INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

XIAOYU TONG DISSECTING INCONGRUITY

Dissecting Incongruity

Metaphor and Humor Understanding of Large Language Models

XIAOYU TONG

Dissecting Incongruity

Metaphor and Humor Understanding of Large Language Models

ILLC Dissertation Series DS-2026-09

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

The research for this doctoral thesis received financial assistance from China Scholarship Council.

Cover design by Xiaoyu Tong.
Cover typeface in Compagnon and Feroniapi distributed by velvetyne.fr.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6536-159-8

Dissecting Incongruity:
Metaphor and Humor Understanding of Large Language Models

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 3 juli 2026, te 11.00 uur

door Xiaoyu Tong
geboren te Hubei

Promotiecommissie

Promotor:	prof. dr. L.W.M. Bod	Universiteit van Amsterdam
Copromotores:	dr. E.V. Shutova	Universiteit van Amsterdam
	dr. M.A.F. Lewis	Universiteit van Amsterdam
Overige leden:	prof. dr. K. Markert	Universität Heidelberg
	dr. D. Elliott	University of Copenhagen
	prof. dr. K. Sima'an	Universiteit van Amsterdam
	dr. J. Bloem	Universiteit van Amsterdam
	prof. dr. T. Blanke	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

Acknowledgments

The completion of this thesis would not have been possible or as rewarding without the support of my supervisors, collaborators, colleagues, family, and friends. First and foremost, to my supervisors, *Ekaterina Shutova* and *Martha Lewis*: Thank you both for guiding me through my PhD journey. It has been a wonderful experience working with you.

Katia, you expressed interest in supervising me two years prior to my PhD study, when I was peeking into natural language processing from a cognitive science vantage point simply because "programming is fun". Your suggestion has led to the most uplifting six years I have ever been through. During the course of my PhD study, you encouraged me through challenges I thought I was not ready for, told me "You're not alone" when I was struggling, and frequently reminded me of how remarkable my achievements are. In addition to the papers we published together, working with you has taught me an important lesson: I should have faith in myself. I feel blessed to continue the adventure of my life with this lesson engrained in me.

Martha, your warm guidance has been an anchor in times of uncertainty—which essentially describe the entirety of my PhD study. You ensured regular meetings with me during my first year, when Katia was on leave and you were in a different time zone. You were always at LAB42 during the last weeks of my thesis writing, offering immediate help when I was stressed out by the impossible deadline. With patience and expertise, you accompanied me through innumerable critical moments, and were always among the first person that cheered with me for every milestone I passed (including when I shaved my head). I feel greatly indebted to your kindness and willingness to work with me.

Second, I would like to thank my collaborators and colleagues: *Rochelle Choenni*, for literally sitting by my side as we worked toward an intimidating deadline, prompting me to stay calm and take the necessary steps to finalize the paper; *Gianluca Michelli*, for being responsive, cordial, and organized as we collaborated remotely; *Zhi Zhang*, for your attention to details and practical approach as we combined our expertise in pursuit of a topic that was novel to both of us; *Pia Sommerauer*, for generously sharing your knowledge and experience in metaphor research, finishing the annotation task despite being on holidays, and giving positive feedback on my work; *Ivo Verhoeven*, for joining the project on short notice and magically turning our curiosity into concrete numbers overnight; and for inviting me to group lunches whenever I was in the office; *Srishti Yadav*, for going through all the image labeling tools you knew until we found the perfect one for my project; *Milan Miletic*, as well as many other lab members, who caught up with me regularly and made me feel like part of the team; and everyone at LAB42 or on Slack that answered my urgent questions.

Furthermore, I am grateful for my parents, *Min Yu* and *Jun Tong*. The dynamics of our family is not the healthiest, but I can see that you are trying your best to support me in every way possible, even when you do not necessarily agree with my choices. Thank you both for all the efforts you have made. I would also like to thank my friends: *Xuan*, for the relaxing moments we shared in Chinese restaurants and your apartment; *Xuelai*, for your hospitality when I frequented your place, and your earnestness when I had an existential crisis regarding all my friendships; and *Jiangyun*—

We had a two-hour Zoom meeting in the last weeks of my thesis writing to brainstorm data analysis methods for the culture chapter; you also gave me practical advice when I was stressed about work. These only scratch the surface of why I am grateful for our friendship.

I will always remember that sunny Friday afternoon, when we ran into each other on a tram, and exchanged names and WeChat IDs. I secretly cherished the discovery that both of our names involve the water element. Despite the multiple layers of serendipity, I did not expect our friendship to last—until it did. We stayed friends even after I became estranged from most of our mutual friends, even when my behavior might, at times, come across as inconsiderate. We spent countless weekends together, exploring new places and fun activities, indulging in bubble teas and various cuisines, laughing at my silly comments on great art pieces. There was never a dull moment with you; even rambling about our respective academic lives felt soothing. While we also spent holidays separately, you always brought me souvenirs that I actually like (it felt like a superpower to me). It was heartwarming every time you came back from a trip and showed me that our friendship had not changed.

I will refrain from logically explaining why all these are related to this thesis. Let me just say that, when you started your new position in Norway, and I found myself planning zero weekend outing for months, it really hit me—that we created a sunlit café that runs 24/7 in the background of my PhD journey.

Amsterdam   Xiaoyu Tong
May, 2006.

Chapter 1

Introduction

1.1 Motivation

> *A snake walks into a bar. The bartender says, "How the hell did you do that?"*[1]

It takes no more than 3 seconds for an English speaker to read the above lines and laugh. Within these few seconds, the brain unpacks relevant common sense knowledge: Walk is a kind of movement of one's feet; snakes do not have feet; a bar that can be walked into and has a bartender present is a place that serves drinks to people who have money to buy drinks; bartenders talk to customers who buy drinks at the bar they work; snakes are not people, cannot order drinks at a bar, and probably do not speak English.

Simultaneously, the brain recognizes the implied *incongruities*—i.e., things that violate the given common sense knowledge: *1)* a snake walking into a bar, and *2)* the bartender regarding the snake as a (human) customer. Such incongruities are an inherent part of both *metaphor* and *humor*. Metaphors are conceptual mappings between a target domain (the domain to be conceptualized) and a source domain (the domain used to conceptualize the target domain) (Lakoff and Johnson, 1980b). The bartender talking to the snake (the second incongruity) utilizes a metaphor that compares SNAKE (target domain) to HUMAN (source domain). Like metaphor, humor is a universal human experience in which incongruity plays a central role (Attardo et al., 2024). The humor of the snake-bartender joke lies in the witty mixture of a snake and the action of walking (the first incongruity) and the bartender's unexpected response to the situation (the second incongruity).

Metaphor and humor are indispensable parts of human cognition and communication. In English language use, metaphor can occur as frequently as in every

[1] https://www.boredpanda.com/bar-jokes/

8 words.[2] The ANIMALS ARE HUMANS metaphor as manifested in the snake-bartender joke is a type of personification that has always been common practice when we conceptualize the behavior of animals or interact with them. The snake-bartender joke is also an instance of "bar joke", a massive collection of jokes that start with "...walks into a bar". A person who is familiar with English jokes would already be expecting a joke once they hear "A snake walks into a bar".

We are so used to metaphor and humor processing that interpreting the snake-bartender joke as intended is easier than finishing a piece of cake, but is this also true for AI agents? The emergence of modern large language models (LLMs) such as LLaMA (Touvron et al., 2023) and GPT-3 (Brown et al., 2020) has significantly expanded the capabilities of AI agents. In zero- and few-shot settings, these general purpose LLMs reach state-of-the-art performance of smaller, fined-tuned models on classic natural language processing (NLP) tasks such as machine translation and question answering. People from all over the world now leverage LLMs for all sorts of complex tasks, including programming, data analysis, and tutoring. While humans rely on shared knowledge to communicate with each other, LLMs are pretrained with an enormous amount of data. But when they interact with human users in real time and receive user input, are they able to link it to relevant knowledge in their system? Are they able to detect incongruities and correctly infer whether the incongruities are intended or mere nonsense, as well as what the intentions are (e.g., as part of metaphor use, or for a humorous effect)? These are all necessary steps even for a one-liner like the snake-bartender joke. Given the ubiquity of metaphor use and the importance of humor for all human beings, failing such steps and, in turn, failing metaphor and humor processing, can gravely compromise the usefulness of LLMs.

1.2 A walk through the chapters

This thesis establishes a series of benchmarks and human baselines to answer the question: **How well do LLMs understand metaphor and humor?** We divide this broad question into the following research questions—

RQ1: How well do LLMs understand metaphorically used words?

We start with metaphors in language use. The groundbreaking finding that metaphors are fundamental components of human cognition (Goatly, 1997; Lakoff and Johnson, 1980b) inspired a new wave of metaphor research in linguistics, psychology, neuroscience, as well as the NLP community. Large metaphor corpora were created, providing metaphor annotations on either word or sentence level

[2]The calculation is based on the VU Amsterdam Metaphor Corpus (Steen et al., 2010b).

(Mohammad et al., 2016; Mohler et al., 2016; Steen et al., 2010b). These datasets were widely used for automatic metaphor detection, but they do not contain information regarding how the annotated metaphors are interpreted. While other works also developed datasets aimed for metaphor interpretation (Bizzoni and Lappin, 2018; Joseph et al., 2023; Shutova, 2010), they were often small in scale (containing 200–1000 instances) and were not designed to test the reasoning process by which a metaphor is interpreted, which remains an open question.

Chapter 3: We design and build a dataset to tap into LLMs' reasoning process during metaphor understanding: the **M**etaphor **Un**derstanding **Ch**allenge Dataset (MUNCH). The dataset provides over 10k paraphrases for sentences containing metaphor use, as well as 1.5k instances containing inapt paraphrases. The inapt paraphrases were carefully selected to serve as control to determine whether the model indeed performs full metaphor interpretation or rather resorts to lexical similarity. All apt and inapt paraphrases were manually annotated. The metaphorical sentences cover natural metaphor uses across 4 genres (academic, news, fiction, and conversation), and they exhibit different levels of novelty. Experiments with LLaMA and GPT-3.5 demonstrate that MUNCH presents a challenging task for LLMs.

RQ2: How well do LLMs infer intentions behind metaphor use?

Understanding the contextual meaning of metaphorical words is not the end of the story. Metaphor use serves a variety of functions in communication: Novelists use metaphors in their writing to invite the reader to visualize the scenes they describe; politicians use metaphors in public speech to persuade the audience to support them; teachers use metaphors in classrooms to explain abstract concepts and complex phenomena. There has been extensive work in the literature linking metaphor to the fulfillment of individual intentions, but no comprehensive taxonomy of such intentions, suitable for NLP applications, is available to date.

Chapter 4: We propose a novel taxonomy of intentions commonly attributed to the use of linguistic metaphors, which comprises 9 categories. We also release the first dataset annotated for intentions behind metaphor use. We use this dataset to test the capability of LLMs in inferring the intentions behind metaphor use, in zero- and in-context few-shot settings. Our experiments show that this is still a challenge for LLMs.

RQ3: How well do LLMs understand humorous multimodal metaphor use?

Humor, as we discuss in Chapter 4, is among the intentions commonly attributed to the use of linguistic metaphors. Yet this relation transcends text and resides in communication in other modalities as well. Given the central role of incongruity in both metaphor and humor, it should come as no surprise that visual metaphors are found to be one of the most common humorous mechanisms in cartoons (Tsakona, 2009). However, the interplay between metaphor and humor, either in text or other modality, is largely overlooked in the NLP community.

Chapter 5: We take inspiration from the Incongruity Theory of humor, the Conceptual Metaphor Theory, and the annotation scheme behind the VU Amsterdam Metaphor Corpus, and develop a novel annotation scheme for humorous multimodal metaphor use in image-caption pairs. We create the HUMMUS Dataset of **Hu**morous **M**ultimodal **M**etaphor **Us**e, providing expert annotation on 1k image-caption pairs sampled from the New Yorker Caption Contest corpus. Using the dataset, we test state-of-the-art multimodal LLMs on their ability to detect and understand humorous multimodal metaphor use. Our experiments show that current LLMs still struggle with processing humorous multimodal metaphors, particularly with regard to integrating visual and textual information.

RQ4: How do cultures differ in humor appreciation?

Humor is a universal human experience, but how humans experience it can be culture-specific. Bar jokes like the snake-bartender one, for example, are strongly tied to Anglophone humor, and language is a mirror of culture. One of the most famous bar jokes uses a pun and an idiom: "A horse walks into a bar. The bartender asks, 'Why the long face?'"[3] The humor will be lost in translation unless the target language/culture shares the idiomatic meaning of *a long face*. The bar joke format is so popular in English-speaking cultures that it has even become a meta humor (e.g., "A panda, a cowboy, a man with a cat on his shoulder, and a time-traveler walk into a bar. 'What is this,' the bartender yells. 'Some kind of joke?'"[4]). People who are unfamiliar with the culture are unlikely to understand why it is humorous.

Cultural alignment in LLMs has gained much attention in the NLP community as LLMs enter the lives of people all over the world. However, the topic has not been studied in the context of humor processing—e.g., whether LLMs align with culture-specific expectations with regard to what is considered humorous, and how to respond to humorous vs unhumorous content. One of the obstacles is the lack

[3] https://en.wikipedia.org/wiki/Bar_joke
[4] https://www.scarymommy.com/bar-jokes

of datasets that reflect cultural differences in humor processing. Previous cross-cultural humor studies typically relied on self-reported humor styles and attitude towards humor (Chen and Martin, 2007; Jiang et al., 2011; Yue et al., 2016), instead of participants' real time reaction to humorous and unhumorous stimuli. Zhang et al. (2024) collected more than 250M human ratings on New Yorker cartoons with English captions, but cultural background was not considered as a variable in their study.

Chapter 6: As a first step towards an evaluation framework for LLMs' cultural alignment in humor processing, we build the first dataset that captures cultural differences in real-time human reactions to multimodal humor. We consider 3 variables: humor appreciation (whether one finds a stimulus humorous), emotion, and metaphor use. Humor and metaphor processing are closely related mental processes, and both are strongly associated with emotional processes. From 4 diverse cultures, Chinese, Mexican, Polish, and the U.S., we obtain 25,600 funniness ratings and emotion responses to 800 New Yorker cartoons (image-caption pairs) in the Hummus dataset, about half of which exhibit humorous multimodal metaphor use. Our data analysis reveals that while the correlation between humor appreciation and various emotions is similar across cultures, there is substantial difference in terms of what is considered humorous within each culture.

1.3 List of publications

During my PhD, I (co-)authored the following papers:

1. **Xiaoyu Tong**, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. https://aclanthology.org/2021.naacl-main.372/.

2. Aarohi Srivastava, Abhinav Rastogi, ... **Xiaoyu Tong**, ... and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research.* https://openreview.net/forum?id=uyTL5Bvosj.

3. **Xiaoyu Tong**, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor Understanding Challenge Dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. https://aclanthology.org/2024.acl-long.193/.

4. Gianluca Michelli*, **Xiaoyu Tong***, and Ekaterina Shutova. 2024. A framework for annotating and modelling intentions behind metaphor use. Under review at **SEM2026*. https://doi.org/10.48550/arXiv.2407.03952.

5. **Xiaoyu Tong**, Zhi Zhang, Pia Sommerauer, Martha Lewis, and Ekaterina Shutova. 2026. Hummus: A Dataset of Humorous Multimodal Metaphor Use. Conditionally accepted by *Transactions of the Association for Computational Linguistics* with minor revisions. https://doi.org/10.48550/arXiv.2504.02983.

6. **Xiaoyu Tong**, Ivo Verhoeven, Martha Lewis, and Ekaterina Shutova. 2026. Evaluating LLMs' cultural alignment in humor appreciation. In preparation.

* denotes joint first authorship.

1.4 List of released datasets and benchmarks

During my PhD, I created the following datasets:

1. MUNCH: github.com/xiaoyuisrain/metaphor-understanding-challenge

2. Hummus: github.com/xiaoyuisrain/humorous-multimodal-metaphor-use

I also co-created the following datasets and benchmarks:

3. BIG-bench metaphor understanding: github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/metaphor_understanding

4. Dataset of intentions behind metaphor use: github.com/GMichelli/intentions-behind-metaphor

Chapter 2

Background

2.1 Theoretical accounts of metaphor

2.1.1 Conceptual metaphors

According to Conceptual Metaphor Theory (CMT), linguistic metaphors have their roots in **conceptual metaphors**, cross-domain mappings in one's conceptual system (Lakoff and Johnson, 1980b). Based on comparable properties and relations in the **target domain** and the **source domain**, conceptual metaphors invite one to conceptualize the former through the latter. For example, Sentence (1) instantiates the conceptual metaphor ARGUMENT IS WAR, in which ARGUMENT is the target domain and WAR the source domain. It uses an event in the domain of WAR to describe an event about ARGUMENT.

(1) He *attacked* every point in her argument. (Macmillan)

The target and source domains of a metaphor usually involve abstract and concrete concepts respectively, which has to do with concept representation in the brain. According to the theory of embodied cognition (Barsalou, 1999), concepts are represented within sensorimotor circuitry: GRASPING, for instance, is represented in areas that control hand movements (Gallese and Lakoff, 2005). When using expressions such as *grasp a point*, the same areas are involved for processing the metaphor (Aziz-Zadeh and Damasio, 2008).

Conceptual metaphors that are entrenched in one's conceptual system are termed **conventional metaphors**; those that are not entrenched are termed **novel metaphors**. ARGUMENT IS WAR is a typical conventional metaphor; the expression *attack one's argument*, for example, has entered contemporary dictionaries of English and is unlikely to be considered novel by the native speakers. On the contrary, Sentence (2) instantiates a novel metaphor, FEAR IS A SNAKE. The verb has a single sense in the Macmillan dictionary: a particular movement of "something long and thin". Fear is an abstract concept, incapable of any movement to be seen. The usage is therefore novel.

(2) Fear *coiled* around his heart.

2.1.2 Linguistic metaphors and figures of speech

Informed by CMT, Steen et al. (2010b) proposed Metaphor Identification Procedure VU University Amsterdam (MIPVU) for manually identifying **metaphor related words (MRWs)** in texts, which are word uses that can be attributed to underlying conceptual metaphors. What are usually called metaphorically used words are a subset of MRWs termed **indirect metaphors**. The source-domain word *attack* in Sentence (1) is an indirect metaphor: It refers to a target-domain act, and is therefore indirectly related to the underlying conceptual metaphor, ARGUMENT IS WAR.

Other types of MRWs include direct metaphors, implicit metaphors, and metaphor flags. **Direct metaphors** refer to source-domain entities or events. They often co-occur with **metaphor flags**, signals that a metaphor is used. In Example (3), the linking verb *is* signals a use of the BEAUTY AS FLOWER metaphor. The noun *flower* is a direct metaphor, referring directly to a source-domain concept.

(3) Beauty is but a flower / which wrinkles will devour; [..] (Thomas Nashe)

Implicit metaphors are substitutions or ellipses that are directly or indirectly related to underlying conceptual metaphors. In Example (4), the pronoun *that* co-refers with the preceding metaphorically used *antidote*; the determiner *this* in the following sentence is also used metaphorically: it refers to what is talked about in the previous sentence, which is an abstraction of its basic, physical meaning (Steen et al., 2010b). Metaphorical usage of substitutions are important for rendering cohesive discourse (Steen et al., 2010b).

(4) Fortunately, there is a single *antidote* effective against both these myths; and *that* is to start all over again .... *This* antidote is effective against the romantic-individualist myth .... (BNC)

Discourse-level information is essential for identifying **extended metaphors**, sustained use of the same metaphors in a discourse fragment. A typical example is the "All the World's a Stage" speech written by William Shakespeare. As is presented below, the speech begins with a WORLD AS STAGE metaphor, and proceeds with exploring various target-source pairs within the metaphor. Note that the last two lines could be mistaken as literal if presented as independent sentences.

(5) All the world's a stage, / And all the men and women merely players: / They have their exits and their entrances; / And one man in his time plays

many parts, [..] (William Shakespeare)

Figurative language. Metaphor, simile, personification, and zoomorphism are different types of figurative language, but they are all considered as metaphors under CMT. **Similes** are linguistic metaphors that use metaphor flags such as *like* and *as* to highlight a cross-domain mapping (e.g., *my love's like a red, red rose*); metaphor as a figure of speech uses categorization statements, making the cross-domain mapping more implicit (e.g., *my love is a red, red rose*). **Personification** is metaphor that compares something non-human to humans (e.g., pets as family members). **Zoomorphism** is metaphor with ANIMALS or a species of animal as the source domain (e.g., *the roar of the ocean*).

Idioms are sometimes called dead metaphors, expressions that have lost their metaphoricity over time. There is empirical evidence that people process idioms and metaphors differently (Desai, 2022; Gibbs, 1992). We therefore treat idiom and metaphor as different phenomena.

Puns play on the double meaning of the same word or similar-sounding words. While metaphors also essentially bring together two different meanings (target domain and source domain), puns are not necessarily metaphorical as the two meanings at play do not have to be the target and source domains of a metaphor. Example (6-a) uses a pun on *bar*: It achieves a humorous effect by activating both the meaning of "a place to buy drinks" and the meaning of "a long straight piece of metal or wood". However, the word use is not metaphorical because the two meanings do not share any similarities for them to be compared in a metaphorical way. Example (6-b), on the other hand, uses a metaphorical pun. The word *running* means both "functioning" and the physical movement of running; the "functioning" meaning is a conventional metaphor based on the physical meaning. An OBJECTS ARE HUMANS metaphor is used at the same time, as the refrigerator is personified.

(6) a. A man walks into a *bar*. He said, "Ouch."
b. Is your refrigerator *running*? Better catch it before it gets away.

2.1.3 Non-verbal and multimodal metaphors

In addition to linguistic metaphors, conceptual metaphors also give rise to metaphor use in other modes of communication. Representing an idea as a light bulb, for example, is a common **visual metaphor**. There are also **metaphoric gestures**: One may talk about events that happen in a sequence while shifting one's body from left to right, as if the events can be placed next to each other in the current space. Such use of the TIME IS SPACE metaphor is very common in gestures, and is likely related to a left-to-right writing direction adopted by the speaker's native language (e.g., most Indo-Enropean languages) (Cienki and Müller, 2008).

"Three yea's, six ney's, and Anderson is still up in the air on this one."

Figure 2.1: An example of multimodal metaphor.

There are also metaphors that involve more than one mode of communication. We call these **multimodal metaphors**. For example, the image in Figure 2.1 shows business people having a meeting; one of them seems to be defying gravity. The caption implies that the person is literally in the air because they have not made a decision. The key phrase, *up in the air*, is an idiom or dead metaphor. When the caption is considered as a standalone text, the idiom does not count as metaphor use. However, when the caption and the image is combined as a whole, the underlying conceptual metaphor, MAKING A DECISION IS SETTLING DOWN, is revived. The target domain of DECISION MAKING is represented in the text; the source domain of SETTLING DOWN is both depicted in the image and implied by the text. As the metaphor makes use of multiple modes (text and image), it is an instance of multimodal metaphor use.

2.1.4 Metaphor comprehension

There has been a debate about whether metaphor comprehension is a comparison or a categorization process. Consider again the poem in Example (3). According to the comparison view (e.g., Ortony, 1979), when one processes the poem, one searches for the shared properties and/or relational structures of the BEAUTY domain and the FLOWER, such as being pleasing to the eye, and the possibility of being fragile. Scholars advocating the categorization view (e.g., Glucksberg and Keysar, 1990), however, would argue that the object, *a flower*, refers to a super-ordinate category that both BEAUTY and FLOWER belong to.

The Career of Metaphor Theory (Bowdle and Gentner, 2005) suggests that both comparison and categorization are possible paths of metaphor processing; which path is chosen depends on an interaction between the conventionality of the metaphor and its linguistic realization. More specifically, novel metaphors are processed through comparison; as a metaphor becomes conventionalized, people tend to process it through categorization, which is less cognitively demanding than a comparison process (Bowdle and Gentner, 2005).

An inevitable result of metaphor processing is the emergence of features not inherent in the target or the source domain (Tourangeau and Rips, 1991). This notion of **emergent meaning** corresponds to the connotations or inference of metaphorical language in linguistics literature. Consider the difference in meaning between *attack one's argument* and *criticize one's argument*. The first contains far richer shades of meaning than the second. It might be possible to express a similar meaning without using non-literal language, but the expression is unlikely to be as concise as the use of a single word, *attack*.

2.1.5 Metaphor use in communication

The linguistic expression of emotional states often employs metaphor (Fainsilber and Ortony, 1987; Fussell and Moss, 1998). As emotion is an abstract domain, it goes with CMT and embodied cognition that we employ more concrete domains, such as physical or bodily experience, to conceptualize it. Moreover, since metaphor gives rise to emergent meaning, metaphorical language has a stronger emotional effect than literal language, regardless of the source and target domains involved (Blanchette et al., 2001; Crawford, 2009; Mohammad et al., 2016). For instance, Citron and Goldberg (2014) found that metaphorical expressions involving taste (e.g., *She looked at him sweetly*) would evoke a higher emotional response than their literal counterparts (e.g., *She looked at him kindly*).

Metaphor has also proved to be an effective persuasive device (Sopory and Dillard, 2002; van Stee, 2018). The persuasive power of metaphors is pronounced in metaphoric framing effect. Since metaphors encourage a particular way to conceptualize the target domain, repeated use of the same metaphors throughout discourses in mass media tends to affect how the public perceives and reacts to societal issues that belong to the target domain (Komatsubara, 2024; Lakoff, 1991; Lakoff and Wehling, 2012). For instance, participants in a series of studies (Thibodeau and Boroditsky, 2011, 2013) favored different social solutions to crime after reading articles that associate CRIME with different source domains. Moreover, while the participants could identify the implicitly advocated solutions given a metaphor, they were unaware of the influence of the metaphors on their own preference.

The Deliberate Metaphor Theory is an attempt to link metaphor use in communication with the cognitive process of metaphor comprehension (Steen, 2008, 2017). The theory defines **deliberate metaphor use** as the intentional introduction of a topic shift or perspective change to the discourse. Deliberate metaphor use is associated with online metaphor processing, the construction of conceptual metaphors during text comprehension (Steen, 2017). Examples of deliberate metaphors include conventional metaphors instantiated as copula metaphors and all novel metaphors. A systematic procedure for the identification of potential deliberate metaphors has also been proposed (Reijnierse et al., 2018).

2.2 Computational approaches to linguistic metaphors

2.2.1 Datasets

Metaphoricity annotation. Tsvetkov et al. (2014) released a dataset (henceforth: TSV) consisting of an equal number (884) of metaphorical and non-metaphorical adjective-noun (AN) phrases collected from the web. The phrases were stated to be verified by multiple annotators, but the criteria for metaphor annotation were not provided.

The dataset released by Mohammad et al. (2016) (henceforth: MOH) consists of 1639 (1230 literal and 409 metaphorical) sentences extracted from WordNet, manifesting the use of 440 verbs. Metaphoricity annotation of the verb uses was obtained through crowdsourcing. Note that the WordNet sentences are mainly instances of conventional metaphor. The specified association between word senses and metaphoricity makes it easier to determine the source and target domains involved.

The Language Computer Corporation metaphor datasets (LCC; Mohler et al., 2016) contain linguistic metaphors in four languages: English, Spanish, Russian, and Farsi. The metaphors were extracted from web corpora and cover a small set of target domains. Metaphoricity was annotated at sentence level; conceptual metaphors and affect information was also specified.

VU Amsterdam Metaphor Corpus (VUA; Steen et al., 2010b) includes 115 text fragments from the British National Corpus (BNC Consortium, 2007), spanning four genres: academic, news, fiction, and conversation. The corpus was annotated using MIPVU, which was developed by the same authors. Every single lexical unit ($\sim$190,000 in total) was annotated in terms of whether it was an MRW, including function words.

The TOEFL metaphor corpus (Beigman Klebanov et al., 2018) contains 240 argumentative essays sampled from the ETS Corpus of Non-Native Written English (Blanchard et al., 2014). The native language of the writers is either Japanese, Italian, or Arabic. Metaphoricity was annotated at word level, following a protocol proposed by Beigman Klebanov and Flor (2013) for identifying metaphors that help advancing an argument.

Metaphor paraphrase datasets. Mohammad et al. (2016) also obtained literal paraphrases of the sentences, in which the metaphorically used verbs are replaced by their synonyms (171 pairs of sentences in total), selected by the authors.

Note that the literal paraphrases were considered to convey less emotion than the original metaphorical sentences in their experiment. It is therefore questionable to what extent the paraphrases capture the connotations of the metaphorical sentences.

Bizzoni and Lappin (2018) built a metaphor paraphrase dataset (MPEC) containing 200 sets of 5 sentences; 4 paraphrases at varying levels of aptness are provided for each metaphorical sentence. Apart from verbs, the dataset also includes metaphorical uses of adjectives, copula metaphors, and multi-word metaphors. The dataset takes into account the connotations of the metaphorical sentences to some extent. For instance, candidate paraphrases for the copula metaphor *My job is a dream* include *I love my job* and *I hate my job*, which indicate opposite sentiment poles. A metaphor processing system will need to infer the sentiment to select the apt paraphrase.

Taking inspiration from the MPEC dataset, I proposed a method to construct apt and inapt paraphrases systematically, which essentially treat metaphor paraphrasing as a word sense disambiguation task (Tong, 2021). All candidate paraphrases replace the indirect metaphor in a reference sentence with a single word, and the replacement word corresponds to *1)* the contextual meaning of the metaphor (target domain), *2)* the more basic meaning (source domain), and *3)* a sense of the metaphorical word that is irrelevant to the metaphor. A selection of the dataset became part of the metaphor-understanding benchmark task in BIG-bench (Srivastava et al., 2023), together with MPEC. A major limitation of this method, however, was that it was rather costly to scale up.

2.2.2 Metaphor identification

Neural architectures

Most neural models treat metaphor identification as a sequence labelling task, outputting a sequence of metaphoricity labels for a sequence of input words (usually a sentence) (Bizzoni and Ghanimifard, 2018; Chen et al., 2020; Dankers et al., 2019; Gao et al., 2018; Gong et al., 2020; Mao et al., 2019; Mykowiecka et al., 2018; Pramanick et al., 2018; Su et al., 2020; Wu et al., 2018). The first sequence labelling systems typically represented an input sentence as a sequence of pre-trained word embeddings and produced a task- and context-specific sentence representation through bidirectional long short-term memory (BiLSTM) (Dankers et al., 2019; Gao et al., 2018; Mykowiecka et al., 2018; Pramanick et al., 2018). Bizzoni and Ghanimifard (2018) experimented with separating long sentences into smaller chunks, which led to a 6% increase in F-score when using a BiLSTM architecture. Their BiLSTM system outperformed their compositional system, which employed a sequence of fully-connected neural networks (NNs) and essentially performed bigram phrase composition to modulate the representation of input words with respect to their neighbors. BiLSTM models also outperformed bidirectional gated recurrent unit (BiGRU) models in the study of Mykowiecka et al. (2018). From Gao et al. (2018), the contextualized Embeddings from Language Models (ELMo) (Peters et al., 2018) began to be used in addition to the context-free Global Vectors (GloVe) (Pennington et al., 2014) for representing input sentences (Dankers et al., 2019; Gao et al., 2018; Mao et al., 2019). The most recent systems adopted a fine-tuning approach, employing pre-trained contextual language models such as Bidirectional Encoder Representations from Transformers (BERT) (Chen et al., 2020; Dankers et al., 2019) and RoBERTa (Gong et al., 2020).

Several BiLSTM-based systems consider both contextualized and pre-trained representations in the classification layers (Mao et al., 2019; Swarnkar and Singh, 2018). The Di-LSTM Contrast system (Swarnkar and Singh, 2018) encodes the left- and right-side context of a target word using forward and backward LSTMs. The classification is based on a concatenation of the target-word vector and its difference with the encoded context. Mao et al. (2019) combined GloVe and BiLSTM hidden states for sequence labelling, which outperformed the best model in the 2018 VUA All POS track.

Wu et al. (2018) and Su et al. (2020) employed separate encoding of local and long-range context. Wu et al. (2018) used a convolutional neural network (CNN) and a BiLSTM to extract local and sentence context respectively. The DeepMet architecture proposed by Su et al. (2020) uses separate Transformer encoder layers to encode global and local text features for each word. The two systems achieved the best performance on the VUA All POS tracks in their respective shared tasks (Leong et al., 2018, 2020).

Modeling metaphor in discourse. Several approaches have also incorporated wider discourse properties in their models. Mu et al. (2019) focused on the metaphoricity of verbs. They used general-purpose word, sentence and document embedding methods (e.g. GloVe, ELMo, doc2vec (Le and Mikolov, 2014), skip-thought (Kiros et al., 2015)) to represent the surrounding paragraphs. Their system feeds into a gradient boosting decision tree classifier (Chen and Guestrin, 2016) a concatenation of three feature vectors, representing *1)* the lemma of the target verb, *2)* its subject and direct object, and *3)* its surrounding paragraph. All the representations are learned from embedding methods. Representing the features with ELMo led to the highest F-score (0.668), using the VUA Verbs 2018 shared task data.

Dankers et al. (2020) fine-tuned a BERT model that receives a discourse fragment as input. Using hierarchical attention (which computes both token- and sentence-level attention) after the encoded layers achieved better performance than applying general attention to all tokens. Both Dankers et al. (2020) and Mu et al. (2019) thus demonstrated the importance of context beyond sentence for word-level metaphor identification. Their qualitative analysis shows that coreference resolution is one of the driving factors in the performance increase.

The above systems use discourse to aid in detecting linguistic metaphors whose metaphoricity is otherwise ambiguous, but do not monitor whether a metaphor is sustained throughout a fragment of discourse. Thus, they are unlikely to be directly applicable to the identification of extended metaphors or metaphoric frames.

Cognitively-inspired approaches

Categorial features Metaphor processing is concerned with how concepts are organized in the brain/mind, and is closely related to categorization. It therefore makes sense to employ categorial features for metaphor identification. Tekiroğlu et al. (2015) tested the use of sensorial categories (the five human senses) for identifying AN synaesthetic metaphors (e.g., *sweet music*, *soft light*). Using sensorial categories in addition to WordNet supersenses, concreteness, and imageability led to improved performance (accuracy 0.890 vs 0.845 on TSV). Mykowiecka et al. (2018) used another general-purpose resource, the Harvard IV psychosocial dictionary[1], which includes categories of emotions, people and animals, objects, places, etc. However, it did not lead to consistent improvement in model performance. Bulat et al. (2017) compared property-based and linguistic embeddings for input word representation. They obtained property-based word embeddings by mapping linguistic word embeddings onto a conceptual space, using as training data a frequency-based human property-norm dataset (McRae et al., 2005). Using property-based word embeddings led to a 4% increase in F-score on TSV.

Sensory features An important function of conceptual metaphors, according to CMT, is to use bodily experience to understand abstract concepts; concreteness features have therefore also proved useful for automated metaphor identification (Bizzoni and Ghanimifard, 2018; Turney et al., 2011). Shutova et al. (2016) tested combination of visual and linguistic embeddings on MOH and TSV. The multimodal system outperformed the monomodal systems in both tests. Gong et al. (2020) also included both categorial and sensory features. In addition to RoBERTa, the system employs concreteness features, topic distributions, WordNet classes, VerbNet classes, verb clusters, and part of speech (POS), which led to improvements in performance on VUA All POS.

Word-context incongruity The neural model of Swarnkar and Singh (2018) computes the difference between a target word and its context. This operation can be associated with the comparison view of metaphor, including CMT, and is reflected in MIPVU. Shutova et al. (2016) used cosine similarity between word or phrase embeddings to predict the metaphoricity of verb-noun (VN) or AN pairs;

[1] http://www.wjh.harvard.edu/~inquirer/homecat.htm

a word pair is marked metaphorical if the cosine similarity is below a trained threshold. The systems reached an F-score of 0.71 on MOH and 0.76 on TSV using linguistic embeddings alone. Rei et al. (2017) proposed a supervised similarity network, which learns to calculate weighted cosine similarity in a task-specific vector space. This allows the model to learn which dimensions of similarity are most relevant in particular metaphoric comparisons. The system outperformed Shutova et al. (2016) without the use of visual representations. Mao et al. (2018) dealt with word-level metaphor identification in sentences. Given a target word in a sentence, the system searches WordNet for the synonym or direct hypernym of the target word most similar to the context words. The target word is metaphorically used if its cosine similarity with the selected word is below a threshold.

Metaphor and emotion Motivated by the close relationship between metaphor use and the expression of emotions, Gargett and Barnden (2015) successfully used emotion features, amongst others, for metaphor identification. Kozareva (2013) and Strzalkowski et al. (2014) modeled the affect carried by metaphors in texts in different languages. Most recently, Dankers et al. (2019) employed multi-task learning (MTL) to train models of metaphor identification and emotion prediction jointly. Models were based on BiLSTM and BERT, with a range of MTL architectures. The emotion prediction task used the Valence-Arousal-Dominance model (Mehrabian, 1996), and each of these was considered separately in a MTL setup. The best performance was achieved with the BERT architecture. Dankers et al. (2019) found that while predicting dominance was the most challenging task on the emotion side, it also provided the greatest and most consistent improvements to metaphor identification, and vice versa.

2.2.3 Metaphor interpretation

Recent research on automated metaphor interpretation mainly followed Shutova (2010) in treating the problem as a paraphrasing task. Su et al. (2017) proposed a property transfer process for the interpretation of copula metaphors in Chinese and English. Given a target-source pair, the system extracts source-domain properties from hand-crafted databases[2], represented as adjectives. It then selects the property that contributes the most to the semantic relatedness of the target-source pair. The resultant pair of target-domain word and property is taken as interpretation of the copula metaphor. The metaphor LOVE IS TIDE, for instance, was interpreted as *The love is unstoppable*. Su et al. (2017) thus took into account the emergent meaning of metaphors. Note that the literal explanations they obtained can be regarded as explanations of conceptual metaphors as

[2]For instance, the adjective taxonomy provided by Sardonicus: http://bonnat.ucd.ie/sardonicus/.

well.

The above-mentioned system of Mao et al. (2018) performs both metaphor interpretation and identification: the synonym or hypernym of the metaphoric target word that matches the context can be considered the literal counterpart of the target word. The output interpretations were evaluated through a English-Chinese machine translation task: words classified as metaphorically used were paraphrased prior to translation. The system improved the accuracy of both Google Translation (0.60 vs 0.34) and Bing Translation (0.66 vs 0.42) on the metaphorical sentences in MOH. The experiment thus demonstrated the value of metaphor interpretation for machine translation.

Bizzoni and Lappin (2018) presented a neural model that detects paraphrases of sentences containing metaphor use. Given a metaphorical sentence and a candidate paraphrase, the system uses parallel CNN-LSTM blocks to encode the two sentences separately. The sentence representations are then merged and passed through fully-connected layers to produce a prediction. The system reached an F-score of 0.746 in a binary classification task and a Pearson correlation of 0.553 in a paraphrase ordering task, on the authors' own dataset.

2.3 Computational approaches to visual and multimodal metaphors

2.3.1 Videos

Alnajjar et al. (2022) released a metaphor corpus containing 27 YouTube videos ($\sim$4 hours long in total), with metaphoricity annotation on the texts. Words or phrases involving metaphor use were highlighted and labelled as either vehicle (the source domain) or tenor (the target domain). A total of 371 metaphorical expressions were identified. The study found in the annotation process that gestures in the videos could make certain metaphors more salient (e.g., accompanying the expression *sprinkling keywords* with a sprinkling gesture). However, since the videos themselves are not annotated, the resulting corpus does not contain information about how the metaphorical expressions in text are related to other modes.

Alnajjar et al. (2022) adopted the DeepMet architecture proposed by Su et al. (2020) for the metaphor detection task. The best performance (F1 0.62) was achieved when the model was trained on VUA and later fined-tuned using their own corpus. The authors also experimented with an audio model (Facebook's XLSR-Wav2Vec2 pretrained in multiple languages (Baevski et al., 2020; Conneau et al., 2020)), a video model (18 layer deep R(2+1)D network (Tran et al., 2018) pretrained on the Kinetics human action video dataset (Kay et al., 2017)), and three mulimodal models that combined the text-only model with one or both of the other models. However, none of these architectures outperformed the text-only model.

2.3.2 Memes

Liu et al. (2022a) created the FigMemes dataset for figurative language classification in politically-opinionated memes. A total of 5,141 memes were collected from an online bulletin board[3] and classified in terms of figurative language use. Six categories of figurative language were identified: allusion, exaggeration/hyperbole, irony/sarcasm, anthropomorphism/zoomorphism, metaphor/simile, and contrast. Around 70% (3,600) of the memes were annotated as using one of the 6 types of figurative language. Metaphor/simile was found to be the most frequent (30%). The dataset was used to train and test text-only, image-only, and multimodal models on a figurative language classification task. The highest performance in identifying the use of metaphor/simile (F1 44.87) was achieved by an image-only model that uses the CNN component of the Contrastive Language-Image Pre-Training (CLIP) model (Radford et al., 2021) with linear probing (only the classification layer was trained). CLIP (the multimodal version with the image and text representations concatenated for classification) with full fine-tuning reached the highest F1 score (41.76) on identifying the use of anthropomorphism/zoomorphism.

Xu et al. (2022) created the MET-Meme dataset, which contains 6,045 Chinese and 4,000 English memes with metaphor, sentiment, intention, and offensiveness annotations. Metaphor annotations include *1)* whether or not the meme uses metaphor; *2)* whether the metaphor resides in the image, text, or both modes; *3)* the name of the target/source domain and whether it is represented in the image or text. Sentiment annotation involves 7 categories: happiness, love, anger, sorrow, fear, hate, and surprise. Intention annotation considers 5 categories: interactive, expressive, purely entertaining, offensive, and other. Offensiveness is annotated on a scale from 0 (non-offensive) to 3 (very offensive). A total of 3,441 (34%) memes were annotated as metaphorical. The dataset was used to train and test text-only, image-only, and multimodal models on metaphor understanding, sentiment analysis, intention detection, and offensiveness detection. Multilingual BERT[4] with VGG16 (Simonyan and Zisserman, 2015) for image processing reached the highest F1 score (0.82) on understanding the metaphors in English memes. Multilingual BERT with Resnet50 (He et al., 2016) reached the highest F1 score (0.77) on understanding the metaphors in Chinese memes.

2.3.3 Visual advertisements

Akula et al. (2023) built a dataset of visual metaphors in advertisements, contain-

[3] boards.4chan.org/pol/
[4] https://github.com/google-research/bert/blob/master/multilingual.md

ing 5k samples that were identified as metaphorical and 6.5k distractors, which comprise a balanced number of symbolic and non-symbolic samples. The ads were obtained from an existing dataset of image and video ads (Hussain et al., 2017). Each of the metaphor samples is also annotated with a one-sentence metaphor interpretation statement in the form “<target> is as <property> as <source>”, as well as bounding boxes around representations of target- and source-domain concepts in the images. Based on the dataset, the authors proposed MetaCLUE, a set of computer vision tasks concerning visual metaphors: *1)* classification (whether a given image uses metaphor), *2)* understanding (interpretation statement), *3)* localization (bounding boxes), and *4)* generation (given an interpretation statement, generate a visual metaphor).

Akula et al. (2023) evaluated state-of-the-art models on MetaCLUE, including EfficientNet (Tan and Le, 2019) and Vision Transformer (ViT; Dosovitskiy et al., 2021) for task 1; CLIP, ALBEF (Li et al., 2021), and PaLI-17B (Chen et al., 2023) for task 2; CLIP based phrase localization model (Li et al., 2022) for task 3; Imagen (Saharia et al., 2022) and Stable Diffusion (Rombach et al., 2021) for task 4. The experiments indicated that the tasks were challenging for the models. For example, both fine-tuned EfficientNet and ViT models struggled with differentiating metaphor samples from symbolic distractors (highest accuracy 0.67, achieved by ViT-B/16). Both CLIP and ALBEF models struggled in metaphor understanding when asked to choose between the correct interpretation statement and a distractor that swapped the target and the source domain (highest accuracy 0.50, achieved by ALBEF with Flickr30k image-encoder in zero-shot setting).

2.4 Humor and metaphor

2.4.1 Humor theories

Humor theories are usually divided into 3 groups: incongruity theories, hostility/disparagement theories, and release theories (Attardo, 1994; Morreall, 1983; Raskin, 1984). **Incongruity** theories of humor consider perceived incongruity as the cause of laughter. Beattie (1779), who is among the first to explicitly articulate this idea, wrote the following:

> May we not then conclude, that “Laughter arises from the view of two or more inconsistent, unsuitable, or incongruous parts or circumstances, considered as united in one complex object or assemblage, or as acquiring a sort of mutual relations from the peculiar manner in which the mind takes notice of them?”

Suls (1972) proposed a two-stage model for humor appreciation: perception of incongruity and resolution of incongruity. Consider Example (7). The first sentence introduces an incongruity: It is such a strange question. A guy who is really loud can be called any name. The answer "Mike" is a witty pun that resolves the incongruity—a mike amplifies one's voice, and Mike happens to be a name that is usually given to boys. We use this model in Chapter 5 to understand the humor in New Yorker cartoons.

(7) What do you call a guy who's really loud? Mike.[5]

Attardo (1994) argued that the incongruity view is essentialist, and therefore "not incompatible" with the other two groups of humor theories. **Hostility/disparagement** theories regard humor in the context of social interaction, positing that laughter expresses a sense of superiority over the object being laughed at. Thomas Hobbes famously described the passion of laughter as "sudden glory" as one recognizes one's "eminency" compared to the "infirmity" of others or one's former self (quoted from Raskin, 1984). Hutcheson (1750) critiqued this view by arguing that feelings of superiority are neither necessary nor sufficient for laughter to arise. **Release** theories focus on the psychological effect of laughter. In essence, these theories contend that humans are constantly under constraints (e.g., social norms and expectations), and humor offers a way to release the nervous energy that has built up over time. Advocates of this view of humor include Spencer (1875) and Freud (1960). Modern empirical studies continue to investigate the use of humor as a strategy to cope with stress or regulate emotions (e.g., Kugler and Kuhbandner, 2015; Kuiper et al., 2004; Mesmer-Magnus et al., 2012).

2.4.2 Humor styles

Martin et al. (2003) classified one's use of humor along two axes: *1)* whether one's use of humor serves to enhance oneself or one's relationship with others, and *2)* whether one's relationship with others or oneself is harmed in the process. Four humor styles were thus identified: *1)* **Self-enhancing humor** enhances oneself without harming one's relationship with others; *2)* **aggressive humor** enhances oneself at the expense of interpersonal relationships; *3)* **affiliative humor** enhances one's relationship with others while also being kind to oneself; *4)* **self-defeating humor** enhances one's relationship with others at the expense of oneself.

Martin et al. (2003) also developed the Humor Styles Questionnaire (HSQ), which has been widely used in the literature to assess one's humor style. Kuiper et al. (2004), for example, adopted HSQ along with two other self-report measures (Coping Humor Scale (Martin, 1996); Humorous Behavior Deck-Revised (Kirsh and Kuiper, 2003)) to study the relationship between one's sense of humor and psychological well-being (e.g., self-esteem, depression level). They found that self-enhancing and affiliative humor were associated with better psychological well-being as compared to self-defeating humor. Aggressive humor, on the other

[5] https://parade.com/1287449/marynliles/short-jokes/

hand, was not related with psychological well-being. HSQ has also been used to study cultural differences in the usage of humor (e.g., Chen and Martin, 2007; Hiranandani and Yue, 2014; Kalliny et al., 2006). Schermer et al. (2023) collected HSQ responses from 28 countries and found affiliative humor to be the most popular humor style across all countries.

2.4.3 Figurative humor

Humor has a close relationship with figurative language. Figurative language is "omnipresent across humorous genres (Godioli and Chłopicki, 2024)". Types of figurative languages that are frequently used for humorous purposes include metaphor, pun, hyperbole, irony, sarcasm. Visual metaphors are among the most common humorous mechanisms in cartoons (Tsakona, 2009).

In the NLP community, the relationship between humor and metaphor use has not received much attention. The FigMemes dataset and the MET-Meme dataset provide metaphor annotation for memes, which can be humorous (Tanaka et al., 2022), and both are included in a recent survey of computational modeling of humor (Lemmens and De Marez, 2026). However, neither studies make an explicit connection between humor and metaphor use. The datasets do not contain information about, for example, whether the identified metaphors are used for a humorous purpose.

Chapter 3

The Metaphor Understanding Challenge Dataset

3.1 Introduction

LLMs such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and LLaMA (Touvron et al., 2023) have become a common paradigm in NLP research. Several benchmarks have been proposed to investigate the capabilities of LLMs (Hendrycks et al., 2021; Liang et al., 2022; Srivastava, 2022); and comprehensive analyses have been conducted, evaluating their performance on a range of NLU tasks (Bang et al., 2023; Kocoń et al., 2023; Qin et al., 2023; Ye et al., 2023; Zhong et al., 2023). The community has extensively examined LLM performance on question answering, summarization, sentiment analysis, natural language inference; a few studies have also shed light on LLMs' analogical reasoning capabilities (Czinczoll et al., 2022; Webb et al., 2023). However, the ability of LLMs to comprehend metaphor—a fundamental linguistic and cognitive tool—is still poorly understood.

Metaphors are linguistic expressions based on conceptual mappings between a target and a source domain (Lakoff and Johnson, 1980b). The verb phrase *to <u>stir</u> excitement*, for example, is based on the conceptual metaphor FEELING IS LIQUID, with FEELING (excitement) being the target domain and LIQUID (something that can be stirred) the source domain. The metaphor compares FEELING with LIQUID, introducing vividness into the description of an otherwise intangible emotional impact. Such cross-domain mappings are sets of systematic ontological correspondences, mapping concepts and their relational structure across distinct domains. Performing this mapping is an essential part of reasoning involved in the interpretation of metaphorical language (Gentner and Markman, 1997; Grady et al., 1999; Lakoff, 2014).

Humans use metaphors so naturally and frequently that they largely fly under our radar. In VUA (Steen et al., 2010b), one of the largest metaphor corpora annotated by linguists, every 8th word is metaphorical, as averaged over four dif-

⊙	Mark's promotions and progress up the company ∗ladder∗.	No golden light ∗bathed∗ the red brick of the house.
⊛	hierarchy, payroll	covered, enveloped, illuminated, immersed, reached
⊜	✓ hierarchy, ✗ steps	✓ covered, ✗ cleaned

Table 3.1: MUNCH dataset samples. Each metaphor sample (⊙) has a ∗highlighted∗ word that is metaphorically used. The dataset provides up to 5 crowdsourced paraphrases (⊛) for each metaphor sample: Substituting the highlighted word (e.g., "ladder") with one of the provided words (e.g., "hierarchy" or "payroll") should result in an apt paraphrase. For a selection of metaphor samples, the dataset also provides expert annotation (⊜) of a pair of correct (✓) and incorrect (✗) substitution words.

ferent genres, including academic and conversation. LLMs, therefore, require the ability to comprehend metaphor in order to have a full command of language. As such, metaphor understanding is an essential task for evaluating the capabilities of LLMs.

Several corpora have been created that contain metaphor annotations at either word or sentence level. These include the VUA corpus (Steen et al., 2010b), the LCC metaphor datasets (Mohler et al., 2016) and the metaphor-emotion dataset of Mohammad et al. (2016), among others. These datasets have been widely used to develop and evaluate automated metaphor identification systems (see Tong et al. (2021) for a survey), but they do not contain information of how the annotated metaphors are interpreted. On the other hand, several works developed datasets with a focus on interpretation, typically casting the problem as a paraphrasing task (Bizzoni and Lappin, 2018; Joseph et al., 2023; Shutova, 2010). Yet, those datasets were often small in scale (containing 200–1000 instances) and were not designed to test the reasoning process by which a metaphor is interpreted, which remains an open question.

This study presents a novel Metaphor Understanding Challenge Dataset for LLMs (MUNCH). It provides over 10k paraphrases for metaphorical sentences and 1.5k triples of a metaphorical sentence and two candidate paraphrases, which could be apt or inapt (for dataset examples see Table 3.1; for statistics see Appendix A.1). The metaphorical sentences were extracted from VUA texts, spanning four genres (academic, news, fiction, and conversation) and featuring metaphors at different levels of novelty. Each metaphorical sentence contains a content word that is marked as metaphorically used. A candidate paraphrase replaces the metaphorical word with another word, so that the resulting sentence is the same as the reference sentence except for that one word, therefore representing a lexical substitution task. An apt paraphrase shows correct contextual

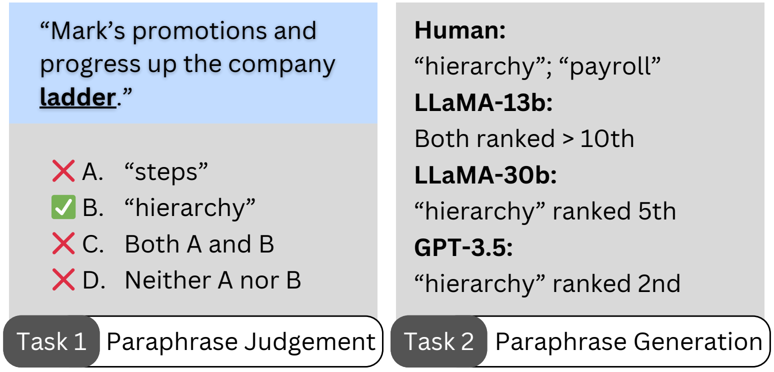


Figure 3.1: Two tasks for MUNCH: Given a sentence containing a metaphorically used word, a model is prompted to *1)* select correct paraphrases from two given candidates (Paraphrase Judgement), and *2)* paraphrase the sentence by replacing the highlighted metaphorically used word (Paraphrase Generation).

interpretation of the metaphor while an inapt paraphrase uses a word that is related to a literal, source domain sense of the metaphorical word (see the examples of correct and incorrect substitution words in Table 3.1). Such a setup of the task is inspired by Conceptual Metaphor Theory (Lakoff and Johnson, 1980b) and allows us to investigate whether the model performs full metaphor interpretation by cross-domain mapping or rather resorts to more shallow lexical similarity. In order to investigate this in a more controlled fashion, we opted for a lexical substitution task. Specifically, we test whether the model consistently chooses the correct target domain paraphrase (therefore, fully interpreting the metaphor) or rather bases its decisions on lexical similarity and chooses the inapt paraphrase that is similar in meaning to the literal use of the metaphorical word.

We set up a fill-in-the-blank task to crowdsource apt paraphrases, and manually selected the best paraphrases using expert knowledge. We also manually created inapt paraphrases from WordNet synsets, so that the apt and inapt paraphrases reflect the target and source domains of the metaphors respectively. Specifically, the inapt paraphrases are synonyms or hypernyms associated with the word's literal use (the source domain).

Using this dataset, we tested the metaphor understanding capabilities of LLaMA-13B, 30B, and GPT-3.5 zero-shot in two tasks: paraphrase judgement and paraphrase generation, as illustrated in Figure 3.1. Our results show that both tasks are challenging for the models. In particular, the models are prone to confuse the target and source domains of the metaphors, as they often fail to distinguish the inapt paraphrases from the apt paraphrases or reference sentences. Our experiments also reveal that LLMs' metaphor understanding capabilities are associated with genre, metaphor novelty, and POS of the metaphorical word. The MUNCH tasks thus allow us to gain insight into how LLMs process metaphors as well as how this remarkable ability can be improved in the future.

3.2 Related work

Steen et al. (2010b) created the VUA corpus, which marks out MRWs (Section 2.1.2) in a 4-million-word subset of the British National Corpus. It has been widely used in studies on automated metaphor detection (Choi et al., 2021; Leong et al., 2018, 2020; Li et al., 2023; Zhang and Liu, 2022). However, the corpus does not specify the conceptual metaphors indicated by the MRWs or provide annotation for interpreting the metaphors. The corpus is not directly applicable to automated metaphor interpretation.

Shutova (2010) defined automated metaphor interpretation as a paraphrasing task: Given a metaphorical expression where a word is marked as metaphorically used, the model should replace this word with another word to render a literal paraphrase of the expression. For example, the verb phrase *stir excitement*, where *stir* is used metaphorically, should be paraphrased as *provoke excitement*.

Bizzoni and Lappin (2018) created a Metaphor Paraphrase Evaluation Corpus (MPEC), which provides correct and incorrect paraphrases for ∼200 short sentences containing metaphor use; paraphrases could greatly differ from the reference sentences. Joseph et al. (2023) created the NewsMet dataset, which consists of 1k verbal metaphors in news headlines as well as their literal equivalents; incorrect paraphrases are not provided.

Several recent studies approached metaphor understanding as an inference task. The IMPLI dataset (Stowe et al., 2022) includes entailed and non-entailed sentences for ∼900 metaphorical sentences. The FLUTE dataset (Chakrabarty et al., 2022b) provides entailment and contradiction pairs for 1500 metaphorical sentences (including 750 similes). Fine-tuned transformer-based models reached > 0.8 accuracies in these 2 studies in predicting the class of a given sentence pair.

Recent studies also employed multiple-choice and generative tasks to assess LLMs' ability to reason with metaphorical language. The MiQA benchmark (Comșa et al., 2022) uses such tasks to test whether models can distinguish metaphorical and literal uses of the same words; 150 conventional metaphors are involved. The Fig-QA task (Liu et al., 2022b) includes 10k similes (a type of direct metaphor) and requires models to distinguish a pair of metaphors of opposite meanings. Chakrabarty et al. (2022a) examined LLMs' figurative language understanding by asking them to generate text after encountering an idiom or simile.

The MUNCH dataset provides 3k samples of indirect metaphors, 10k correct paraphrases, and 1.5k incorrect paraphrases. It is therefore one of the largest datasets for paraphrasing of indirect metaphors. The candidate paraphrases are also systematically different from the ones in previous datasets, as we tailored the dataset for testing LLMs' understanding of metaphors as cross-domain mappings and correctly capturing the underlying relational structures. We summarize differences between MUNCH and previous datasets and provide more details for the latter in Appendix A.1.

3.3 Data collection: metaphor samples

The metaphor samples in our dataset were selected from the VUA corpus. Each metaphor sample is a sentence containing a highlighted MRW, the metaphorical word to be interpreted; a paraphrase uses a single word to substitute the metaphorical word. We use two criteria for selecting metaphorical sentences: novelty of the metaphor and possibility of single-word substitution. We explicate our selection process below.

The novelty criterion. We employed novelty scores from Do Dinh et al. (2018) to increase the proportion of novel metaphors in our dataset. Scores range from -1 to 1. VUA contains a large proportion of conventional metaphors: The metaphorical use of the word can be found in a dictionary of contemporary language use (Steen et al., 2010a). As LLMs might have encountered enough data for such conventional metaphor uses during pre-training, the understanding of such metaphors should be relatively easy. To render a more challenging dataset, we excluded MRWs with novelty scores below -0.3. Metaphors with a novelty score higher than -0.3 could still be conventional: The crowd workers who provided the novelty annotations in Do Dinh et al. (2018) relied on their intuition instead of a dictionary like Steen et al. (2010a). And metaphorical uses included in dictionaries may still be considered novel by lay people. We chose -0.3 as the threshold in order to collect a large and diverse dataset as a starting point.

The single-word criterion. To ensure that the metaphorical sentences can be paraphrased via single-word substitution, we excluded MRWs that are marked as direct metaphors, as well as a portion of indirect metaphors. Directly used MRWs usually occur in a sequence, such as *I knew the pathway like the <u>back of my hand</u>*. They are thus not suitable for single-word substitution. Also, the direct metaphor *back of my hand* refers literally to the back of the speaker's hand—its contextual meaning is directly associated with the source domain. This is contrary to our task setup, where apt paraphrases (contextual meaning) should be associated with target domains. We therefore opted to focus on indirect metaphors in this study.

Within the category of indirect metaphors, we filtered out new-formations, consecutive MRWs, and proper names. New-formations are words that do not have an entry in dictionary, so VUA annotated the parts that do have corresponding entries. For example, in the phrase *a rose-tinted vision of the world*, the word *rose-tinted* was a new-formation. Consequently, *rose* and *tinted* are marked as separate MRWs in VUA and received their separate novelty scores in the study of Do Dinh et al. (2018). We filtered these out because a single metaphorical word should have a single novelty score (*rose-tinted* has two), yet it is hard to paraphrase *rose* or *tinted* instead of *rose-tinted* altogether.

Likewise, we excluded cases where multiple content words marked as indirect metaphors occur consecutively, such as *take place*, *long road home*, *great leap forward*. These often involve fixed expressions or phrases that either should be replaced as a whole or should not be marked as consecutive indirect metaphors. We also excluded metaphorical words that are part of a proper name, which, like fixed expressions, need to be treated as a whole. For example, the proper name *Nord Stream* would lose its meaning if one changed the metaphorical word *stream* into another word.

3.4 Annotation of apt paraphrases

Crowdsourcing task. We constructed a fill-in-the-blank task to crowdsource (apt) paraphrases for the metaphorical sentences. Each task included 30 sentences to be paraphrased, so that the task can be finished within 30 minutes. Under each sentence, the workers were presented with a copy of the sentence where the metaphorical word is replaced with a blank; they were asked to fill the blank with a single word so that the new sentence is a semantically and grammatically apt paraphrase of the reference sentence. If they were not able to paraphrase the sentence, they were asked to explain why it was difficult. Examples of good and bad answers were provided in the instructions (see Appendix A.2).

The workers were recruited via Prolific. We set prescreening criteria to only include adult (age $>$ 18) native English speakers who were living in an English-speaking country and did not have any language-related disorders. The workers were asked to confirm within the task that they met these criteria. After giving consent to participate and reading the instructions, they were also required to correctly paraphrase a trial sentence in order to access the task. More details (worker's consent, the trial sentence) are given in Appendix A.2.

We released 99 tasks in total and collected 5 data points for each of the 2970 reference sentences. We received single-word substitutions for 2953 sentences (the other 17 are presented and explained in Appendix A.2), and 61% of them got repeated answers—multiple workers submit the same paraphrase despite the question being open-ended. This confirms the reliability of our task.

Expert validation. For a selection of the reference sentences for which we later annotated inapt paraphrases (Section 3.5), we further validated the crowd-sourced paraphrases to determine the best paraphrase for creating the triples (one metaphor sample, two candidate paraphrases).

We used both majority vote and expert knowledge to find one best paraphrase for each reference sentence. First, we ranked the single-word substitutions collected for each reference sentence in order of popularity (i.e., how many participants proposed each substitution). We then went through the list and verified each option in terms of whether it would result in an apt paraphrase of the reference sentence. Among the verified paraphrases, we selected the one that was proposed by the largest number of participants as the best paraphrase for that reference sentence.

When multiple apt paraphrases have the same number of votes, we chose the one that is clearly within the target domain—that is, the paraphrase clearly shows that the metaphorical word is interpreted in its contextual sense. For instance, we received 5 different single-word substitutions for the metaphorical word *attack* in Sentence (1). These are *remarks*, *views*, *offense*, *incursion*, and *disagreement*, each proposed by a single participant. All of them can be considered apt paraphrases. We chose *remarks* because it clearly shows the metaphorical word *attack* is interpreted in the ARGUMENT domain. The meaning of *offense* and *disagreement* are more abstract and could involve other conceptual domains; the paraphrases that replace *attack* with *views* and *incursion* respectively are still metaphorical, as *view* can be associated with VISION and *incursion*, like the metaphorical word itself, is still in the domain of BATTLE. These four are thus less preferable with regard to the purpose of our dataset.

(1) They have also aroused Protestant anger against Dr Runcie, at the same time as he has become involved in a row over his *attack* on the "Pharisees" of British society. (VUA)

While we managed to find one best paraphrase for most reference sentences, there are 45 for which we selected two paraphrases as the best, as the two received the same votes and are equally apt. There are also 11 sentences for which no paraphrase was selected. These are cases where the given context is insufficient for determining the contextual meaning of the metaphorical word.

3.5 Annotation of inapt paraphrases

Tong (2021) shows that incorrect paraphrases based on the basic sense of the metaphorical word are the least distinguishable from correct ones (i.e., paraphrases based on the contextual sense) with respect to aptness. To render truly challenging inapt paraphrases for our task, we therefore created inapt paraphrases exclusively from basic senses.

We employed WordNet for identifying basic senses and obtaining sense-specific synonyms, following the annotation guidelines presented in Appendix A.3. For each metaphorical word, we first locate the WordNet synsets that correspond to its more basic meaning (relative to its contextual meaning in the reference sentence). Then we go through the synonyms (or hypernyms when no synonyms are provided) under the basic-sense synsets and select those that are clearly associated with the metaphor's source domain and would render a grammatical (but inapt) paraphrase.

We went through all 2970 sentences released for the crowdsourcing task and

	ACPROSE	NEWS	FICTION	CONVRSN	TOTAL
	1061	922	593	377	2953
N	50%	38%	35%	25%	40%
V	35%	42%	39%	51%	40%
A	15%	20%	26%	24%	20%

Table 3.2: Number of metaphor samples per genre (academic, news, fiction, conversation), and the percentage of sentences where the metaphorical word is a noun (N), a verb (V), or either an adjective or an adverb (A).

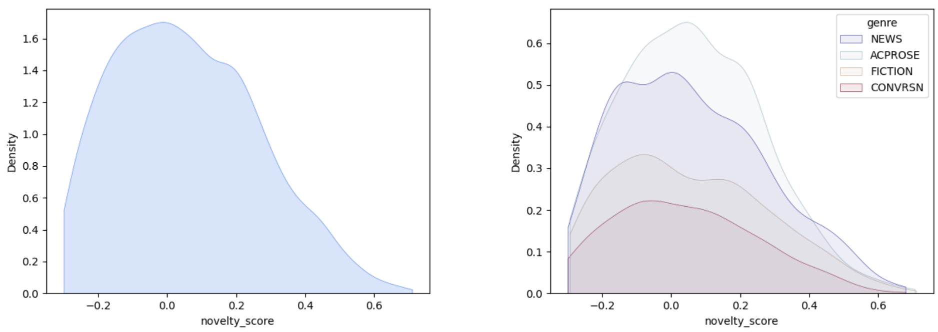


Figure 3.2: Novelty distribution of the metaphor samples, across all genres (left) and in different genres (right). The novelty scores are extracted from Do Dinh et al. (2018).

found inapt paraphrases for 991 of them. After removing items lacking apt paraphrases (either because no single-word substitutions were crowdsourced or because none of the collected ones are of sufficient quality), we created 1492 triples for 728 metaphorical sentences, including 1072 triples with an apt and an inapt paraphrase, 375 triples with two inapt paraphrases, and 45 with two apt paraphrases.

Inter-annotator agreement. From the 991 sentences for which inapt paraphrases were identified, we randomly selected 200 to be annotated by a second annotator. The annotator was a PhD candidate in linguistics specializing in metaphor research. We explained the annotation process to the second annotator through a meeting and the guidelines in Appendix A.3. The Gwet's gamma coefficient for the agreement between the two expert annotators is 0.84.

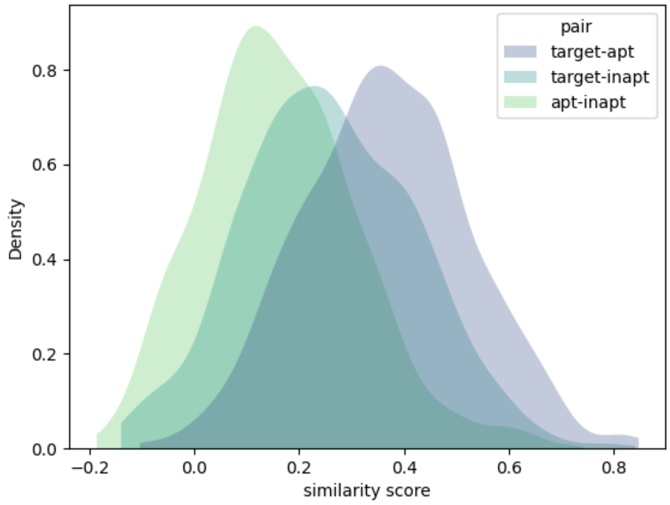


Figure 3.3: Distribution of the cosine similarity between target-apt, target-inapt, and apt-inapt pairs.

3.6 Data analysis

The dataset contains approximately the same number of samples from academic and news genres, and fewer samples from fiction and conversation, as shown in Table 3.2. These metaphor samples cover metaphorical use of content words in all four parts of speech. Noun and verb MRWs are of a higher proportion compared to adjectives and adverbs. In news and fiction, these two categories have similar percentages. The academic genre contains more noun MRWs than verbs whereas in conversation the situation is reversed: Half of the metaphorical words are verbs, while the percentage of nouns is similar to that of adjectives and adverbs.

As we excluded MRWs of novelty scores lower than -0.3, the metaphor samples exhibit a wider range of novelty scores above 0 than below 0 (Figure 3.2). Meanwhile, a large proportion of the metaphor samples could be considered only slightly novel or conventional (novelty scores between -0.3 and 0.3). Metaphor samples of the highest novelty scores can be from any of the four genres. Despite their different proportion in the entire dataset (Table 3.2), all four genres include metaphor samples across all levels of perceived novelty.

We also examined the cosine similarity between the metaphorical words and apt and inapt substitutions. Since the inapt paraphrases were based on the more basic meaning of the MRWs (Section 3.5), we expected inapt substitution words to be more similar to the metaphorical words than apt substitution words. We computed the cosine similarity scores using `glove-wiki-gigaword-300` embeddings (Pennington et al., 2014), accessed through the `gensim` Python library. Figure 3.3 shows the distribution of cosine similarity scores for 1006 triples, ex-

Select words that can replace the highlighted word in the given sentence without altering the sentence's meaning.
Sentence: ... *extending* the Government's borrowing power ...
Option A: increasing
Option B: exserting
Option C: Both Option A and Option B
Option D: Neither Option A nor Option B
Correct answer: Option

Figure 3.4: Example prompt for the *Word-judgement* task (the *Implicit* condition). The given sentence is shortened for illustration.

cluding the ones containing out-of-vocabulary words. Surprisingly, the target-apt pairs tend to have higher cosine similarity scores than the target-inapt pairs. The plot suggests that the 3 pairs are distinguishable in terms of cosine similarity scores, with target-apt pairs being the most similar, and apt-inapt the least similar. This might be associated with the fact that our metaphorical sentences were sampled from VUA, which, being representative of metaphor use in natural discourse, includes a large percentage of conventional metaphors. Nonetheless, the majority of the cosine similarity scores are above 0, and the 3 pairs still share a wide range of similarity scores. The distribution plot is therefore also suggestive of the reliability of our dataset, as well as its potential challenge for LLMs.

3.7 Model evaluation

We evaluated LLaMA-13B, LLaMA-30B, and GPT-3.5 (`text-davinci-003`) on two tasks: (Task 1) **paraphrase judgement**, which requires a model to select correct paraphrases for a given reference sentence from given candidates; and (Task 2) **paraphrase generation**, which asks a model to generate correct paraphrases for a given reference sentence. The paraphrase judgement task used the 1492 triples that include inapt paraphrases; the generation task used all 2953 metaphorical sentences. Details regarding computational budget is given in Appendix A.4.

3.7.1 Paraphrase judgement

We evaluate the LLMs in a prompting setup. We test the models' ability to interpret metaphor under different conditions. In the first scenario, we prompt the model by providing the reference sentence with the metaphorical word highlighted and two candidate replacement words for it (*Word-judgement*). In the second scenario, each of the candidate replacement words is embedded in the sen-

	LLaMA-13B	LLaMA-30B	GPT-3.5
Word-judge			
Implicit	0.28 (0.18)	0.21 (0.10)	0.23 (0.10)
M-Sent	0.30 (0.16)	0.19 (0.09)	0.20 (0.10)
M-Word	0.33 (0.18)	0.21 (0.08)	0.20 (0.08)
Sent-judge			
Implicit	0.13 (0.06)	0.14 (0.03)	0.17 (0.07)
M-Sent	0.12 (0.07)	0.17 (0.03)	0.16 (0.06)
M-Word	0.10 (0.08)	0.27 (0.05)	0.21 (0.02)

Table 3.3: Mean (SD) accuracies across 3 prompts for each paraphrase judgement condition.

tence (*Sentence-judgement*). In both cases the model needs to solve a multiple choice task. Besides providing the apt and inapt paraphrases (Options A and B) as answer options, we also complement them with Option C, that both candidates are correct, and Option D, that neither are correct. See Figure 3.4 for an example. We expect *Word-judgement* to be more challenging, as the model would need an additional inference step compared to sentence judgement, to replace the metaphorical word with the two given options and (implicitly) form the intended paraphrases.

For both *Word-judgement* and *Sentence-judgement* setups, we further investigate whether it makes a difference if the model is explicitly "told" that the task is to paraphrase a metaphor or not. This results in three further conditions: *Implicit* (not mentioning metaphor in the prompt), *Metaphor-Sent* (revealing that the reference sentence contains a metaphor), and *Metaphor-Word* (revealing that the specific highlighted word in the sentence is metaphorically used). The *Implicit* condition corresponds best to the real-life application of LLMs, where the model needs to be able to interpret metaphors without being instructed that metaphors are there.

We tested LLaMA-13B and 30B, and GPT-3.5 in each of the 6 conditions, using 3 prompts for each condition (the prompts are listed in Appendix A.4). Table 3.3 shows the mean accuracy and standard deviation for each model in each condition. The random baseline achieves an accuracy of 0.25, as there is always only one correct option out of the given four. The performance of all three models was below the random baseline in most cases, except for LLaMA-13B in the *Word-judgement* tasks and LLaMA-30B in the *Metaphor-Word* condition of the *Sentence-judgement* task. Meanwhile, the accuracy of LLaMA-13B varied a great deal across different prompts in the *Word-judgement* tasks.

The *Sentence-judgement* task seems to be more challenging than *Word-judgement* for the models. For LLaMA-30B and GPT-3.5, the task was particularly

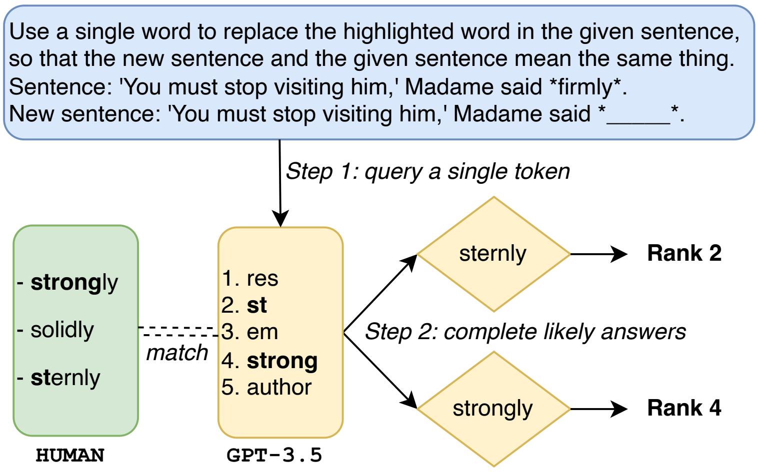


Figure 3.5: Procedure of the paraphrase generation task, using GPT-3.5 prompt and outputs as example. We first ask the model to generate a single token to get a glimpse of its top 5 answers. For each token that matches the beginning of a human answer, we let the model complete it to see whether it is a complete match.

difficult when they were not instructed to focus on the metaphorical word, and were not informed that the word is metaphorically used (the *Metaphor-Word* condition). For LLaMA-13B, all three *Sentence-judgement* conditions are similarly difficult. However, its higher accuracies in the *Word-judgement* tasks also indicate the benefit of instructing the model to focus on the metaphorical word.

3.7.2 Paraphrase generation

The purpose of this task is to compare model and human performance in paraphrasing metaphorically used words. The prompts were thus designed to be semantically close to the instructions in our crowdsourcing task (Section 3.4). The model answers were generated in two steps (see Figure 3.5). We first let the models generate a single token—this allowed us to access the models' ranking of all tokens in their vocabulary. Of these, we selected the ones that match human annotations and let the models complete them into words. The completions were then compared with human annotations to determine the rank of each expected answer.

We tested the models on 3 prompts (see Appendix A.4) and their mean performance in terms of mean reciprocal rank (MRR), recall at top 5 paraphrases and recall at top 10 paraphrases is shown in Table 3.4. GPT-3.5 performed best and LLaMA-30B came second. The models' performance was also more stable across different prompts as compared to the paraphrase judgement task. Nonetheless, all three models clearly preferred different answers as compared to human

	MRR	Recall@5	Recall@10
LLaMA-13B	0.33 (0.02)	0.22 (0.02)	0.33 (0.02)
LLaMA-30B	0.47 (0.03)	0.28 (0.02)	0.40 (0.03)
GPT-3.5	0.54 (0.02)	0.32 (0.01)	N/A

Table 3.4: Mean (SD) performance across 3 prompts in the paraphrase generation task. Recall@10 does not apply to GPT-3.5 as the OpenAI API only allows access to the top-5 answers.

annotators.

3.8 Discussion

Paraphrase judgement. We looked into the type of errors the models made in paraphrase judgement. The number of each combination of expected and predicted answers for each model is in Figure 3.6. We found that LLaMA-30B and GPT-3.5 could ignore the semantic differences between a given sentence and an inapt paraphrase, as they tend to predict both candidates as correct when presented with one or more inapt paraphrases. LLaMA-13B, on the other hand, tends to assume that the two given candidates always contain one apt and one inapt paraphrase. Nonetheless, it did not seem capable of distinguishing the two, as it made a similar number of Option A and Option B predictions.

Paraphrase generation. We examined the top-ranked answers of the models and found 4 categories that the "incorrect" or unexpected answers could fall into. **1) Nonsensical:** For the sentence "...for this number line I would say ...", GPT-3.5 gives *thus* as the best substitution word, ignoring the meaning of *line*, whereas LLaMA-13B repeats *number*. **2) Lack of contextual understanding:** In "...he touched both sides of the coin ...", GPT-3.5 replaces the word *sides* with *facets*, suggesting that it neglects details of the meaning of the sentence (that a coin only has 2 sides). **3) Ungrammatical:** On the other hand, the model may have understood the metaphor, but fails to convert its understanding into a suitable substitution word. In "...they all shared the emphasis on 'her' ...", LLaMA-30B suggests *concurred* as the best answer, implying that the meaning of *shared* is understood, but that grammatical agreement has been sacrificed. **4) Preference:** Finally, the disagreement between the models and human annotators may simply be a matter of preference. For "For a man whom Rebecca West ...called 'repulsive' and 'treacherous' ...", crowd workers provide 4 possible answers: *revolting*, *disgusting*, *awful*, and *grotesque*. Both LLaMA-30B and GPT-3.5 give *odious* as the best answer. Here, both the human annotators and

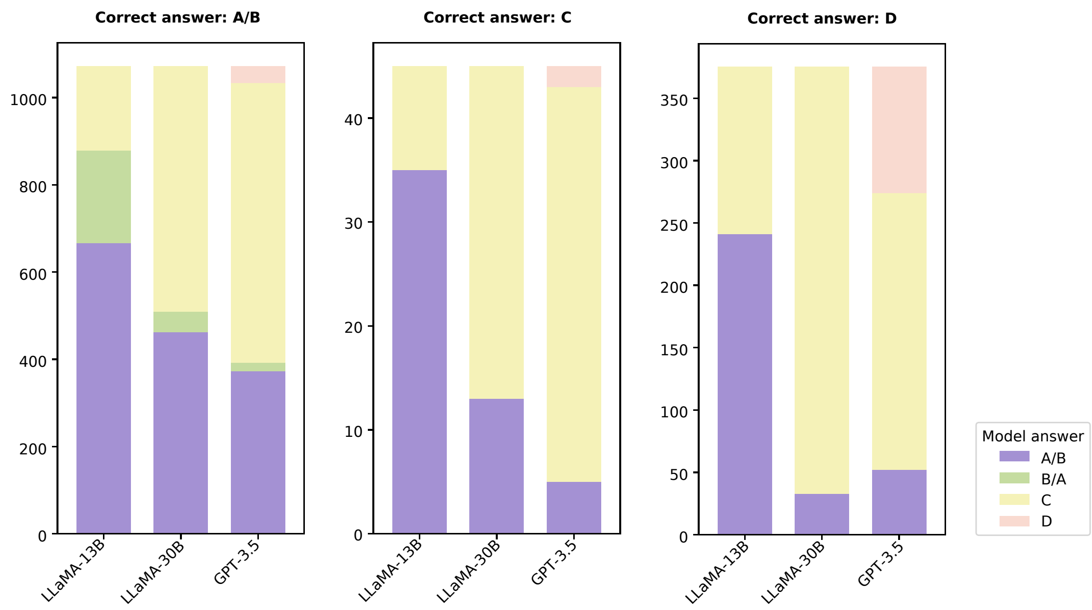


Figure 3.6: Number of each combination of correct answer and model prediction. The counts are based on the predictions of the models when they reach their respective highest accuracy in our experiments.

the models understand and can paraphrase the sentence, and it is hard to say whose answer is best.

Factors associated with model performance. We also examined the association between model performance and 3 factors: genre, metaphor novelty, and the POS of the metaphorical word. The details are available in Appendix A.5. We found metaphors of higher novelty scores to be more difficult for LLaMA-30B in paraphrase judgement, and for GPT-3.5 in paraphrase generation. The association between genre or POS and model performance tends to differ per model and task. The fiction genre, for example, is the easiest for the LLaMA models in paraphrase generation; yet it is the most difficult for GPT-3.5 in the generation task and for LLaMA-30B in the judgement task. Similarly, noun metaphors are the easiest for LLaMA-30B in the generation task and for GPT-3.5 in the judgement task. Meanwhile for GPT-3.5 in the generation task, adverb metaphors become the easiest.

To sum up, the results of the two paraphrase tasks indicate that the LLMs are unable to (fully) understand some of the metaphors in our dataset. The paraphrase judgement task further reveals that the models have difficulty distinguishing the metaphors' source domains (implied by the inapt paraphrases) and target domains (implied by the reference sentences and apt paraphrases). This further suggests that the models are unlikely to perform reasoning across semantic domains; when they succeed in understanding the metaphor, they may still reason in ways that are different from humans. This means that for downstream NLP tasks such as opinion mining, bias detection, humour detection, and intent recognition, the LLMs could overlook the entailment of a metaphor. In machine translation as well as summarization of highly figurative or poetic texts, the problems may manifest as incorrect or peculiar explanation of metaphors.

A direction for improvement is to mark out metaphor uses in texts and direct the model's attention to them: In the paraphrase judgement task, the models reach higher accuracies when the metaphorical word is marked out (in the *Word-judgement* task or in a *Metaphor-Word* condition). However, since the models generally performed poorly in the experiments, the LLMs may need to be fine-tuned in order to better understand metaphors. When fine-tuning, one can consider increasing the proportion of certain metaphor types in training data, as genre, metaphor novelty, and POS of the metaphorical word are all associated with model performance. Future studies could first employ MUNCH to detect the weak points of an LLM and then curate training data accordingly.

3.9 Conclusion

We release a dataset of manually created apt and inapt paraphrases for metaphorical sentences and present two metaphor understanding tasks, which we demonstrate to be challenging for current LLMs. The errors the models make in the paraphrase generation task indicate various levels of misunderstanding of the metaphors. In the paraphrase judgement task, the models' accuracy was lower than the random baseline in the majority of the cases; a closer look at their errors reveals that the models had difficulty in detecting the inaptness of the inapt paraphrases. The experiments also show that the models performed better when being instructed to focus on the metaphorical word, and that genre and the POS and novelty of the metaphorical word are all potential factors that affect model performance.

3.10 Limitations

We designed the metaphor understanding tasks to be representative of a lexical substitution task: The metaphorical word is the only difference between a reference sentence and a candidate paraphrase. This setup makes it possible to examine whether LLMs indeed perform metaphor interpretation or resort to lexical similarity when they encounter metaphorically used words. At the same time, however, it also means that the models' understanding of multi-word metaphors and direct metaphors (e.g., similes) could not be tested using our dataset.

We tested the latest and state-of-the-art LLMs available at the time that our study was ongoing, but newer LLMs such as GPT-4 and Llama 2 emerged shortly after the completion of our study. We suggest running a data contamination test before evaluating newer LLMs using our dataset.

This study reveals that LLMs have difficulty distinguishing the target and source domains of linguistic metaphors. A more extensive analysis is desirable to uncover more differences between LLMs and humans in terms of metaphor interpretation.

3.11 Ethics statement

The metaphor resources used in this study are publicly available. The crowdsourcing task was approved by an ethics committee. The crowd workers received fair payment (9 GBP per hour), and no personal information was collected or stored in our database.

The metaphor samples in our dataset come from excerpts of natural discourse. They may therefore involve bias, taboo, violence, or other aspects of everyday language use that could be considered negative (we also pointed this out to the crowd workers before they gave their consent to participate, as presented in Appendix A.2). Nonetheless, these are integral parts of language use, and should be properly understood by NLP systems, which is precisely what this study aims at.

Acknowledgements

The crowdsourcing task (for collecting apt paraphrases) was funded by the Faculty of Engineering at the University of Bristol, through the faculty's Pump-Priming Fund. We thank Dr. Jie Fu for co-annotating the inapt paraphrases. We also thank all anonymous reviewers who provided constructive feedback for previous versions of this report.

Chapter 4

A framework for annotating and modeling intentions behind metaphor use

4.1 Introduction

The meaning of a metaphorical word always has to be considered within the context. Similarly, adequate understanding of metaphor use always has to take into account what the metaphor is used for. The literature relating metaphor and intention is rich but generally fragmented. With some exceptions (Roberts and Kreuz, 1994), metaphor scholars tend to focus only on isolated intentions. Hence, there is still a lack of a systematic and comprehensive account of intentions behind metaphor use and an operationalized framework enabling annotation of such intentions in linguistic data.

In this chapter, we fill in this gap by systematizing the existing literature on metaphor and intention, and proposing a first-of-a-kind unified taxonomy of intentions behind metaphor use. We further propose an annotation procedure and release a first dataset annotated for intentions behind metaphor use. We show that the proposed taxonomy is thus suitable for annotating metaphors in unrestricted text. We make our dataset publicly available.[1]

Using our dataset, we tested GPT-4 Turbo and two Llama2-Chat models (the 13B and 70B versions) on their ability to infer the intentions behind metaphor use. The task required the models to select one category from the taxonomy for a given metaphorical expression in a sentence. The best-performing model, GPT-4, reached an average accuracy of 43.30% in the zero-shot setting and a slightly higher accuracy of 45.09% in the five-shot setting, demonstrating that inferring the intentions behind metaphor use is a challenging task for state-of-the-art LLMs.

I co-developed the taxonomy, co-annotated the dataset, and conducted the computational experiments and error analysis independently.

[1] https://github.com/GMichelli/intentions-behind-metaphor

4.2 Related work

Conceptual and deliberate metaphor. In a seminal paper, Ortony emphasized the necessary role of metaphor in everyday language (Ortony, 1975). Proponents of CMT reinforced this idea, highlighting that our own conceptual system is, at least partly, metaphorically structured (Lakoff and Johnson, 1980b). Abstract concepts, e.g., emotions like love, are understood through various kinds of conceptual metaphors.

While CMT revealed the pervasive nature of metaphor in human cognition, several authors emphasized the importance of communicative aspects in the analysis of metaphors. In particular, Steen (2008) stressed the significance of discerning deliberate metaphors from non-deliberate ones. Deliberate Metaphor Theory (DMT) departs from CMT by recognizing that only metaphors intentionally used *as* metaphors involve online cross-domain mappings (Steen, 2023, 2017), and non-deliberate metaphors can be processed differently—by lexical disambiguation. DMT faced criticisms, however. For instance, Gibbs (2011) highlighted the difficulty of identifying deliberate metaphors without specific linguistic markers and the unreliability of producers' conscious judgments on their own intentions. In order to address these challenges, advocates of DMT developed the Deliberate Metaphor Identification Procedure (DMIP) and clarified the distinction between deliberate and conscious use of metaphors (Reijnierse et al., 2018; Steen, 2014).

Intentions in language use. The notion of communicative intention (CI) holds a central position in pragmatics. CI is the speaker's intention to convey non-natural meaning through their utterances (Grice, 1957). Subsequent research has shown that one can distinguish among different intentions, varying in nature—prior intention vs intention in action (Searle, 1983), temporal aspect—proximal vs prospective (Haugh and Jaszczolt, 2012), and social dimension—individual vs social (Ciaramidaro et al., 2007).

As stressed by Gibbs (1999), conceiving of intentions as individual mental states makes them opaque since agents are not always aware of the causes of their behavior. Intentions should be viewed as *social judgments* instead. Inspired by the philosophy of action proposed by Anscombe (1957), we conceive intentions here as features attributed to linguistic acts. More specifically, intentions are those *reasons* that speakers may provide once asked why they resorted to certain metaphors. They serve as interpretive tools for understanding human behavior broadly, and linguistic behavior in particular.

Intentions and metaphor. Although there is not a common notion of intention shared among all metaphor scholars, in the literature intentions are typically formalized as prior intentions, i.e., as representations in the speaker's mind of their goals. Roberts and Kreuz (1994) build a first taxonomy of intentions for various forms of figurative language, including metaphor. This taxonomy was developed through experiments where participants were asked to provide reasons for using each figure of speech. We believe that this study has some limitations. Around 20 participants were asked to provide intentions only for 10 metaphors each, and there is no information on the typological variation of the selected items[2]. In contrast, we propose a metaphor-specific taxonomy that draws on a larger and more varied set of linguistic data.

Previous work has partially explored the relation between metaphor and individual intentions. Researchers have observed how metaphors can convey emotions (Fainsilber and Ortony, 1987; Fussell and Moss, 2014), persuade (Sopory and Dillard, 2002; van Stee, 2018), contribute to argumentation (van Poppel, 2021; Wagemans, 2016), serve didactic purposes (Cameron, 2003), add humor (Attardo, 2015), and cultivate intimacy between interlocutors (Cohen, 1978; Goatly, 1997).

4.3 Taxonomy of intentions

To compile our taxonomy, we reviewed existing literature from psychological (Roberts and Kreuz, 1994) and linguistic perspectives (Goatly, 1997). While we drew inspiration from individual categories, no single taxonomy served as a basis for our own. An initial version was informally tested against the VUA dataset, ensuring each category was represented and that most instances in the corpus would fit naturally within the taxonomy. The iterative process of refinement through successive revisions eventually stabilized. Hence, the resulting taxonomy reflects a continuous exchange with real linguistic data.

We now introduce each intention category, motivating it through theoretical considerations, previous literature and examples from available material.

Lexicalized metaphor. These metaphors are associated with a plain communicative intention, and the utterance is judged as meant to convey just its propositional message. For lexicalized metaphors, the question of why a metaphor was preferred over a literal paraphrase does not arise in interpretation. In Cameron's words, the metaphoric expression is "just the way to say it (Cameron, 2003)".

(1) a. I *fell in* love.
b. Summer *bedding* is looking tired.

Sentence (1-a) is an example of how the language of emotions often relies on metaphors. This observation, already noted by Fainsilber and Ortony (1987), aligns with the idea that emotions may be conceptualized metaphorically, as maintained by CMT. Example (1-b), instead, shows how the language we use to talk about some activities tends to have its own metaphorical jargon. This is true

[2] We note that the authors presented participants with a comparativist definition of metaphor (i.e., metaphor as implicit comparison), resulting in potentially biased judgments.

for academic domains such as mathematics, physics and the like, but also for non-academic domains like sports or hobbies. Both examples are cases of lexicalized metaphors which constitute the most conventional way of talking about the target domain.

Artistic use of metaphor. These metaphors are used to attribute at once a whole set of features to the target domain. These features need not be clearly determined in advance. Ultimately, the intention is to stimulate the receiver's creative interpretation.

(2) a. It is the east, and Juliet is *the Sun*.
b. Fermi's *mantle* in physics had fallen on his young shoulders.

Some metaphors are not easily paraphrasable because they could be paraphrased in a number of different, yet equally valid, ways. The ambiguity of the metaphorical meaning can be inherent to the target domain of the metaphor or it can be related to the set of features that the metaphor attributes. At least in poetry and literature, interpreters tend to activate multiple mappings at once (Rasse et al., 2020) and ambiguity in interpretation is shown to correlate with aesthetic liking (Jacobs and Kinder, 2017).

Visualization. The utterer might resort to a metaphor whose source domain is easier to visualize than the TARGET. The goal is to prompt an intuitive mental representation of the latter.

(3) a. It was *like a very bright light was just shining outward*.
b. It would bounce up and down *like a yo-yo*.

Metaphors often hinge on a highly concrete/imaginable source domain to address an abstract target domain[3]. This is particularly true for subjective feelings, as in Example (3-a). Fussell and Moss (2014) provide evidence for the ability of metaphors to express precise emotional states. More recently, Broadwell et al. (2013) developed a prototype model for automated metaphor identification partly based on imageability.

Some metaphors do not constitute mappings from the concrete to the abstract, but just from the familiar to the unfamiliar, like Example (3-b). As already stressed by Ortony (1975), metaphoric expressions are often perceived as more vivid than their literal paraphrases. Thus, they can foster the formation of a more insightful mental image. Vivid metaphors can be instrumental not only for descriptive purposes. As reported in (Cameron, 2003), they can also be used to

[3]In psycholinguistics literature, imageability refers to the property of words to easily evoke a mental image of their meaning (Paivio et al., 1968). Imageability and concreteness, thought positively correlated, might be two distinct constructs (Dellantonio et al., 2014; Gargett and Barnden, 2015).

express more clearly some commands (c.f., a PE teacher explaining their pupils how to perform a dance: *you are spokes in a wheel*).

Persuasiveness. Using a metaphor to refer to the target domain—in a political speech, for instance—the author can give it a non-neutral connotation. This connotation is not motivated by explicit arguments. The intention is for the audience to adopt the utterer's perspective or stance towards the target domain.

(4) a. The islamic *wave.*
b. This *slender* and *anaemic* first novel by a notable poet.

As already stressed by Lakoff and Johnson (1980b), metaphors generally highlight some aspects of the target domain, while at the same time hiding others. This process of highlighting and hiding causes a framing effect on the receiver, whereby the target domain is seen, as it were, through the distorting lens of the source domain. The availability of several experiments and of meta-studies (Sopory and Dillard, 2002; van Stee, 2018) makes the Persuasiveness category one that is most supported empirically.

Explanation. This type of metaphors are used for didactic purposes. The intention is to explain a new or already familiar concept to the addressee. There is some knowledge asymmetry in the discourse from specialists to non-specialists, e.g., from teacher to students.

(5) a. The atmosphere is *the blanket* of gases that surrounds the earth.
b. When the neutron falls apart, *spits out* an electron, it becomes a proton.

The clarifying effect of metaphor has been recognized in the existing study of intentions behind it by Roberts and Kreuz (1994). The role metaphors play in educational settings—viz., in primary education—has been analyzed in detail by Cameron (2003). Moreover, there is some empirical evidence for the usefulness of certain (deliberate) metaphors in undergraduate lectures (Beger and Jäkel, 2015). However, the use of metaphors in education does not go without risks of blocking further understanding, as highlighted by Spiro et al. (1989).

Argumentative metaphor. These metaphors are part of explicit arguments intended by the author to convince the audience of a certain claim. The intention is to make the argument more compelling.

(6) a. But the villages are dying, becoming suburbs or *dormitories* where few people work but many sleep.
b. If so, it will be a gamble, because he *flopped on* his only previous international appearance in Saudi Arabia.

As pointed out by van Poppel (2021), among others, argumentative metaphors can be used to make an effective statement, either as a standpoint or as a starting point (premise) for an argument. Moreover, they can also actively contribute to the flow of argumentation, like Example (6).

Social interaction. These metaphors focus on interpersonal relations, group or cultural conventions. The aim is to create or reinforce a bond between producer and receiver.

(7) a. *Sleepy* Joe, *Crooked* Hillary.
b. She *passed away.*

A metaphor can bring closer its maker and appreciators in a number of different ways. First, it can exploit the fact that they belong to the same group—e.g., Trump's supporters, as in Example (7-a). In such cases, a social metaphor is used to isolate the desired receiver from the general public (Cohen, 1978), thus reinforcing the in-group/out-group dynamic. Second, metaphor can be used to conceal a TARGET that is experienced as negative. If they understand this, the receiver becomes aware of the additional care put by the producer in their utterance. The shared awareness fosters intimacy building between the pair and stimulates empathetic effects, like Example (7-b).

Humor. The intention is to entertain the addressee, to be funny. Metaphoric language is exploited for its divertive effects, which would fade in literal paraphrases.

(8) a. I'm *a doormat in the world of boots.*
b. You walked into what I would call *a cupboard* but they classed it as the bathroom.

Language is not only used to communicate. Among the many and varied uses of language, there is also the one of entertaining others, and being entertained in return. Steen (2008) and Steen (2014) cite typical cases of humorous metaphors: sports newspaper headers, jokes, riddles and so on. The incongruity theories of humor offers a possible explanation for the divertive potential of certain metaphors (Dynel, 2009; Oring, 2003).

Heuristic reasoning. The intention is to provide an interpretative model for a theory, an artwork, etc., typically an abstract domain which is otherwise difficult to structure and conceive of. The metaphoric expression is used to organize the addressee's conceptualization of the target domain, based on their prior knowledge about the source domain. The discourse generally remains among specialists.

(9) a. A gas is *like a collection of billiard balls in random motion.*

b. It is her body *as the canvas*, her appearance *as art*.

Metaphor is a matter of seeing something as something else, that is, of interpreting things from a certain perspective. In cognitive terms, we map the source domain to the target domain in order to better understand it. Thus, a primary intention of metaphor, especially within academic contexts, is to provide an interpretation for the products of science, like Example (9-a), as illustrated by Hesse (1966), or of art, like Example (9-b), and literature (Ricœur, 1975).

4.4 Data collection and annotation

Collecting the data. In order to empirically test the proposed taxonomy, we collected and annotated data ($\sim$ 1.2k metaphors) from the VU Amsterdam Metaphor Corpus (VUA; Steen et al., 2010b)[4]. This freely-accessible corpus was chosen since it contains fine-grained metaphoricity annotations at word level; it includes different genres; it contains metaphors in different grammatical constructions; and it has been extended in subsequent work with other relevant annotations, such as metaphor novelty scores (Do Dinh et al., 2018). Metaphor-related words (MRWs) in the VUA are identified following the MIPVU identification procedure (Steen et al., 2010b). The core idea behind the procedure is the distinction between *contextual* and *basic* meaning of words. Text fragments are collected from the British National Corpus (BNC) Baby (The BNC Baby, 2005), a 4-million-words corpus of English language covering 4 registers (Academic, News, Fiction, Conversation). The VUA encodes multiple information at word level, including information on metaphor type, distinguishing among *direct* and *indirect* metaphors.

Direct metaphors are expressions whose dictionary meaning coincides with the contextual meaning. For example, the word *ferret* in the phrase *he's like a ferret* is a direct metaphor. Indirect metaphors, instead, are defined as expressions having a more basic dictionary meaning, differing from the contextual meaning. Cf., the use of *valuable* in the sentence *teachers do a valuable work.*

Our corpus consists of 1214 MRWs collected from the VUA. We annotated all unique instances of direct metaphors found in the corpus (301 MRWs), and a subset of indirect metaphors (913 MRWs). The VUA contains redundant instances of the same direct metaphor—several MRWs correspond, e.g., to the phrase *like a piñata above the teeming streets of the city.* However, for the purpose of annotating intentions the most natural unit of analysis is the phrase since the same intention is typically attributed to all MRWs in it. Thus, for each direct metaphor we assigned an intention only to one MRW. Annotators manually selected which word to annotate, based on their intuition of which lexical unit contributes the most to the metaphoricity of the phrase.

[4] http://www.vismet.org/metcor/about.html

To select a subset of indirect metaphors to annotate we used the novelty scores collected by Do Dinh et al. (2018). We divided all indirect metaphors into 5 bins according to their novelty scores. We opted to focus only on the top two bins—MRWs with novelty scores in [1,0.6] or (0.6,0.2]—which correspond to the most novel metaphors. Our rationale was that more creative uses of metaphor would yield more interesting material for investigating intentions. Within these indirect metaphors, we annotated 913 MRWs.

Some further cases were excluded from the annotation of intentions: cases where there was *not sufficient context* to fully interpret the metaphor and assign an intention; cases of *idiomatic* use[5]; some highly conventionalized *interjections*. Instances marked as cases to be excluded were not considered in subsequent study phases. The final dataset comprises 988 MRWs, each annotated with at least one intention from the taxonomy.

Annotation procedure and guidelines. The annotation procedure consists of two key steps:

1. The annotator should distinguish lexicalized metaphors from other types of metaphors. If they perceive some intention behind the metaphor other than communication of information, then they shall move on to step 2.

2. The annotator is asked to assign up to three intentions to the metaphor under analysis. In order to complete the task, they are provided with a table listing the taxonomic categories, each with its description and some examples.

The full guidelines can be found in Appendix B.1. In the guidelines, we provide a detailed description of the sequential steps to be followed during annotation. We also work out at length an example of annotation performed following the guidelines.

Corpus annotation. The annotation was carried out by an author of this paper, who was a Master's student in logic and philosophy of language. In addition to the 9 intention categories in the taxonomy, we also include a "dummy category" to keep track of cases where an intention could not be attributed.

Inter-annotator reliability. Another author, a metaphor researcher, annotated a subset of the data (360 MRWs). This subset is representative of the whole annotated corpus and replicates its proportions between different metaphor types: direct metaphors, indirect metaphors with novelty score in [1-0.6] and in (0.6-0.2].

[5]Idiom differs from metaphor: while the former is a relatively fixed and stable expression within a linguistic community, metaphor is more productive and can show variation.

	Direct	Indirect [1,.6]	Indirect (.6,.2]	**Total**
Lexicalized metaphor	9	19	379	407
Artistic metaphor	19	13	43	75
Visualization	53	11	132	196
Persuasiveness	2	15	51	68
Explanation	9	3	30	42
Argumentative metaphor	4	7	48	59
Social interaction	5	2	26	33
Humour	12	10	28	50
Heuristic reasoning	16	3	39	58

Table 4.1: Distribution of intentions by metaphor type, with the most frequent instances boxed in red.

We calculate inter-annotator reliability for 301 of the 360 items, to which both annotators assign at least one intention category. Their agreement in terms of Krippendorff's α (Artstein and Poesio, 2008) is 0.77. More details about the metric used, as well as a discussion of the resulting IAA score, can be found in Appendix B.2.

4.5 Corpus analysis

We analyzed our corpus to shed some light on the relationship between intentions and metaphor type, genre and novelty. Only the first attributed intention was considered for data analysis since no other intention was selected in most cases (827/988). Distribution of intention categories in the whole corpus and per metaphor type is shown in Table 4.1, and further analysis of metaphor type can be found in Appendix B.3.

Genre. The genre of a discourse can offer clues about a metaphor's presumed intention, with certain intentions more likely to appear in some genres than others. This is suggested also by Steen (2008), who claims in passing that the function of a deliberate metaphor depends on the function of the discourse in which it is found. In the VUA, four intuitive tags provide information on the genre of each fragment: FICTION, NEWS, CONVRSN, ACPROSE.

In Figure 4.1, we report how individual intention categories (the vertical bars) are distributed over the four genres (the coloured parts in each bar). Our findings support the assertion that intentions behind metaphor use seem to correlate with the discourse genre in which the metaphor is found. For instance, Artistic metaphor and Visualization are found mostly in Fiction; Persuasiveness and Ar-

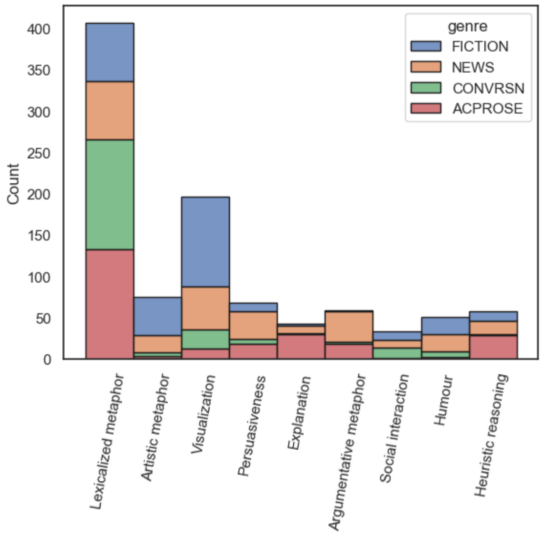


Figure 4.1: Distribution of intention categories per genre.

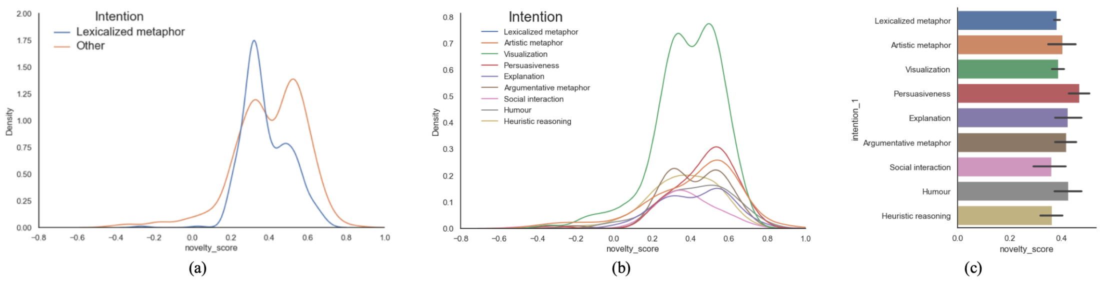


Figure 4.2: Distribution of novelty scores: (a) comparison between Lexicalized vs other metaphors, (b) individual distributions, and (c) mean novelty scores. Figures (a), (b) show probability densities.

gumentative metaphor in News; Explanation and Heuristic reasoning in Academic texts; Social interaction in Conversation. All of these results agree with what one would intuitively expect. However, reality is complex and suggests that drawing one-to-one correspondences would be too simplistic. In most cases, instances of the same intention are found in all four registers. Genre can thus help to track most common uses but not all uses.

Novelty score. Information on the novelty vs conventionality of metaphors is crucial for understanding how different intentions are reflected in different language choices. While certain intentions seem to correlate with highly conventional metaphors, others result in more original ones.

Figure 4.2(a) contrasts the distribution over Lexicalized metaphor (the blue line) vs all other intentions merged together (the orange line). In Figure 4.2(b), we zoom in and plot individual distributions. Each coloured line corresponds to the distribution of a single intention category. Finally, we have computed mean novelty scores per intention with confidence intervals, as shown in Figure 4.2(c).

In terms of novelty, metaphors with different perceived intentions show different degrees of conventionality. Taking into account average novelty scores and estimated distributions, categories such as Persuasiveness, Explanation, Humour and Artistic metaphor are generally more original, while Lexicalized metaphor, Social interaction and Heuristic reasoning are more conventional.

4.6 Evaluation of LLMs

We use our dataset to test GPT-4 Turbo (`gpt-4-0125-preview`; OpenAI et al., 2023), Llama2-13B-Chat, and Llama2-70B-Chat (Touvron et al., 2023) in terms of their ability to predict the intentions behind metaphor use (for details regarding model access, parameters, and computational budget, see Appendix B.4). The task requires the models to choose a single intention category from our taxonomy, given a highlighted metaphorical expression in a sentence.

We test the models in zero-shot, as well as five- and nine-shot in-context learning settings. In the zero-shot setting, a short explanation for each intention category is provided. We compute the average performance of the models across 3 different prompts, as shown in Appendix B.5.

The few-shot settings provide randomly selected examples for each test item. As there are nine categories in total, the nine-shot settings select one example for each category; in the five-shot settings, the category of the test item is always exemplified.

Since the in-context examples implicitly explain the intention categories, we test the models under two conditions in the few-shot settings: one provides explanations for the intention categories, just like the zero-shot setting (5/9-shot); the other removes those explanations from the prompt (5/9-shot-short). The latter setup tests whether the models can correctly infer what each intention category means from in-context examples.

Results. Table 4.2 shows the models' performance in these tasks in terms of accuracy. All three models reach accuracies that are above the random baseline in the zero-shot experiments. The accuracies are still relatively low, however, demonstrating that this is a challenging task for the LLMs.

The models reach their respective best performances under different conditions: GPT-4 in 5-shot (45.09%), Llama2-70B-Chat in 9-shot (39.00%), and Llama2-13B-Chat in 9-shot-short (30.88%). Llama2-13B-Chat is the only model whose accuracy increases with the number of in-context examples, albeit at the expense of success rates. The accuracy of Llama2-70B-Chat and GPT-4 drops in the 5-shot and 9-shot experiments respectively.

Model	0-shot	5-shot	9-shot	
Llama2-13B	24.88 (2.48)	27.16	29.10	
		23.76	30.88	ș
Llama2-70B	27.29 (5.54)	21.63	39.00	
		14.62	24.39	ș
GPT-4	43.30 (1.58)	45.09	39.00	
		41.61	34.42	ș
Random	13.01	20.00	13.01	

Table 4.2: Model accuracy (%) in zero- and few-shot settings, compared to random baseline. Zero-shot accuracy is averaged over 3 runs that use different prompts, with standard deviation in parentheses. The ș rows show 5-shot-short and 9-shot-short results, settings that remove intention category explanations from the prompts. Success rates under 100% are highlighted (90–94% , 95–99%).

In the few-shot settings, the models perform worse when explanations for the intention categories are removed from the prompts. The only exception is Llama2-13B-Chat in the 9-shot setting: There is a 1.78% increase in its accuracy when the explanations are removed. The result indicates that, overall, the models struggle to infer a correct characterization of the intention from the in-context examples.

Error analysis. Figure 4.3 shows the mean F_1 score for each intention category in the zero-shot experiments, averaged across the three prompts (for analysis of the few-shot experiments, see Figure B.3 in Appendix B.6). GPT-4 reaches the highest F_1 scores when it comes to Lexicalized metaphor and Visualization, closely followed by Llama2-70B-Chat with regard to Visualization. On the other hand, all three models show great difficulty in dealing with metaphors in the Social interaction and Heuristic reasoning categories, which are also the least represented categories in our dataset (Section 4.5). A possible explanation, therefore, is that the models encounter few of such data in the training phase.

Both GPT-4 and Llama2-70B-Chat mistake Lexicalized metaphor as Visualization. These concerns conventional metaphors whose TARGET domains pertain to visible objects or the action of seeing, e.g., *a glimpse of the impact of the 1980-1 riots*, and *channels of communication*.

These two models also mistake Visualization for other intentions, such as Artistic metaphor or Persuasiveness, e.g., *as enjoyable as feeling gently hungry or amorous*; *the wide sleeves of limp cotton hung from her freckled arms like rags thrown over a stick*. These errors can be attributed to LLMs' lack of embodied ex-

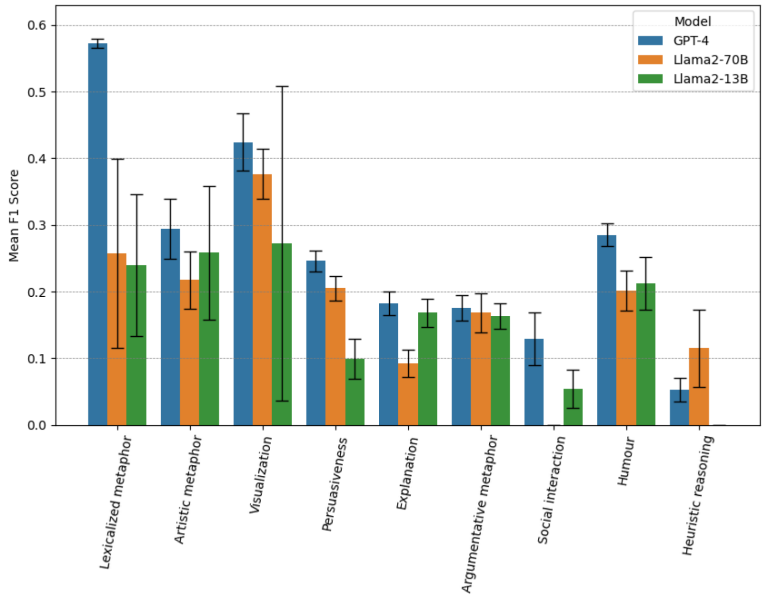


Figure 4.3: Model F_1 score in the zero-shot experiment, averaged across three prompts. Confidence intervals are computed with standard deviation across the prompts.

perience. These metaphors naturally evoke a mental image or sensory experience in humans. LLMs, on the other hand, do not automatically form representations of the meaning of a text in another modality.

4.7 Conclusion

We have gathered evidence from existing literature and incorporated it into a novel taxonomy of intentions commonly attributed to metaphor. The taxonomy can be used to annotate metaphors in unrestricted text, as demonstrated by our corpus annotation effort. Data collected from the VUA helped to better understand the nature of the different intentions and how these are realized in linguistic metaphors varying in their type, genre, and novelty score.

We have created and released a first dataset with metaphors annotated according to the taxonomy, Our experiments show that inferring intentions behind metaphor use is still a challenging task for current LLMs, proving that our dataset is a valuable resource for the community. As addressed in our error analysis, we anticipate future work that provides data for the less represented categories in our dataset, as well as employment of multimodal LLMs to tackle the issue of embodiment in metaphor processing.

4.8 Limitations

This study inevitably has some limitations; we discuss three of them here. First, the corpus used for the annotation, the VUA, contains mostly indirect metaphors, which are generally quite conventional. Adopting a corpus with more direct and novel metaphors would probably yield interesting results in terms of attributed intentions. However, such a corpus, comparable in size and range to the VUAMC, is missing. Second, while the output of the reliability study is encouraging, future work that includes more annotators could further validate the robustness of our annotation procedure. Third, the current experimental setup asks LLMs to select only one intention category per metaphor, whereas our annotation guidelines allow for up to three intention categories per metaphor. We opted for this experimental setup to make the task easier for the models. Nevertheless, we acknowledge that this choice does not reflect the complexity inherit to the analysis of metaphors in language use.

4.9 Ethical considerations

The metaphorical sentences annotated in this study are sampled from the VUA corpus, which is licensed under CC BY-SA 3.0 and suitable for research purposes. The two annotators are authors of the paper and volunteered to annotate the dataset.

Chapter 5

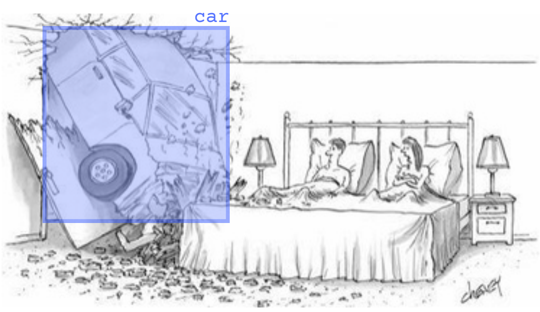
Hummus: A Dataset of Humorous Multimodal Metaphor Use

5.1 Introduction

Conceptual Metaphor Theory (CMT; Lakoff and Johnson, 1980b) contends that the metaphors we use in language are based on cross-domain mappings in our mind; metaphor is essentially a cognitive process of understanding one thing in terms of another. For example, when one says *Our marriage has gone off the track*, one is using the LOVE IS A JOURNEY metaphor, conceptualizing LOVE or RELATIONSHIP in terms of a JOURNEY. CMT marks a turning point in metaphor research, shifting focus towards how metaphors are represented or processed in the mind and the brain (see Holyoak and Stamenković, 2018 for a survey). There has also been an increased interest in nonverbal and multimodal manifestations of metaphor, such as metaphor in films, cartoons, adverts, and gestures (Forceville, 2015, 2017; Kappelhoff and Müller, 2011; Tsakona, 2009).

The influence of this cognitive turn in metaphor research has extended to the field of natural language processing. VUA, which is created by cognitive linguists and employs an annotation scheme that closely follows the CMT, has inspired research on automatic detection of metaphors in text (see Tong et al., 2021 for a survey). There is also recent work on computational modelling of visual and multimodal metaphors in adverts, memes, and videos (Alnajjar et al., 2022; Xu et al., 2022; Zhang et al., 2021, 2023).

Metaphor use serves many purposes in communication, including the delivery of humor (Attardo, 2015; Michelli et al., 2024). Metaphor and humor share a lot of common ground. The Incongruity Theory explains that humor arises from the perception of incongruity, something that violates one's expectations (Clark, 1970; Morreall, 2024). Similarly, metaphor researchers consider incongruity as a major contextual clue for identifying metaphors in text (Cameron, 2003; Steen et al., 2010b). Metaphor and humor have even been suggested to share neural pathways in the brain (Hellberg, 2018), and visual metaphor is found to be one



"I thought our sex life was a <i>train wreck</i>."

Conceptual metaphor:
LOVE IS A JOURNEY

Explanation:
The couple's sex life is compared with a crash. This metaphorical comparison makes the caption a sarcastic comment on the couple's sex life and provides an unexpected resolution to the unusual scene in the image (a car crashing into a bedroom).

Figurative devices:
sarcasm; pun; exaggeration

Figure 5.1: HUMMUS metaphor sample, with annotation for image (bounding box and label), caption (`<i></i>`), conceptual metaphor, explanation of how the metaphor use contributes to humor, and additional figurative devices.

of the most common humorous mechanisms in cartoons (Tsakona, 2009).

This chapter is concerned with the humorous capacity of multimodal metaphors, and sheds light on how well multimodal large language models (MLLMs) understand humorous multimodal metaphor use. We created a Dataset of Humorous Multimodal Metaphor Use (HUMMUS), providing expert annotation on 1k image-caption pairs sampled from the New Yorker Caption Contest (CapCon) corpus (Hessel et al., 2023). Each image-caption pair is annotated in 6 aspects:

1. Whether it contains humorous multimodal metaphor use;
2. The conceptual metaphors involved;
3. Image areas related to the metaphor use;
4. Parts of caption related to the metaphor use;
5. How the metaphor use contributes to humor;
6. Use of other figurative devices (e.g., idiom, irony, hyperbole).

An example is provided in Figure 5.1. Our final dataset contains 558 metaphorical and 382 non-metaphorical items.

Based on the annotations, we created 6 tasks to test MLLMs' ability to identify and understand humorous multimodal metaphor use: Classification, Naming (of conceptual metaphors), ImageBbox, ImageLabel, CaptionHL, and Explanation. We tested four open-source and two closed source MLLMs: LLaVA-NeXT-110B/8B, Qwen2-VL-72B/7B-Instruct, GPT-4o, and GPT-4 Turbo, finding that the models struggle with distinguishing metaphorical and non-metaphorical items, as well as understanding the humorous multimodal metaphor use involved. Our ablation study and error analysis reveal that the models' struggles are likely to be caused by difficulties in integrating visual and textual information.

5.2 Related work

Linguistic metaphor. A large proportion of research on computational modelling of metaphor is restricted to linguistic metaphors, and automatic metaphor detection in particular (Li et al., 2023; Maudslay and Teufel, 2022; Shutova, 2015; Tong et al., 2021; Wang et al., 2023; Zhang and Liu, 2022). Most of these studies employ the VUA corpus (Steen et al., 2010b), where every single word in 115 text fragments is annotated by metaphor experts with regard to whether it is related to an underlying conceptual metaphor. The underlying conceptual metaphors themselves, however, are not specified.

Nevertheless, recent years have seen an increased interest in metaphor understanding. More paraphrase datasets are released (Joseph et al., 2023; Tong et al., 2024), following earlier works that treat metaphor understanding as a paraphrasing task (Bizzoni and Lappin, 2018; Shutova, 2010). Other studies frame the issue as an inference task, providing entailed and non-entailed (Comșa et al., 2022; Stowe et al., 2022) or contradictory statements (Chakrabarty et al., 2022b; Liu et al., 2022b) for metaphorical sentences. A framework for intentions behind metaphor use has also been proposed (Michelli et al., 2024).

Visual and multimodal metaphor. Automatic identification of text-based metaphors in videos (Alnajjar et al., 2022) and figurative language use in memes (Liu et al., 2022a) has been tackled. For metaphor understanding, the V-FLUTE dataset (Saakyan et al., 2024) provides automatically generated explanations for visual metaphors. Other datasets provide detailed annotations for monomodal and multimodal metaphors in social media and adverts (Zhang et al., 2021, 2023) or memes (Xu et al., 2022). Their annotations concern metaphor occurrence, modality (whether the metaphor is text-based, image-based, or multimodal), target domain, source domain, intent, and sentiment category.

Akula et al. (2023) annotate visual metaphors in adverts and introduce MetaCLUE, a set of five visual metaphor processing tasks: classification (whether a given image contains metaphor), localization (identifying image regions), understanding (interpreting metaphors in the form "`<target>` is as `<property>` as `<source>`") and generation (generating visual metaphors from text prompt).

Humor and metaphor. Hessel et al. (2023) propose a set of humor understanding tasks based on their New Yorker Caption Contest (CapCon) corpus. The corpus contains earlier releases (Jain et al., 2020; Radev et al., 2016; Shahaf et al., 2015) of *The New Yorker* cartoons and captions. It also provides manually created explanations for 651 image-caption pairs.

Chang et al. (2024) annotate a subset of the CapCon corpus and propose NYK-MS, a benchmark for metaphor and sarcasm understanding. While our work also uses the CapCon corpus, our contributions are different in the following ways: *1)* Every metaphorical item in HUMMUS qualifies as multimodal metaphor use, and its multimodality is reflected in the annotations; NYK-MS, on the other hand, is primarily concerned with metaphorical words in captions. *2)* We annotate the conceptual metaphors underlying the identified multimodal metaphor use; NYK-MS lacks this type of annotation. *3)* We provide bounding box annotations for image areas related to the identified metaphor use, which is also absent in NYK-MS. *4)* We explain our metaphorical items in relation to the humorous effect of the image-caption pairs, whereas NYK-MS does not consider the interplay between metaphor and humor. *5)* HUMMUS is annotated by a linguist specializing in metaphor research while the ground truth of NYK-MS is based on GPT-4V generations.

5.3 Dataset creation

HUMMUS is built upon the image-caption pairs in the CapCon corpus (Hessel et al., 2023). Each week, *The New Yorker* publishes a captionless cartoon and receives caption submissions from readers. The submitted captions are then rated by readers, based on which the magazine selects 3 finalists. Apart from the official caption contest, Jain et al. (2020) collected through crowdsourcing 119M funniness ratings for ~1M captions received in 176 contests. The CapCon corpus includes 2578 funny caption submissions for 679 cartoons of *The New Yorker*, including both the 3 official finalists per cartoon and, if available, 3 additional finalists per cartoon according to the crowdsourced ratings. We randomly sampled 251 cartoons from the corpus, for which a total of 1000 image-caption pairs are available. For the cartoon image in Figure 5.1, for instance, the following 3 captions are available:

(1) I thought our sex life was a train wreck.

(2) Well, at least he made curfew.

(3) We should tell the G.P.S. people that they changed the off-ramp.

The same image yields different jokes when combined with the different captions, and only Example (1) uses the *train wreck* metaphor. We thus treat each image-caption pair as a unique instance in terms of its humor and possible metaphor use.

We designed an annotation scheme informed by theories of humor understanding and metaphor use, and manually annotated the image-caption pairs on Labelbox. Our annotation scheme can be summarized into two stages: *1)* humorous metaphor identification, which takes less than a minute per item, and *2)* detailed metaphor annotation, which takes ~15 minutes per item.

Figure 5.2: Metaphor sample for which multiple conceptual metaphors are annotated: PUB IS A COAL MINE; HUMANS ARE ANIMALS.

5.3.1 Humorous metaphor identification

Humorous metaphor use is identified in two steps: humor understanding and metaphor identification. Image-caption pairs are tagged "Yes", "No", "WIDLII (While In Doubt, Leave It In)", or "Discard" at this stage. Given an image-caption pair, the annotator first employs the incongruity-resolution approach to understand the humor (see Section 2.4): They note down all possible incongruities in the image and see how the caption resolves those incongruities. If the annotator fails to understand the humor (incongruities remain unresolved), the item is tagged as "Discard".

If the annotator understands the humor, they proceed to determine whether the humor involves metaphor use—whether it can be attributed to any cross-domain mapping, or a process of conceptualizing/depicting one thing in terms of another. If the answer is a definite yes or no, the item is marked as "Yes" or "No" respectively. If the annotator is uncertain, the item is marked as "WIDLII" instead. We adopt this strategy from the annotation scheme behind the VUA corpus (Steen et al., 2010b). Both "Yes" and "WIDLII" items participate in the subsequent stage of detailed metaphor annotation; this allows us to avoid mistakenly marking an interesting or implicit case of metaphor use as non-metaphorical.

5.3.2 Detailed metaphor annotation

Image-caption pairs that are tagged as "Yes" or "WIDLII" in terms of the involvement of humorous metaphor use are further annotated in 4 aspects:

1. Which conceptual metaphor(s) are involved;

2. How the conceptual metaphor(s) are reflected in the image (highlighting the relevant objects);
3. How they are reflected in the caption (highlighting relevant words);
4. How the metaphor use contributes to the humorous effect of the image-caption pair (a concise explanation).

WIDLII items could be re-labelled as No or Discard items; an explanation is provided if a WIDLII item is eventually Discarded.

Conceptual metaphor. Following the tradition of previous research (Lakoff and Johnson, 1980a; Lakoff et al., 1991), underlying conceptual metaphors are annotated in "TARGET DOMAIN IS SOURCE DOMAIN" format (e.g., "LOVE IS A JOURNEY" for the image-caption in Figure 5.1).

An image-caption pair could employ multiple conceptual metaphors to achieve humorous effects. The annotator is thus asked to be as inclusive as possible, specifying all conceptual metaphors they can identify. Take the image-caption pair in Figure 5.2 as an example. The image shows two people having drinks in a pub. One of them wears a hazmat suit, which is unheard-of in pubs. The caption justifies this strange clothing choice (thus resolving the incongruity) by revealing that the person considers himself in a coal mine. There is thus a metaphorical comparison between PUB and COAL MINE. The caption also explicitly compares the man in an ordinary suit to a canary in a coal mine (in its literal sense), which is an instantiation of the HUMANS ARE ANIMALS metaphor.

Image and caption annotation. The annotator also marks out image and text fragments that are related to the annotated conceptual metaphor(s). Metaphor-related image areas are annotated using both bounding boxes and texts—a word or phrase that specifies which part of the image is highlighted. Each bounding box has a corresponding textual description. Metaphor-related words or phrases in the caption are surrounded by `<i></i>` tags.

A conceptual metaphor is a mapping between two conceptual domains. To annotate the image and the caption thus requires the annotator to determine how the two domains are reflected in the two modalities. Usually, representations of the two domains are unbalanced: The image should have a recognizable setting that represents the target or the source domain of the metaphor, while a small part of the image (e.g., a particular object) points to the other domain, thus creating incongruity and cross-domain mapping. Similarly, the context of the caption can be assigned to the target or the source domain, while a particular word or phrase suggests the involvement of the other domain. The annotator's job is therefore to mark out the less represented domain in the image and the caption respectively.

Let us return to the image-caption pair in Figure 5.1, which illustrates the LOVE IS A JOURNEY metaphor. The image predominantly belongs to the LOVE domain, with a bedroom as the setting and the couple sitting in bed occupying more than half of the image. The car crashing into the room on the left side of the image evokes the JOURNEY domain, creating incongruity and encouraging cross-domain mapping. While the cracks in the ceiling, the damaged door and the mess on the ground come along with the car crash, they are visual representations of the result of the cross-domain mapping, as opposed to belonging to the source domain, JOURNEY, itself. The car is thus annotated as the metaphor-related fragment in the image, with both a bounding box surrounding the car and a corresponding textual description, "car". The caption talks about the couple's sex life and compares it explicitly with a train wreck. The phrase *train wreck* is the "incongruous" part of the caption, evidence of the JOURNEY domain in the context of the LOVE domain. It is therefore marked out as metaphor-related.

Explanation. For each image-caption pair that contains humorous multimodal metaphor use, we also provide a short explanation about how the metaphor use contributes to the humor. This is different from the explanations provided in the CapCon corpus: While the CapCon corpus focuses on humor understanding and only explicitly mentions metaphor use for two image-caption pairs, our explanations, as exemplified in Figure 5.1, focus specifically on the interplay between humor and metaphor use.

5.3.3 Annotation of figurative devices

For both metaphorical and non-metaphorical items, we also provide annotation of the use of other figurative devices, such as pun, idiom, irony, hyperbole. If we identify the use of personification or zoomorphism, which we regard as subtypes of metaphor use (see Section 2.1.2), we also note it down.

This part of the annotation is not used for testing MLLMs in this research, as our focus is on the models' performance in metaphor processing. However, the annotation is included in the dataset and made publicly available for future research, as it provides insights into the relation between humor and the use figurative expressions in general, as well as co-occurrence of metaphor and other figures of speech in delivering humor. Moreover, it makes the dataset equally valuable for researchers who do not necessarily classify metaphor and related figurative devices (analogy, personification, idiom, etc.) in the same way as this research.

5.3.4 Inter-annotator agreement

A second expert annotated 300 image-caption pairs randomly sampled from the 1000. Like the first annotator, the second annotator is a linguist with knowledge of CMT and experience in linguistic metaphor annotation.

To ensure timely completion of the task, the procedure was slightly simplified to leave out the explanation and figurative device annotations. The second expert was thus asked to produce the following:

1. Classification of the items into Yes/No/WIDLII/Discard categories in terms of humorous metaphor use;
2. Identification of conceptual metaphors used in the Yes/WIDLII items;
3. Bounding box annotation for image;
4. Caption annotation.

To facilitate comparison with model performance, we use the same metrics for inter-annotator agreement calculation and model evaluation: multicalss F1 score for humorous metaphor identification, cosine similarity for conceptual metaphor identification, Intersection over Union (IoU) for image annotation, and Jaccard index for caption annotation.

The co-annotation process started with us introducing the second annotator to our annotation scheme. We provided them with the written annotation guidelines and verbally walked them through every step of the scheme (excluding explanation and figurative device annotation), exemplifying both metaphorical and non-metaphorical instances. After the first meeting, the second annotator was assigned the first 50 items to complete independently. Upon finishing the first 50 items, the two annotators inspected their disagreements together and discussed necessary amendments to the original annotation guidelines. After the second meeting, we updated the guidelines and let the second annotator finish the remaining 250 items independently. The annotation was completed in 2 months. With sufficient familiarity with the procedure, annotation for a metaphorical item can be completed in 7 minutes. The guidelines are provided in Appendix C.1.

The final agreement scores were calculated without further discussion of disagreements: 0.73 average F1 score for humorous metaphor identification, with 0.75 for metaphorical (Yes/WIDLII) items and 0.71 for non-metaphorical (No) items; 0.63 mean similarity score for conceptual metaphor identification ($SD = 0.20$); 0.73 mean IoU score for image annotation ($SD = 0.32$); 0.65 mean Jaccard index for caption annotation ($SD = 0.41$).

5.4 Data analysis

Hummus provides annotations for 1,000 image-caption pairs, including 558 items that contain humorous multimodal metaphor use (331 “Yes” and 227 “WIDLII”), 382 “No” items, and 60 items marked as “Discard”.

Figure 5.3: Conceptual metaphors identified in our dataset. Font size corresponds to frequency of occurrence.

Multimodality of the metaphor samples. Are the metaphors annotated in our dataset identifiable from the captions alone? To answer this question, we employ FrameBERT, an automatic textual metaphor detector proposed by (Li et al., 2023). The model reaches an F1 score of 0.73 when tested on the VUA-20 dataset (Leong et al., 2020). We consider it as a "Yes" answer if the model predicts that a given caption contains a content-word real metaphor or borderline metaphor. Compared with ground truth, the model reaches an accuracy of 0.53 and a F1 score of 0.54, both of which are close to a random guess, and also lower than the model's performance on detecting linguistic metaphors. This can be considered as an indicator of the multimodal nature of the metaphors annotated in our dataset.

Conceptual metaphor. Around 390 conceptual metaphors are identified in the dataset. As shown in Figure 5.3, The most frequently occurring ones are ANIMALS ARE HUMANS (23% metaphor samples) and HUMANS ARE ANIMALS (8%). In addition to comparison with ANIMALS, the dataset also includes mappings from HUMANS as a source domain to other target domains such as OBJECT, GOD, and ALIENS. Personification thus appears to be the most frequently used type of metaphor in these humorous image-caption pairs (51%).

"How many times do I have to tell you, don't call me at work!"

(a) ANIMALS ARE HUMANS.

"Let's keep this brief. I've got to get back to staring out the window."

(b) ANIMALS/HUMANS ARE HUMANS/ANIMALS.

Figure 5.4: Example of (a) unidirectional and (b) bidirectional metaphorical mappings between ANIMALS and HUMANS in our dataset.

Directionality of metaphors. A crucial and inevitable step in identifying any (conceptual) metaphor is determining the direction of the metaphorical mapping —in other words, which of the two conceptual domains at play is the target domain, and which is the source domain. Metaphorical mappings are unidirectional: Properties and relations of the source domain are projected onto the target domain, not the other way around. One cannot reverse the direction of a mapping without creating an entirely different metaphor. Consider the image-caption pair in Figure 5.4a, for example. It is clear that the joke is based on a metaphor with ANIMALS as the target domain and HUMANS as the source domain: A hippo is given human characteristics—it uses telephones and gets angry when someone calls it at an inconvenient time.

This rule of unidirectionality applies to "prototypical metaphors of all kinds and occurring in all media" (Forceville, 2002) while exceptions also exist (Carroll, 1994; Forceville, 1995). HUMMUS also demonstrates some exceptions to the rule, especially with regard to the HUMANS and ANIMALS domains. The image-caption pair in Figure 5.4b, for example, shows a cat having a meeting with some other animals in a modern conference room. The cat expresses its wish to keep the meeting brief, so that it could go back to staring out the window. The fact that the animals are sitting in a conference room and engaged in a meeting can be considered as employing the ANIMALS ARE HUMANS metaphor. The humor of the image-caption pair thus lies in the cat attaching ample importance to staring out the window despite its position in the company; it also invites the reader to wonder what kind of business this company might be running.

On the other hand, one can also interpret the joke as based on the metaphor

Figure 5.5: Example of metaphorically used idiom (*go through the roof*) in our dataset.

HUMANS ARE ANIMALS—people working in a company are represented as animals in the image. By representing the person in a higher position as a cat, the metaphor satirically emphasizes the absurdity of a common scenario in society: A person has an important position in a company while all they care about is something as unproductive as staring out the window. For this image-caption pair, therefore, it is difficult to distinguish HUMANS and ANIMALS in terms of target and source domains, unless we ask the caption writer directly; but it could also happen that the caption writer intended the metaphor to be bidirectional in the first place. Nonetheless, the double interpretation adds to the depth of such image-caption pairs, making them particularly interesting cases for both metaphor analysis and model evaluation.

Co-occurrence with other figurative devices. More than half (65%) of the metaphor samples in HUMMUS feature metaphor use co-occurring with the use of other figurative devices (excluding personification and zoomorphism, which are considered metaphors in this study). The most frequently used ones include pun (27%), exaggeration (10%), and satire (10%).

We also find a small percentage of metaphor samples (5%) where the humorous metaphor use concerns an idiom—recall that idioms are usually considered dead metaphors (Section 2.1.2). Consider the image-caption pair in Figure 5.5. When we look at the caption alone, its use of the idiom *go through the roof* is non-metaphorical: One understands it refers to a high cholesterol level without visualizing cholesterol actually going through the roof. When the caption is combined with the image, however, the metaphorical mapping between AMOUNT and

HEIGHT is resurrected, and it is precisely the resurrection of the dead metaphor that brings out humor.

5.5 Model evaluation

We design six tasks for humorous multimodal metaphor processing: Classification, Naming, ImageBbox, ImageLabel, CaptionHL, and Explanation. Table 5.1 provides an overview of the instructions and evaluation metrics for each task. The Classification task includes all items in the test set—that is, the image-caption pair is tagged "Yes", "No", or "WIDLII" in terms of whether or not it involves humorous multimodal metaphor use. The "No" items are considered negative cases in the Classification task; the "Yes" and "WIDLII" items positive, as our dataset provides full metaphor annotation for both categories. All other tasks only involve the positive cases.

The Naming task and the ImageLabel task employ LaBSE (Feng et al., 2022) as the evaluator. Our choice is based on a pilot Naming test that involves the first 100 items in our test set and two models: GPT-4o and GPT-4 Turbo. We calculate cosine similarity scores between the model outputs and ground truth using a variety of SBERT models. We choose LaBSE as its predictions align the most closely to human judgement of good and bad answers.

We evaluate six models, including four state-of-the-art MLLMs, GPT-4 Omni (GPT-4o, `gpt-4o-2024-05-13`), GPT-4 Turbo (`gpt-4-turbo-2024-04-09`), LLaVA-NeXT-110B (Liu et al., 2023a,b, 2024), and Qwen2-VL-72B-Instruct (Wang et al., 2024), and two smaller, open-source models, LLaVA-NeXT-8B, Qwen2-VL-7B-Instruct. The smaller models require less computing resources and can be more versatile than their larger counterparts in certain use cases (e.g., for fine-tuning).

5.5.1 Benchmark results

Table 5.2 and 5.3 show results of testing the six models on the six tasks presented in Table 5.1. We also report success rates, which measure whether the models provide meaningful answers for evaluation (e.g., if a model merely repeats the instructions, it fails to provide a meaningful answer).

The random baseline for the Classification task is calculated by randomly choosing between a "Yes" and a "No" answer for each item, and averaging the results of 100 iterations. For ImageBbox, we follow standard practice in object detection (Everingham et al., 2010; Lin et al., 2014) and consider IoU scores of 0.5 or higher as indicating sufficient overlap between model prediction and ground truth. For Naming, ImageLabel, and Explanation, we decide thresholds for good answers by manually examining model answers in various score ranges. The threshold for Naming and ImageLabel is set at 0.6. For Explanation, model answers reaching a ROUGE-1 score of 0.35 and a ROUGE-2 score of 0.087 are considered acceptable. Model outputs of different score ranges in these three tasks are presented in Appendix C.2.

How well do the models identify humorous multimodal metaphor use? All six models are prone to classifying the image-caption pairs as metaphorical. While they achieve F1 scores higher than the random baseline in identifying positive cases, even the highest F1 score for the negative category (0.47 by GPT-4 Turbo) is merely around the random baseline (0.45).

How well do the models identify the underlying conceptual metaphors? All models struggle with identifying the underlying conceptual metaphors. Only a small proportion of model predictions (21%) reachs the threshold of 0.6 in the Naming task.

How well do the models localize humorous multimodal metaphor use in image and caption? The models are much better at labelling metaphor-related image areas than providing their coordinates: Around 40% model answers reach the 0.6 threshold in the ImageLabel task, whereas the highest mean IoU score in the ImageBbox task is 0.25 (GPT-4o), which suggests barely any overlap with the ground truth.

When highlighting metaphor-related text fragments, the models are able to cover a large proportion of ground truth annotations, but their answers also tend to be longer, including text fragments that are not related to the humorous metaphor use.

How well do the model explain humorous multimodal metaphor use? The Explanation task proves to be difficult for the models. Merely 5% of model predictions reach our threshold of acceptable explanations.

5.5.2 Prompt engineering

We experiment with other ways to formulate the Classification task, to see whether the high probability of "Yes" answers is associated with the prompt we use. Using the first 100 items in the test set, we run a pilot study that tests a wide range of prompts on one of the worst performing models, Qwen2-VL-7B-Instruct. Instead of asking the model to reply with "Yes" or "No", these prompts require different ways to label the given image-caption pair, such as "Metaphorical/Non-metaphorical", "True/False", "A/B".

Based on the performance of Qwen2-VL-7B-Instruct in the pilot study, we select the top-3 prompts that result in the highest average F1 scores. These prompts ask the model to reply with *1)* "Yes (i.e., metaphor use is involved) or No (i.e., metaphor use is not involved)", *2)* "No or Yes", and *3)* "A or B", respectively.

In the third prompt, the order of the two options (the given image-caption pair "involves" or "does not involve" humorous metaphor use) is randomized for each test item. Full forms of the prompts are provided in Appendix C.2.

We test these three prompts on three models: GPT-4o, and the two Qwen2-VL models. GPT-4o is one of the best-forming models over all tasks, and its success rate is always 100%. The two Qwen2-VL models represent the best and the worst open-source models. The 72B model has 100% success rate on all tasks. The 7B model is the most unstable, achieving a 41% success rate on the CaptionHL task.

As shown in Table 5.4, the two Qwen2-VL models, especially the smaller one, are sensitive to different prompts. The performance of GPT-4o, on the other hand, remains stable over different prompts. The experiment thus proves the reliability of our Classification benchmark results.

5.5.3 Comparison with human performance

Is the models' compromised performance really due to the their lack of capacity to process humorous multimodal metaphor use, or is it due to the subjectivity of the tasks? To answer this question, we design a new prompt that is similar to the annotation guidelines we provided to the second expert annotator—the models are instructed to complete Classification, Naming, ImageBbox, and CaptionHL in one go, following the same steps as the expert annotator. We also ask the models to adopt a persona that is representative of the human experts: a linguist who is familiar with CMT and has experience in manual metaphor annotation. The prompt is provided in Appendix C.2. It is tested on the smallest model, Qwen2-VL-7B, before being used for the full experiment.

As shown in Table 5.5, even when following similar guidelines, none of the models reach the level of agreement between the two expert annotators. Model performance is closer to human performance in the Naming task, as compared to the other tasks. In CaptionHL, LLaVA-NeXT-110B reaches a Jaccard index score (0.62) that is close to human (0.65), but with a low success rate (62%): for ~40% of the instances the model identifies as metaphorical, it fails to provide a caption annotation. In Classification and ImageBbox, the human expert surpasses the models by a large margin.

Also note that this new approach does not help with the Classification task despite its more detailed instructions about what is metaphor and how to identify humorous metaphor use: most of the models appear to be better at excluding negative cases in the benchmark task (Table 5.2), although the performance is still much lower than the human expert. Nonetheless, the new approach benefits the Naming task and yields higher Jaccard index scores in the CaptionHL task for GPT-4o, LLaVA-NeXT-110B, and Qwen2-VL-7B, albeit with compromised success rates for the two larger models. Overall, the new approach of following the manual annotation guidelines is better suited for detailed annotation of humorous metaphor use (Naming, ImageBbox, and CaptionHL). As for the Classification

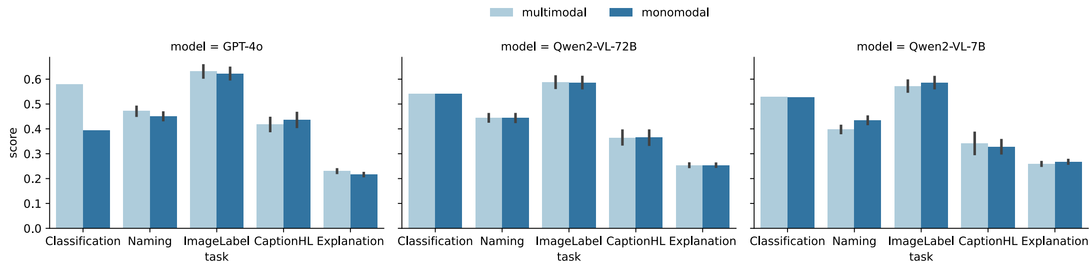


Figure 5.6: Model performance in multimodal versus monomodal experiments: average F1 score for Classification, sentence similarity score for Naming and ImageLabel, Jaccard index score for CaptionHL, ROUGE-1 for Explanation. Success rate is 100% except for Qwen2-VL-7B-Instruct in the monomodal CaptionHL task (97%).

task, the models' inability to identify negative cases seems to go beyond what can be achieved by prompt engineering.

5.5.4 Ablation study

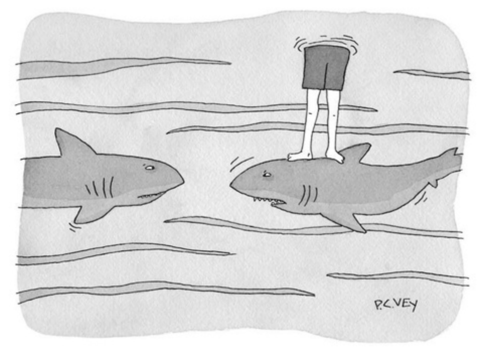
To examine the models' processing of multimodal input in our tasks, we design an ablation study that replaces image input with textual descriptions of the images, making the input data purely textual. We use the `image_description` data provided in the CapCon corpus, which are short, literal descriptions of the scene. For example, the description for the image in Figure 5.1 is as follows: "A man and woman are in bed together under the covers. They are looking towards the bedroom door when a car has crashed into their home. They don't seem too upset about the situation."

We rerun the six tasks on the same three models for prompt engineering: GPT-4o and the two Qwen2-VL models. For Classification, we use the prompt that results in the highest average F1 score for most models in prompt engineering: It asks the models to answer the question with "No or Yes" instead of "Yes or No".

As shown in Figure 5.6, there is not much difference between the models' performance in the multimodal and monomodal experiments, except for GPT-4o in the Classification task. The comparison indicates that the multimodal input data is not adequately utilized by the models. Our error analysis provides more support for this.

5.5.5 Error Analysis

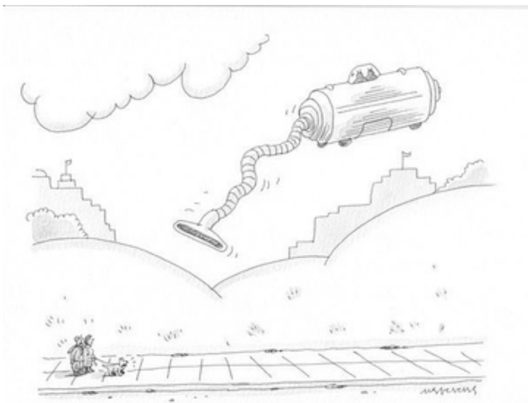
A primary reason for incorrect answers is the models' inability to integrate visual and textual information into a coherent story. For example, the humor of the image-caption pair in Figure 5.7 is based on personification of the sharks: They communicate in human language and can experience and express embarrassment.



"Well, that's embarrassing. How long has it been there?"

Class.	Yes	No
Naming	ANIMALS ARE HUMANS; A PERSON IS AN OBJECT	EMBARRASSMENT IS A SHARK ATTACK
Image.	Person	Shark
Caption.	<i>Well, that's embarrassing. How long has <i>it</i> been there?</i>	Well, that's embarrassing. How long has it been there?
Expla-nation	The shark on the right didn't realize someone is standing on it until another shark point it out for it	The image of a person standing on a shark while another shark approaches from behind creates a metaphor for being in a precarious situation ....

Figure 5.7: Ground truth (left) versus LLaVA-NeXT-110B predictions (right) for the given image-caption pair. Explanations are shortened.



"And they say life can't exist in a vacuum."

Figure 5.8: All models except GPT-4o predict LIFE IS A VACUUM for this item in the Naming task.

Subsequently, the person in the image is compared metaphorically to an object, referred to as *it* in the caption.

LLaVA-NeXT-110B's answers indicate that the model assumes the caption is uttered by the person in the image, thus overlooking the ANIMALS ARE HUMANS metaphor and the subsequent A PERSON IS AN OBJECT metaphor. This example also shows the importance of metaphor processing in humor understanding: One cannot understand the humor of this image-caption pair without recognizing that the sharks are personified.

On the other hand, the models can usually identify common metaphorical or idiomatic expressions in the caption, although it does not guarantee adequate understanding of humorous multimodal metaphor use. For example, the cartoon in Figure 5.8 depicts an alien spaceship as a vacuum cleaner. The humor comes from the pun on *vacuum* in the caption: It refers both to aliens physically existing in a vacuum cleaner, as depicted in the image, and life existing *in a vacuum* in an idiomatic sense. The ground truth annotation of the conceptual metaphor is ALIEN SPACESHIP IS A VACUUM CLEANER, because it is the base of the pun on *vacuum* in the caption. The MLLMs acknowledge the metaphoricity of the expression *life can't exist in a vacuum*, but their prediction in the Naming task, LIFE IS A VACUUM, indicates that they do not relate it to the vacuum cleaner spaceship in the image, thus failing to properly understand the humorous multimodal metaphor use.

Our analysis indicates that the models' performance can potentially be improved when they are instructed more explicitly to combine image and caption information. For example, one can use chain-of-thought prompting to instruct an MLLM to mimic the human annotation process: first process the image, and then integrate it with the caption, before proceeding to metaphor identification and understanding.

5.6 Conclusion and limitations

This study releases a dataset that provides expert annotation on humorous multimodal metaphor use. Using the dataset, we benchmark popular and state-of-the-art MLLMs on their capabilities to identify and understand humorous multimodal metaphor use. Our experiments show that current MLLMs struggle with processing humorous multimodal metaphor use, especially with regard to integrating visual and textual information.

Limitations and future directions. We hope this study will encourage more research on MLLMs' capabilities to process humorous multimodal metaphors. While we have experimented with promoting model performance through prompt engineering, other possibilities such as few-shot prompting and fine-tuning are beyond the scope of the current study.

We also acknowledge that the latest reasoning models such as DeepSeek R1 and OpenAI o1 are not included in this study. They are released after the experiments were conducted and do not necessarily have multimodal capabilities (e.g., DeepSeek R1).

5.7 Ethical considerations

We accessed the GPT-4 models through OpenAI API. The open-source models (LLaVA-NeXT-8/110B and Qwen2-VL-7/72B) as well as the CapCon corpus were accessed through HuggingFace. The CapCon corpus has CC-BY-4.0 license. Our dataset is freely accessible on GitHub.

Our dataset includes jokes that could be considered offensive, and certain jokes may be inappropriate for a younger audience. These data remain in our dataset as they are a nonremovable part of multimodal humor and can be valuable for future research.

Task	**Instructions**	**Evaluation**
Classification	"Does the humor of the given image-and-caption combination involve metaphor use? Answer the question with Yes or No."	Multiclass F1 score
Naming	"The humor of the given image-and-caption combination involves metaphor use. Which conceptual metaphor is used? Answer the question in 'TARGET DOMAIN IS SOURCE DOMAIN' format (e.g., 'LOVE IS A JOURNEY')."	Sentence similarity
ImageBbox	"The humor [...] involves metaphor use. Which object in the image is related to the metaphor? Answer with its label and normalized bounding box coordinates in 'label: [top, left, height, width]' format."	IoU, precision, recall
ImageLabel	"The humor [...] involves metaphor use. Which object in the image is related to the metaphor? Answer the question with a single word."	Sentence similarity
CaptionHL	"The humor [...] involves metaphor use. Which part of the caption is related to the metaphor? Surround it with a pair of <i></i> tag."	Jaccard index, precision, recall
Explanation	"How does metaphor use contribute to the humor of the given image-and-caption combination? Explain in no more than 30 words."	ROUGE-1, ROUGE-2

Table 5.1: Benchmark tasks and corresponding prompts and evaluation metrics. A full prompt includes an image, a caption, and instructions: <image>+Caption:␣<caption>\n\n<instructions>. The first sentence of ImageBbox/Label and CaptionHL tasks are the same as the Naming task. IoU = Intersection over Union.

Model		**Classification**			**Naming**	**Explanation**	
		Pos	Neg	Avg		ROUGE-1	ROUGE-2
GPT-4	Omni	0.70	0.39	0.55	0.47 (0.22)	0.23 (0.08)	0.03 (0.04)
	Turbo	0.64	0.47	0.56	0.49 (0.21)	0.22 (0.08)	0.03 (0.04)
LLaVA-NeXT	110B	0.69	0.36	0.53	0.43 (0.18)	0.27 (0.08)	0.05 (0.04)
	8B	0.74	0.00	0.37	0.46 (0.18)	0.26 (0.08)	0.04 (0.04)
Qwen2-VL	72B	0.64	0.43	0.54	0.44 (0.18)	0.25 (0.08)	0.04 (0.04)
	7B	0.61	0.46	0.53	0.40 (0.18)	0.26 (0.08)	0.04 (0.04)
Random		0.54	0.45	0.50	–	–	–

Table 5.2: Classification, Naming, and Explanation benchmarks. Standard deviations are in parenthesis. Success rate is 100% except GPT-4 Turbo in the Naming task (96%).

Model		**ImageBbox**			**ImageLabel**	**Caption**		
		P	R	IoU		P	R	Jaccard
GPT-4	Omni	0.40 (0.34)	0.43 (0.34)	0.25 (0.22)	0.63 (0.29)	0.47 (0.36)	0.85 (0.32)	0.42 (0.32)
	Turbo	0.38 (0.37)	0.32 (0.34)	0.19 (0.21)	0.56 (0.27)	0.54 (0.39)	0.80 (0.37)	0.47 (0.36)
LLaVA-NeXT	110B	0.17 (0.20)	0.50 (0.38)	0.14 (0.16)	0.58 (0.27)	0.42 (0.38)	0.93 (0.24)	0.40 (0.37)
	8B	0.17 (0.20)	0.64 (0.39)	0.15 (0.16)	0.58 (0.27)	0.33 (0.33)	0.95 (0.20)	0.31 (0.30)
Qwen2-VL	72B	0.17 (0.20)	0.40 (0.35)	0.13 (0.14)	0.59 (0.28)	0.40 (0.37)	0.89 (0.29)	0.36 (0.34)
	7B	0.17 (0.21)	0.40 (0.34)	0.13 (0.14)	0.57 (0.27)	0.36 (0.35)	0.94 (0.22)	0.34 (0.32)

Table 5.3: ImageBbox, ImageLabel, and CaptionHL benchmarks in mean (SD) format. A few tests result in success rates less than 100% : no less than 98% in the ImageBbox task; 97% for LLaVA-NeXT-8B, 89% for LLaVA-NeXT-110B, and 41% for Qwen2-VL-7B in the CaptionHL task.

Model	**Positive**	**Negative**	**Average**
GPT-4o	0.71 (0.01)	0.43 (0.02)	0.57 (0.01)
Qwen2-VL-72B	0.61 (0.06)	0.46 (0.07)	0.53 (0.01)
Qwen2-VL-7B	0.66 (0.06)	0.36 (0.11)	0.51 (0.03)

Table 5.4: Mean F1 scores using 3 different prompts for the Classification task. Qwen2-VL stands for Qwen2-VL-Instruct. Success rate is always 100%.

Model		**Classification**			**Naming**	**ImageBbox**	**Caption**
		Pos	Neg	Avg			
GPT-4	Omni	0.70	0.36	0.53	0.53 (0.23)	0.25 (0.20)	0.50 (0.39)
	Turbo	0.73	0.22	0.48	0.50 (0.23)	0.21 (0.20)	0.33 (0.33)
LLaVA-NeXT	110B	0.73	0.00	0.37	0.49 (0.19)	0.17 (0.17)	0.62 (0.43)
	8B	0.74	0.01	0.37	0.45 (0.17)	0.15 (0.16)	0.30 (0.31)
Qwen2-VL	72B	0.73	0.00	0.37	0.47 (0.19)	0.17 (0.19)	0.39 (0.39)
	7B	0.73	0.00	0.36	0.51 (0.20)	0.13 (0.15)	0.38 (0.36)
Human		0.75	0.71	0.73	0.63 (0.20)	0.73 (0.32)	0.65 (0.41)

Table 5.5: Model performance compared with human performance on Classification, Naming, ImageBbox (IoU), and Caption (Jaccard index) tasks. Prompts for the models are similar to the annotation guidelines. Some of the tests result in success rate less than 100%: $\geqslant$90% in Naming; $\geqslant$90% for the Qwen2-VL models and $\geqslant$80% for the other models in ImageBbox; $\geqslant$90% for GPT-4 Turbo and Qwen2-VL-72B, 79% for GPT-4o, and ~60% for the other models in CaptionHL.

Chapter 6

Cultural differences in humor appreciation

6.1 Introduction

Humor exhibits both universality and cross-cultural variation: While the ability to appreciate and produce humor is shared by the entire human race, it is also shaped by one's cultural background. Previous studies find humor to be perceived more positively in Western cultures compared to Eastern cultures; Westerners also tend to use more aggressive humor while Easterners prefer more affiliative humor (Jiang et al., 2019). As LLMs reach global users, modeling such differences becomes essential for the models to align their responses with the preferences and expectations of specific cultures. Misaligned behavior such as overlooking the humor in user input or interpreting seriousness as humorous may upset the user unintentionally.

Cultural alignment in LLMs has been rigorously researched in recent years, with a heavy focus on cultural values, such as collectivism vs individualism, significance of friendship, confidence in government (AlKhamissi et al., 2024; Kabir et al., 2025; Masoud et al., 2025b). The community has not yet looked at the problem in the context of humor processing. Therefore, while studies have consistently shown Western cultural biases in LLMs, it is unclear whether these biases also apply to humor processing; and if they do, how they manifest in this specific scenario, and how to alleviate them.

This chapter serves as a first step towards a framework for evaluating LLMs' cultural alignment in humor processing. Our research question is: How do cultures differ in humor appreciation? By answering this question, we aim to establish human baselines that LLMs can be compared to. We consider the relationship between humor, emotion, and metaphor use, and how it varies across cultures. Humor and emotion are closely related: Humor appreciation activates brain regions associated with emotional processes (Farkas et al., 2021); humor is also an effective strategy to cope with negative emotions (Kugler and Kuhbandner, 2015; Samson and Gross, 2012). However, humor is not equally appreciated in all cultures (Lu, 2023); culture also influences one's preferred types of humor (Schermer et al., 2023). There could be cultural differences with regard to how humor is associated with different emotions. Moreover, metaphors are frequently used to induce humor (Chapter 4 and 5), and metaphorical expressions tend to be regarded as more emotional than equivalent literal expressions (Mohammad et al., 2016). Metaphor use could therefore add another layer of complexity to emotional responses to humorous content.

We thus seek answers to the following questions in this chapter: *1)* To what extent do cultures differ in what is considered humorous? *2)* How do cultures differ in terms of emotional responses to humorous stimuli? *3)* Is metaphor use associated with humor appreciation in different cultures? We utilized the HUMMUS dataset (Chapter 5) to answer our questions: it contains 1,000 New Yorker cartoons (image-caption pairs), ~560 of which are identified as involving humorous multimodal metaphor use. Through a web-based experiment, we collected funniness and emotion annotations for these cartoons from four diverse cultures: the U.S., Mexican, Polish, and Chinese. These four cultures differ in terms of primary spoken languages, history, and ideologies, providing a fair starting point to building a paradigm that is applicable to a broad range of cultures.

We collected a total of 25,600 data points for 800 cartoons, including 482 metaphorical ones. Our data analyses showed that the cartoons tended to be considered funnier and receive more emotional responses from Mexican and Chinese culture than Polish and the U.S. culture. On the other hand, all cultures tended to give higher funniness ratings and select more emotions for cartoons involving humorous multimodal metaphor use than those not involving the metaphor use. We also found a similar pattern of how funniness and emotion categories were associated in the four cultures, although there were also subtle differences. Through qualitative analysis, we found that one may focus more on the source domain of a metaphor if the metaphor was not an intrinsic part of one's culture or was expressed differently in one's native language. This could lead to negative emotions and low funniness ratings. Nevertheless, familiarity with the setting of a cartoon does not guarantee a positive or negative response; how the humor is appreciated will still depend on how the message of the cartoon is perceived in a culture.

6.2 Related work

Cultural preferences in humor. Previous studies have investigated how cultures differ in terms of attitude towards humor. A common consensus is that Western cultures such as the U.S. have a more positive view of humor than Eastern cultures such as Chinese. Jiang et al. (2011), for example, find that humor is associated with positive adjectives in the U.S. culture and negative adjectives in Chinese culture. Yue (2011) finds that while humor is valued in Chinese culture, Chinese people do not consider themselves humorous. Both Yue (2011) and Yue and Hiranandani (2014) suggest that humor is more associated with comedians than ordinary people in Chinese culture.

Cultural preferences in humor usage has also been investigated. Many studies used the Humor Styles Questionnaire (HSQ; Martin et al., 2003) to measure how one uses humor (Chen and Martin, 2007; Hiranandani and Yue, 2014; Kalliny et al., 2006). Four humor styles are identified: affiliative, self-enhancing, aggressive, and self-defeating (see Section 2.4). Schermer et al. (2023) collected HSQ responses from 28 countries and found affiliative humor to be the most popular humor style across all countries; self-enhancing humor is the least common in Japan. Pozdena (2020) annotated 31 American and 31 Hungarian TED talks in terms of humor type. They considered 3 humor types that corresponded to the 3 groups of humor theories: incongruity-based, superiority-based, and release-based (see Section 2.4). Incongruity-based humor was found to be the most represented in both cultures, followed by superiority-based humor. The author argued that the popularity of superiority-based humor could be explained by both cultures being individualistic.

Cultural alignment in LLMs. Studies on cultural alignment in LLMs have mainly focused on culture values, leveraging culture value questionnaires such as World Values Survey (Haerpfer et al., 2022) and Hofstede's cultural dimension framework (Hofstede et al., 2010) to compare model answers with human baselines (AlKhamissi et al., 2024; Bulté and Terryn, 2025; Masoud et al., 2025b). Western cultural biases are found consistently across studies. Several strategies to improve cultural alignment have been studied, including prompting (AlKhamissi et al., 2024), fine-tuning (Masoud et al., 2025b), and prompt tuning (Masoud et al., 2025a). Kabir et al. (2025) demonstrated that strictly following the close-style culture value surveys was insufficient for understanding cultural alignment in LLMs. They thus called for a more robust evaluation framework.

6.3 Data collection

We designed a web-based experiment to collect humor appreciation data for the captioned New Yorker cartoons in the HUMMUS dataset. The experiment interface is presentedn in Figure 6.1. For each cartoon, we asked two questions: *1)* Do you find the cartoon funny? *2)* What emotions arose after reading the cartoon? A 5-point Likert scale was used for the first question to register answers ranging from "definitely not" to "definitely yes". The second question gave participants 8 options, including 6 primary emotions from Plutchik's Wheel of Emotions (Plutchik, 1980): Joy, Sadness, Fear, Disgust, Surprise, and Anger; as well as an "Other, namely" option for participants to provide emotions outside of the provided 6, and a "None" option. The "None" option was exclusive; otherwise

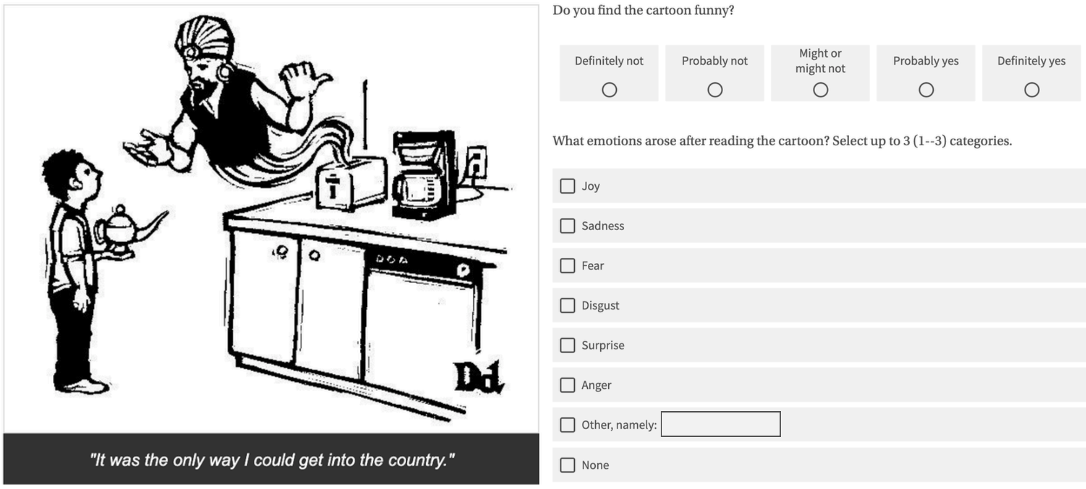


Figure 6.1: Data collection interface.

up to 3 options can be selected for each cartoon.

We included four cultures: Chinese, Mexican, Polish, and the U.S.. For Mexican, Polish, and the U.S. cultures, we used the following screeners to select participants: *1)* The participant is 18 years old or above; *2)* The participant spent most of their time before turning 18 in the target country (Mexico, Poland, the U.S.); *3)* Both the participant's first language and their primary spoken language are the (de facto) official language of the target country (Spanish, Polish, English); *4)* The participant is fluent in English. For Chinese culture, we expanded screener 2 from a single country to four regions: Mainland China, Hong Kong, Macau, and Taiwan. We also changed screener 3. While Mandarin can be considered the official language of both People's Replublic of China (which includes Mainland China, Hongkong, and Macau) and Taiwan, it is far from the first or primary spoken language for a large part of the population. Instead of first language and primary spoken language, therefore, we accepted participants who were fluent in Chinese, Cantonese, Mandarin, or Hakka. These languages/dialects were both widely spoken in the target regions and listed by Prolific as individual languages. There were also other languages that were first languages of people living in certain parts of China, such as Mongolian (in Inner Mongolia) and Korean (in Northeastern Jilin). These were not included as they were primarily related with other regions/cultures.

We collected 32 responses (8 per culture) for each cartoon. Each participant provided response to 20 cartoons; participating for multiple times was not allowed. While HUMMUS included multiple captions per image, we made sure that there was always only one instance of each image in each batch of 20 cartoons. We also did the following to avoid bots and low-effort responses: *1)* A reCAPTCHA was implemented at the beginning of the experiment. *2)* Participants were informed that external resources, including dictionaries, translation tools, and AI tools,

Effect	df	F	p
Metaphor	2, 782.49	15.20	$<$ .001
Culture	3, 1248.30	25.24	$<$ .001
Age	1, 1264.56	3.61	.058
Sex	1, 1265.79	0.29	.592
Education	7, 1264.80	0.63	.730

Table 6.1: ANOVA summary of fixed effects (humorous multimodal metaphor use in cartoon, culture, age, sex, and highest level of education completed) from linear mixed model of funniness ratings with random intercepts for cartoons and participants.

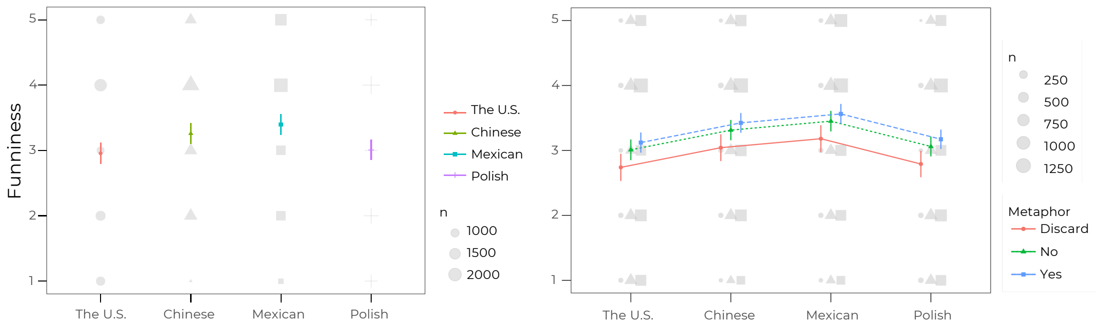


Figure 6.2: Estimated marginal means of funniness ratings by culture (left); by both culture and humorous multimodal metaphor use (right).

were strictly prohibited before they accepted the study on Prolific, before they gave consent to participate, and on every page where they were asked to provide responses to cartoons (including the screening questions to test their English proficiency). *3)* Participants had to pass a small test right after the reCAPTCHA to show that they spoke English and could read the cartoon images. *4)* Right click was disabled throughout the experiment. *5)* Tab switches were monitored throughout the experiment. Experiment would be terminated automatically if the participant switched tabs or minimized the window for 3 times. *6)* Two attention checks for included.

6.4 Data analysis

From 1,280 participants (320 per culture), we collected a total of 25,600 responses for 800 cartoons. Demographic information of the participants is summarized in Appendix D.1.

Effect	df	F	p
Metaphor	2, 773.29	12.83	$<$ .001
Culture	3, 1255.84	21.28	$<$ .001
Age	1, 1267.24	0.38	.540
Sex	1, 1268.21	0.91	.339
Education	7, 1267.57	0.86	.543

Table 6.2: ANOVA summary of fixed effects (humorous multimodal metaphor use in cartoon, culture, age, sex, and highest level of education completed) from linear mixed model of number of emotions selected with random intercepts for cartoons and participants.

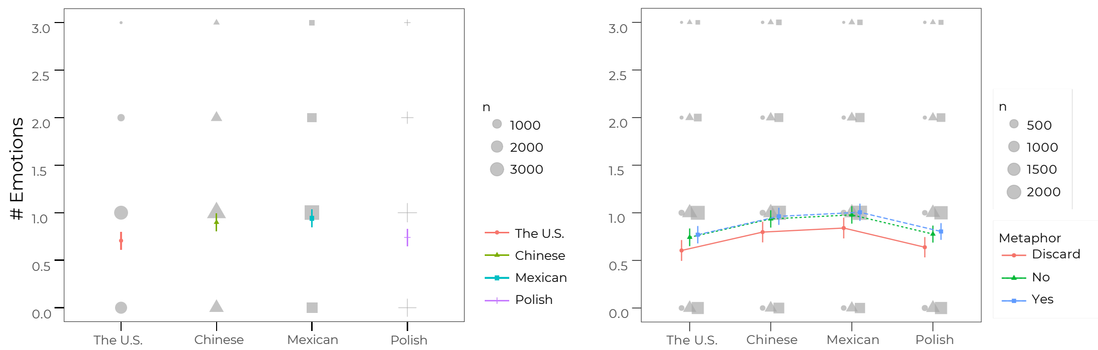


Figure 6.3: Estimated marginal means of number of emotions selected by culture (left); by both culture and humorous multimodal metaphor use (right).

6.4.1 Effect of culture and metaphor use on funniness ratings

We fitted a linear mixed model using restricted maximum likelihood to predict funniness ratings from humorous multimodal metaphor use in cartoon as well as participants' culture, age, sex, and highest level of education completed, with random intercepts for cartoons and participants.[1] An ANOVA on the fixed effects (Table 6.1) showed that both culture and metaphor use were significant predictors of funniness ratings (culture $F(3, 1248.30) = 25.24$, $p < .001$; metaphor use $F(2, 782.49) = 15.20$, $p < .001$).

Figure 6.2 presents estimated marginal means (EMMs) obtained from the model. Mexican culture showed the highest adjusted mean funniness rating (3.40 ± 0.08, 95% CI $[3.24, 3.56]$), followed by Chinese (3.26 ± 0.08, 95% CI $[3.10, 3.42]$) and Polish (3.01 ± 0.08, 95% CI $[2.86, 3.16]$); the U.S. culture had the lowest EMM (2.96 ± 0.08, 95% CI $[2.80, 3.12]$). Pairwise comparison using the

[1] We used the JASP software (JASP Team, 2026) to conduct the statistical analyses.

Emotion	df	χ^2	p
Joy	3	68.28	$< .001$
Sadness	3	61.43	$< .001$
Fear	3	74.46	$< .001$
Disgust	3	18.66	$< .001$
Surprise	3	19.34	$< .001$
Anger	3	28.88	$< .001$
Other	3	18.30	$< .001$
None	3	61.74	$< .001$

Table 6.3: Abridged ANOVA summary of effect of culture from mixed effects (culture, age, sex, highest level of education completed) logistic regression model of each emotion category (with random intercepts for participants and cartoons).

Holm-Bonferroni Method for p-value adjustment indicated that there was no significant difference between the EMMs of Polish and the U.S. culture ($p = .407$). In general, therefore, the cartoons were significantly funnier for participants from Mexican culture than those from Chinese culture; both cultures also found the cartoons significantly funnier than Polish and the U.S. culture.

Regarding humorous multimodal metaphor use in the cartoons, the EMMs indicated that all cultures tended to give higher funniness ratings to metaphorical cartoons than non-metaphorical ones; cartoons not fully understood by the expert metaphor annotator (the Discarded items) also tended to be the least funny across all cultures. Pairwise comparison using the Holm-Bonferroni Method for p-value adjustment showed that the differences in funniness ratings for metaphorical and non-metaphorical cartoons were significant across all cultures ($p = .005$).

6.4.2 Effect of culture and metaphor use on emotional response

Presence of emotions. We fitted a linear mixed model to predict the number of emotions selected from humorous multimodal metaphor use in cartoon, culture, age, sex, and highest level of education completed, with random intercepts for cartoons and participants. The ANOVA summary (Table 6.2) showed that both culture and metaphor use were significant predictors of the number of emotions selected (culture $F(3, 1255.84) = 21.28$, $p < .001$; metaphor use $F(2, 773.29) = 12.83$, $p < .001$).

The adjusted mean number of selected emotions (Figure 6.3) was the highest in Mexican culture (0.94 ± 0.05, 95% CI $[0.85, 1.03]$), followed by Chinese (0.90 ± 0.05, 95% CI $[0.81, 0.99]$) and Polish (0.74 ± 0.05, 95% CI $[0.65, 0.83]$), and the lowest in the U.S. culture (0.71 ± 0.05, 95% CI $[0.61, 0.80]$). The differences in EMMs were significant for all culture pairs ($p < .001$, adjusted using the Holm-Bonferroni Method) except for Mexican-Chinese ($p = .408$) and Polish-the U.S. ($p = .408$). In other words, participants from Mexican and Chinese culture were significantly more likely to select any emotions than those from Polish and the U.S. culture.

On the other hand, participants from all 4 cultures tended to select more emotion categories for metaphorical cartoons than non-metaphorical cartoons (Figure 6.3). However, the differences between the EMMs were not statistically significant ($p = .069$, adjusted using the Holm-Bonferroni method).

Emotion categories. For each emotion category, we fitted a mixed effects logistic regression model to predict whether the category was selected from culture, age, sex, and highest level of education completed, with random intercepts for participants and cartoons. We initially attempted to include metaphor use as a fixed effect variable, but the model for Disgust failed to reach convergence within a reasonable computation time in JASP. A likely cause was a lack of sufficient data (1,472 selections of Disgust in 25,600 observations) to support the complexity of the model. We therefore opted for a simpler setup for more stable estimates.

Effect of culture from each of the models is provided in Table 6.3. Full ANOVA summary of fixed effects from the models, as well as EMMs by culture, are provided in Appendix D.2. The ANOVA results showed that culture had a significant main effect on all of the emotion categories ($p < .001$).

The EMMs indicated both cross-cultural similarities and differences in emotional responses to the cartoons. Joy, Surprise, and None (absense of emotion) had the highest estimated probabilities of being selected across all cultures. Participants from Mexican culture, in particular, were more likely to select Joy and Surprise than the other cultures. Participants from Chinese culture were more likely to select Fear, Sadness and Anger; both Mexican and Chinese culture had a slightly higher probability to select Disgust than Polish and the U.S.. Participants from Polish culture were more likely to select None, while those from the U.S. culture were more likely to provide answers outside of the given categories.

6.4.3 Association between funniness and emotions

For each culture, we performed a chi-square test of independence using the `scipy.stats` module in Python to assess the relationship between funniness ratings and emotion categories. As shown in Table 6.4, there was a significant relationship between the two variables for all cultures.[2] We also calculated the effect size of the associations in terms of Cramér's V (Table 6.4): the effect size was the smallest for Chinese culture (0.34) and the largest for Polish culture

[2]There was one cell with value lower than 5 but larger than 1 for the U.S. and Mexican culture. Since it only accounts for 2.50% of the cells for either culture, we consider the results of chi-square test reliable.

	χ^2	df	N	p	V
Chinese	3494.05	28	7687	$< .001$	0.34
Mexican	4081.49	28	7818	$< .001$	0.36
Polish	4477.82	28	7484	$< .001$	0.39
The U.S.	3512.01	28	7163	$< .001$	0.35

Table 6.4: Results of chi-square test of independence and Cramér's V of funniness ratings and emotion categories from each culture.

(0.39). Taking into account the dimensions of the contingency tables (5 funniness categories × 8 emotion categories), we followed Cohen (1988) and interpreted the results as indicating strong associations between funniness and emotion responses from all cultures.

Figure 6.4 shows standardized residuals calculated from the chi-square tests. Overall, we found a similar relationship between funniness ratings and emotion categories in the four cultures, with slight differences. In all cultures, high funniness ratings (> 3) were most strongly associated with Joy while low ratings were most strongly associated with the None category (absence of emotion). Surprise was also associated with high funniness ratings while Disgust, Anger, and Sadness low funniness ratings. Fear was associated with a lower funniness rating in Chinese and Polish culture than Mexican and the U.S. culture. The Other category was the most associated with low funniness ratings in Chinese culture, and the least so in the U.S. culture.

6.4.4 The "Other" emotion categories

When answering the emotion questions, participant could select "Other, namely" and provide an answer outside of the given options (6 primary emotions and "None"). We received a total of 2,033 "Other" responses from the participants. We corrected spellings and merged answers that only had grammatical differences (e.g., "amused" and "amusement"; "unfunny" and "not funny"). Different words that were arguably similar in meaning (e.g., "amused" and "funny") were still considered as different answers.

As shown in Figure 6.5, the proposed emotions can be positive (e.g., happiness), negative (e.g., annoyance), or neutral (e.g., intrigue). Participants also expressed emotions at different levels of intensity: Apart from "amused" or "amusement", there were also "very amused", "mild amusement", and "very mild amusement".

Overall, the most frequent answers were "confused" (32%), "amused" (11%), and "bored" (5%). A similar pattern was found within the individual cultures (Figure 6.6), although "indifference" outnumbered amusement and boredom in Mexican and Polish culture. Participants from Chinese culture used the word

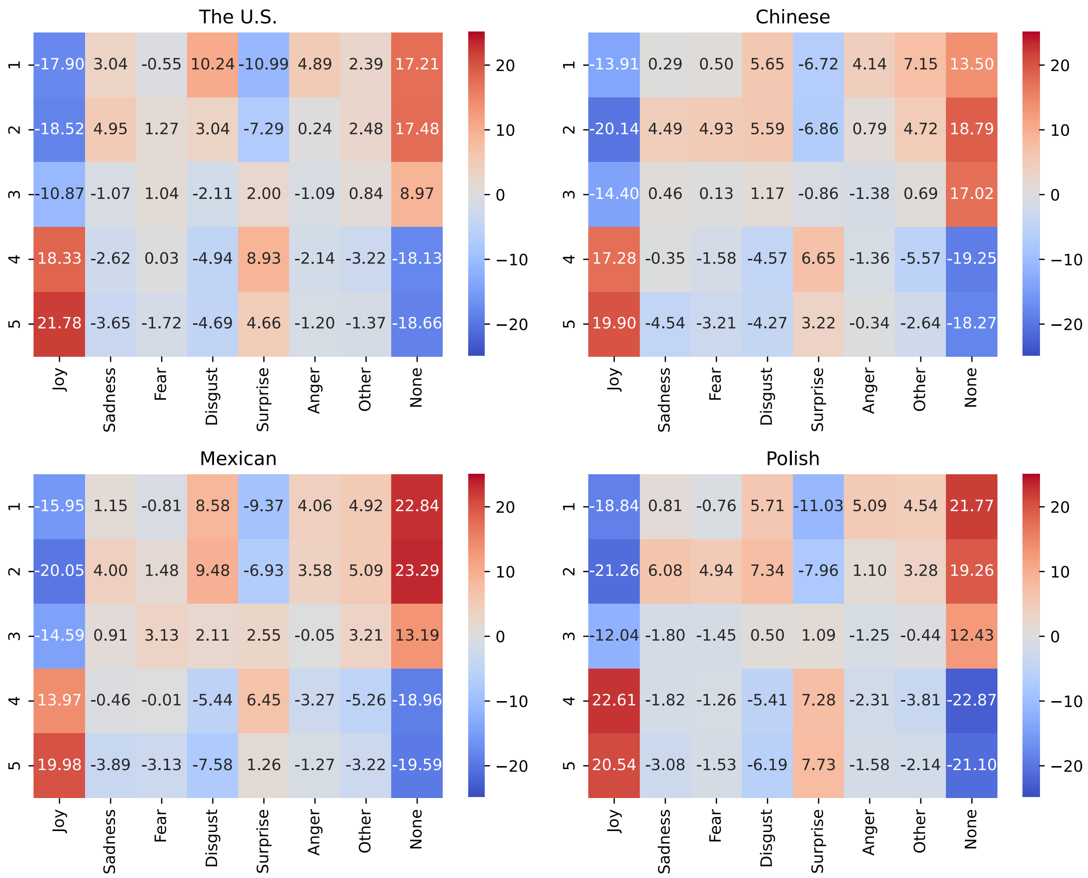


Figure 6.4: Standardized residuals calculated from chi-square test of independence of funniness ratings and emotion categories per culture.

"funny" much more frequently than "amused". "Funny" was also used more frequently in Mexican and Polish culture than the U.S. culture.

The use of the word "funny" in the Chinese, Mexican, and Polish samples indicated an underlying relationship between emotion and verbal expressions of emotion. Participants from the U.S. culture provided the largest number of unique answers ($N = 123$), which corresponded with their higher estimated probability of selecting the "Other" option than the other cultures (Section 6.4.2). Since the experiment was in English, and the U.S. was the only culture among the four where English must be a person's first language, it was probably the easiest for participants from the U.S. culture to articulate their emotional reactions from those from the other three cultures. On the other hand, there were a few occassions where a participant from Mexican culture used Spanish words such as "risa" and "desilusión" in their answers. We treated these as different from their most likely translations in English ("laughter" for "risa", "disappointment" or "disillusion" for "desilusión") because they may contain culture- and language-specific

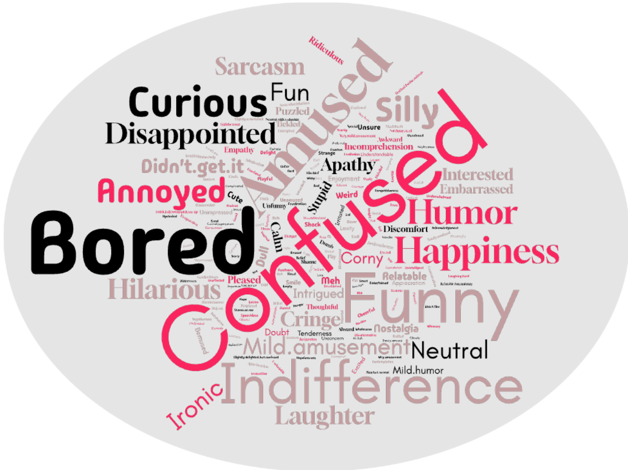


Figure 6.5: "Other" emotions annotated by participants.

(a) Chinese. (b) Mexican. (c) Polish. (d) The U.S..

Figure 6.6: "Other" emotions annotated by participants from each culture.

connotations that cannot be translated into English.

6.4.5 Case studies

To find cartoons where the cultures differed in their responses, we fitted a logistic regression model that predicted funniness (1 for ratings > 3, 0 for ratings $\leqslant 3$) from *1)* cartoon; *2)* interaction of cartoon and participant; *3)* interaction of cartoon, participant, and culture; and *4)* interaction of culture and cartoon. We then used the R package `VEMIRT` (Lyu et al., 2025) to detect differential item functioning in the model. We used the `D2PL_pair_em` function to estimate difference between each culture pair in term of funniness response to each cartoon (β values). Figure 6.7 presents distributions of the estimates.

Using these scores, we found instances where the association between funniness and emotions differed across cultures. For example, the cartoon in Figure 6.8a uses a pun that is based on a PROBLEM IS FORCE metaphor: The word *gravity* in the caption refers to both the seriousness of a situation and literal gravity that makes the brick in the image fall down (and potentially hit the man who is oblivious to the literal gravity of the situation). The β values suggested that

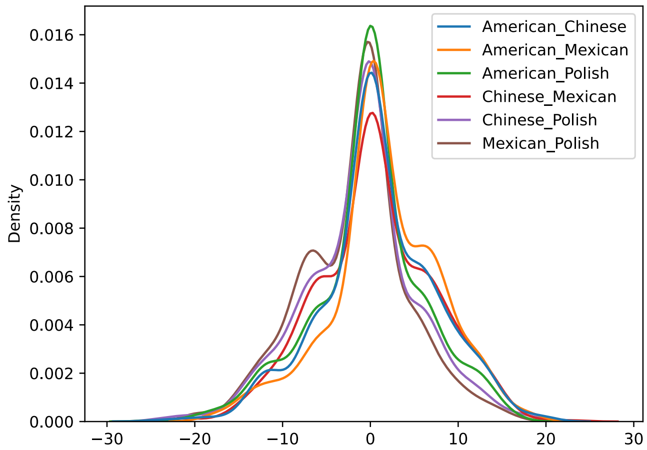
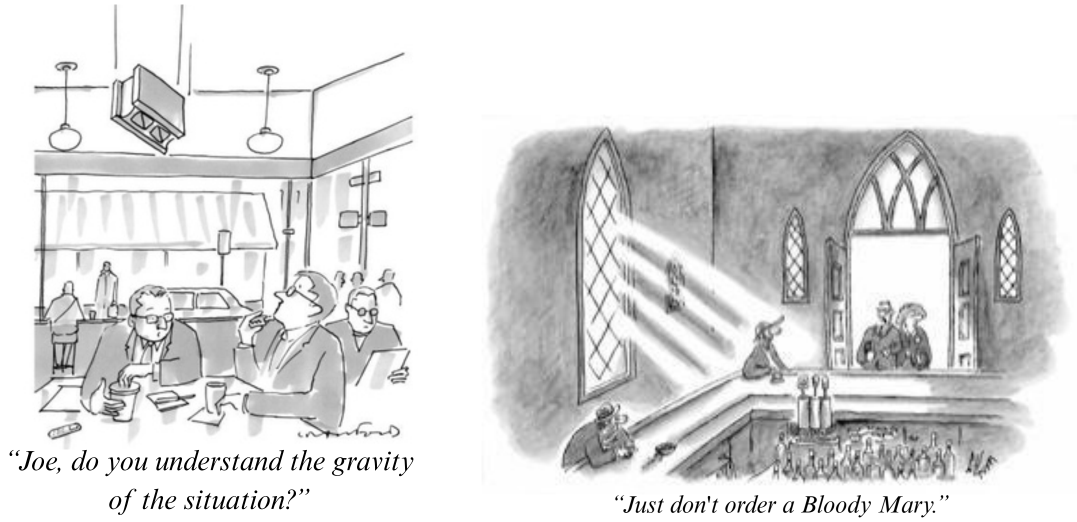


Figure 6.7: Distribution of β values indicating differences between each culture pair in terms of funniness response to the same cartoon. The values were obtained through differential item functioning detection for a logistic regression model predicting funniness response from the interaction of cartoon and culture.

this cartoon was much more likely to be considered funny in Polish and Mexican culture than Chinese culture (β = 11.80 and 10.16 respectively). Emotion responses to this cartoon showed that while most participants from Polish and Mexican culture selected Joy or both Joy and Surprise, half of the participants from Chinese culture selected Fear. On the other hand, their selection of Fear was only associated with low funniness ratings when no other options were selected. When they selected Joy together with Fear, the funniness ratings were > 3. A few participants from Mexican culture also selected Fear, but it was accompanied by Surprise, and the funniness ratings were 3 or above.

The cartoon in Figure 6.8b uses a pun based on a BAR IS A CHURCH metaphor: *Bloody Mary* in this context refers to both the cocktail and the religious figure, Virgin Mary. Participants from Mexican culture had the highest probability to rate this cartoon as funny, and those from Chinese culture the lowest ($\beta = 14.18$ comparing Mexican to Chinese culture); Polish and the U.S. culture had similar probabilities ($\beta = -0.52$). Most participants from Mexican culture selected Joy for this cartoon. Multiple participants from Chinese culture selected Fear; other negative emotions such as Sadness and Disgust were also selected.

Both cartoons in Figure 6.8 use puns, and both puns are based on metaphors. The humor of these two cartoons thus relies on the contrast between the target domain (PROBLEM, DRINK) and the source domain (FORCE, WOMAN) that the metaphors bring out. Participants from Chinese culture may not be familiar with the *gravity* metaphor in Figure 6.8a: While the PROBLEM IS FORCE metaphor is used in Chinese to describe the seriousness of a problem, Chinese equivalence of the word *gravity* emphasizes the meaning of attraction and thus is usually used in



(a) Gravity.

(b) Bloody Mary.

Figure 6.8: Cartoons using puns and metaphors.

a different context. With regard to Figure 6.8b, participants from Chinese culture could be more familiar with the Bloody Mary in British history (it is taught in school) than the cocktail. Even if they knew about Mary's role in the Bible, they might not immediately relate it to the cocktail if they did not know that Virgin Mary was a non-alcoholic version of Bloody Mary.[3] For someone unfamiliar with these metaphors, the contrast between target and source domains might stand out more than someone used to the associations. Thus, the Fear and unfunny responses from participants from Chinese culture might have resulted from a stronger focus on the source domain—the heavy brick falling from the ceiling, and a woman covered in blood.

On the other hand, familiarity is not necessarily associated with high funniness ratings. The cartoon in Figure 6.9a is based on Batman stories. The humor is that Batman, a superhero, has to take the subway to catch Joker. Batman and Joker came from the U.S. culture, but have gained popularity in the other three cultures as well. The subway setting should also be recognizable in all the cultures. Nevertheless, the β values suggested that Chinese and Mexican culture had much higher probabilities to rate the cartoon as funny than the U.S. ($\beta = 13.79/13.46$ comparing Chinese/Mexican to the U.S.) and Polish culture ($\beta = -21.03/-20.70$ comparing Polish to Chinese/Mexican). While participants from Chinese and Mexican culture mostly selected Joy, those from the U.S. and Polish culture mostly selected None.

The humor of the cartoon in Figure 6.9b likely has to do with the "Dad", who seems to be going surfing in a business suit. A parent trying to build a good

[3] I regret the misogyny underlying the naming of these drinks.

"Well, this is just great. The Joker is probably halfway to Brooklyn on the L by now."

(a) Batman.

"Now you're embarrassing me, Dad."

(b) Father and son.

Figure 6.9: Cartoons based in pop culture or universal phenomenon.

relationship with the child should be a universal phenomenon. However, this cartoon showed the highest probability to be considered funny in Chinese culture, followed by Mexican. Both cultures were much more likely to rate the cartoon as funny than the U.S. ($\beta = 19.68/13.32$ comparing Chinese/Mexican to the U.S.) and Polish culture ($\beta = -21.47/-15.11$ compared to Chinese/Mexican). As for emotion categories, most participants from Chinese and Mexican culture selected Joy whereas most from Polish and the U.S. selected None. A few participants from Mexican also selected Sadness and gave the cartoon low funniness ratings, suggesting that they might not appreciate the son's attitute toward his father.

6.5 Discussion and conclusion

In this chapter, we investigated cultural differences in humor appreciation. We collected funniness ratings and emotions data for 800 captioned New Yorker cartoons from four cultures: Chinese, Mexican, Polish, and the U.S.. Using the data, we examined how culture and metaphor use were associated with funniness ratings and emotions, as well as how funniness ratings were associated with emotion categories. We also conducted qualitative analysis on cartoons where the cultures showed differences in their responses. We found a significant effect of both culture and metaphor use on funniness ratings and emotions. Specifically, Mexican and Chinese culture tended to give higher funniness ratings and were more likely to select emotions that were not "None" than Polish and the U.S. culture. Cartoons involving humorous multimodal metaphor use were more likely to receive higher funniness ratings and emotional responses than the non-metaphorical ones in all cultures.

With regard to emotion categories, all cultures showed higher probabilities to select Joy, Surprise, and None than the other categories. Joy and Surprise were associated with high funniness ratings and None low funniness ratings across all cultures. Nevertheless, we also found subtle differences between the cultures. For example, the U.S. culture had a higher probability to select "Other" than the other cultures, and it provided a wider variety of "Other" answers than the other cultures. This was probably because verbal expression of emotions in English was the easiest and the most natural for participants from the U.S. culture. We also found participants from Mexican culture resorting to Spanish words to name emotions outside of the given options.

On the other hand, Chinese culture had a higher probability of selecting Fear than the other cultures; Fear also had a stronger association with low funniness ratings in Chinese culture than Mexican and the U.S. culture. Through case studies we found that negative emotions such as Fear could be caused by an unpleasant mental image that was brought out by a metaphor that the receiver was unfamiliar with. We also demonstrated that familiarity was not necessarily related with high funniness ratings—it would at least partially depend on one's opinion on that specific scenario.

Future directions. Our findings suggest that our current framework, which includes funniness, emotions, and metaphor use, successfully captures cultural differences in humor appreciation, and therefore serves as a solid basis for an evaluation framework of LLMs' cultural alignment in humor processing. As a next step, we will investigate whether there are other factors that help to explain the cultural similarities and differences observed in this study. For example, one may wonder how much the similarities and differences can be explained by Hofstede's cultural dimension framework. The connection will open doors to forming an evaluation framework that can be integrated with existing evaluation frameworks of cultural alignment in LLMs. The joint framework will provide richer context for diagnosing misalignment in humor processing—whether it is due to misaligned values/beliefs or something else.

When asking LLMs to predict the funniness of a cartoon and elicited emotions, we will also consider using open-ended questions or letting the models to explain their answers. As Kabir et al. (2025) have pointed out, closed-style questionnaires can lead to unstable model performance and do not provide sufficient information for understanding the models' cultural alignment. Considering the intricacies of the relationship between funniness and emotional responses that are shown in our study, detailed model outputs will likely be essential for us to measure how well a model aligns with different cultures in humor processing. It will also provide great insights into why any misalignemnt happens, which will shed light on the right directions to for improving the models.

6.6 Limitations

We set aside potential cultural differences in metaphors in this study. The metaphor annotations used in this study require expert knowledge in metaphor; it was infeasible to ask the participants to provide such annotations. While there is the option to ask metaphor scholars from different cultures to do the annotation, it would require the involvement of multiple scholars from each culture—just like we had multiple participants from each culture for the funniness and emotion annotations. That does not fit the scope of the current study. Nevertheless, our metaphor annotation was done by a Chinese scholar, and it had an F1 score of 0.73 compared to the annotation of a Dutch scholar (Chapter 5). We therefore contend that our annotation is representative of universally recognized metaphors.

Due to restrictions of Prolific, it was impossible within the scope of our study to include participants from Chinese culture who were residing in Taiwan or People's Republic of China when the experiment took place. However, all of our participants from Chinese culture were strictly required to have spent most of their time in Taiwan or People's Republic of China before turning 18. They were thus still deeply connected with Chinese culture.

There were also much less available participants from Chinese, Mexican, and Polish culture as compared to the U.S.. This has been reflected in some differences in age and education level distributions. Nevertheless, we have taken age, sex, and education level into account in our analysis of the relationship between funniness, emotions, and metaphor use.

6.7 Ethical considerations

The experiment was approved by the Ethics Committee for Information Sciences at the University of Amsterdam.

Participants were informed of our additional screening procedure and measures to avoid bots before they took part in the experiment. They were also informed beforehand that the cartoon stimuli may contain explicit or sensitive content. Screened out participants received a compensation of 0.10 GBP; full participation was rewarded on the basis of 9 GBP per hour.

Chapter 7

Conclusions

In this thesis, we sought answers to the question: How well do LLMs understand metaphor and humor? We built datasets and established benchmarks that concern the processing of metaphor and humor in both linguistic and multimodal communication. In this chapter, we revisit our key findings and discuss their implications for future research.

7.1 Dissecting incongruity

First, let us take a look at what lies at the core of both metaphor and humor understanding—incongruity. In Chapter 3, we built the MUNCH dataset, providing apt and inapt paraphrases for ∼3k metaphorical sentences. The paraphrases, both apt and inapt, only alter the metaphorical word in the reference sentence while everything else remains intact. Moreover, we made sure that the replacement words in the apt and inapt paraphrases correspond to the target and source domains of the metaphors respectively. A key motivation behind this design was to test whether and to what extent LLMs were able to map source-domain meaning onto the target domain, thus performing full metaphor understanding. Our experiments revealed that the models often confused the two domains of a metaphor. We also found that the models struggled with generating paraphrases for some of the metaphors in our dataset, resulting in nonsensical or ungrammatical outputs. In other cases, the generated paraphrases have opted for words that imply negligence of certain aspects of the target domain (e.g., implying that a coin has more than two sides by paraphrasing *both sides of the coin* as *both facets of the coin*). Nevertheless, the models showed increases in accuracy rate when they were explicitly told which word in the given sentence is metaphorically used.

In Chapter 4, we proposed a taxonomy of 9 intentions behind metaphor use: Lexicalized metaphor, Artistic use of metaphor, Visualization, Persuasiveness, Explanation, Argumentative metaphor, Social interaction, Humor, and Heuristic reasoning. We annotated ∼1.2k linguisic metaphors according to the taxonomy and tested LLMs' capabilities to infer the intentions behind metaphor use in zero- and few-shot settings. We found that while humorous metaphors were slightly less represented in our dataset, LLMs were better at recognizing a humorous intent than other less represented intentions, such as Heuristic reasoning and Social interaction. This could be explained by the central role of incongruity in humor: Unlike use cases such as Persuasiveness or Argumentative metaphor, where the person producing the metaphor may want their metaphor use to stay under the radar (i.e., to hide the incongruity), humor is one use case where detecting the incongruity is essential. The situation thus appears to be similar to the paraphrasing challenge in Chapter 3: Model performance improved when the incongruity was made more salient.

Chapter 5 dealt with multimodal metaphor use in humorous cartoons. We created the HUMMUS dataset, which consists of 1k New Yorker cartoons with English captions, including ~560 that we marked as involving humorous multimodal metaphor use. For these metaphorical items, we provided detailed annotations with regard to which conceptual metaphors are used, how the source and target domains are represented in text and image, and how the metaphor use contribute to the humorous effect of the cartoon. Our experiments with state-of-the-art MLLMs revealed that the models were prone to false positives when asked whether a given cartoon contains humorous multimodal metaphor use. The models also struggled with identifying and localizing the underlying conceptual metaphors. Our ablation study and error analysis suggested that the models' compromised performance could be attributed to inability to integrate visual and textual information.

Implications and future directions. These studies suggest that LLMs are not good at interpreting metaphors as they are—i.e., mappings between distinctive but relevant domains. We observe this issue in the models' processing of both monomodal and multimodal metaphors. Moreover, the false positive predictions of humorous multimodal metaphor use echos a previous study by Akula et al. (2023), which found that vision models struggled with differentiating metaphorical and symbolic advertisements. Our study points out a broader issue: LLMs struggle with processing incongruities in general.

Successful processing of incongruities requires both world knowledge and the activation of relevant world knowledge in real-time interaction. Our studies suggest that the latter is an area we can work on, as we observed improvement in model performance when it was clearer to the models what they should focus on. However, this strategy is likely infeasible in actual human-AI interactions, where the human user is not necessarily interested in teaching the AI how to understand what they have said. Like what E. B. White famously said, "Explaining a joke is like dissecting a frog. You understand it better but the frog dies in the process." When the user has to guide AI through their creative language use, the conversation deviates from its original purpose.

An alternative solution, therefore, is to either train the model beforehand or in the background when a conversation is ongoing. Both options would need a protocol tailored to the purpose (e.g., to detect metaphor use, to analyze the underlying conceptual metaphor, to infer its intention) and datasets that back up the protocol. One such much-needed protocol, as implied by our findings, is to differentiate different types of incongruity—when it should be treated as metaphor use, when it is humorous but non-metaphorical, when it is humrous metaphor use, among others.

7.2 Communicating humanity

In addition to the explicit differentiation of source and target domains, Chapter 3 also investigated whether and to what extent LLMs and humans have similar preferences when paraphrasing metaphorical words. Our experiments showed that LLMs' interpretation of metaphorical words diverged from humans', attributing lower probabilities to paraphrases created by humans than other words in their vocabulary. While some of such cases could be explained by insufficient reasoning ability with the presence of incongruity, other cases appeared to be simply a matter of preference.

In our study of LLMs' capabilities to infer the intentions behind metaphor use (Chapter 4), we observed increased model accuracies in few-shot settings, albeit being accompanied by a higher percentage of nonsensical model outputs. Additionally, the descriptions for each intention category turned out to be indispensable even with few-shot in-context learning—the models appeared to need a combination of high-level explanation and specific examples to gain better understanding of the task.

On the other hand, our experiments showed that the models generally struggled with intentions that were less represented both in our dataset and in natural discourse. There was a huge gap between the mean F1 scores of GPT-4 in identifying Lexicalized metaphor as compared to Heuristic reasoning or Social interaction. Lexicalized metaphor, which typically uses highly conventionalized metaphors, arguably accounts for the majority of metaphor use in natural discourse (Steen, 2017; Steen et al., 2010a). However, this does not make the recognition of less represented intentions less important. Metaphors for Social interaction have their roots in social and cultural norms, conveying how the speaker views the relationship between themselves and the receiver, as well as how the speaker expects the receiver to respond. In human-AI interactions, therefore, recognizing such intentions could be crucial for the AI to be perceived as friendly and helpful.

In Chapter 6, we examined cultural differences in humor appreciation. We collected 25.6k funniness and emotion annotations for 800 New Yorker cartoons with English captions, covering 4 diverse cultures: Chinese, Mexican, Polish, and the U.S.. Our statistical analyses indicated that both culture and metaphor had significant effects on humor appreciation. Cartoons involving humorous multimodal metaphor use tended to be considered funnier and more likely to evoke emotional reactions than cartoons not involving such metaphors. Nevertheless, the same metaphor could lead to different emotional reactions depending on the receiver's culture.

Implications and future directions. Our findings highlight a concern that current LLMs may appear disconnected with human users in human-AI interaction. They tend to have difficulty engaging with highly intellectual discussions (where metaphors could be used for Heuristic reasoning), or recognizing user expectations (which could be implied by metaphorical expressions). While such circumstances may be sparse compared to, e.g., the occurrence of Lexicalized metaphor, failed attempts would still leave a mark—especially considering the close relationship between metaphor and emotion. The situation is even more challenging when we take cultural differences into account. Ignorance or insufficient modeling of the intricate relationship between culture, metaphor, humor, and emotion can easily lead to offensive

Adding to the complexity of the issue is the implication that in-context examples themselves may not be sufficient for LLMs to infer what is left unspoken and make the right decisions. Explicit instructions may still be needed. However, a set of abstract rules is not the same as what happens in real-time communication; a lot of details are lost.

While communication in this broad sense is different from identifying the two domains of a metaphor, the latter (i.e., understanding incongruity) may still form the backbone of more complex tasks that involve intentions, emotions, and cultural differences. Even for metaphor and humor scholars, the analysis always starts with dissecting the incongruity; LLMs may benefit from this thinking process as well.

Appendix A

Appendix to Chapter 3

A.1 Previous metaphor understanding datasets & tasks

Table A.1 summarizes the differences between MUNCH and previous datasets.

Example (1) is extracted from MPEC. The correct paraphrase, Sentence (1-a), is almost completely different from the original sentence. The two distractor sentences that follow indicate different types of misinterpretation: Sentence (1-b) wrongly interprets the meaning of the original sentence, while the last sentence is based on a literal use of the word *wheels*.

(1) the wheels of justice turn slowly
   a. it might take time but eventually justice prevails
   b. ¿ justice prevails in very little time
   c. ¿ the wheels of a car turn slowly

The MPEC corpus is employed by two metaphor understanding tasks in BIG-Bench (Srivastava et al., 2023). The metaphor-boolean task uses a binary classification setup: Given a pair of sentences, is the second sentence a paraphrase of the first? GPT-2 only reached 0.41 accuracy on this task in a zero-shot scenario. The metaphor-understanding task consists of two subtasks: metaphor to paraphrase, which asks the model to select the correct paraphrase from 4 candidates; and paraphrase to metaphor, which requires the model to distinguish the metaphorical sentence corresponding to a given paraphrase from 3 other metaphors. GPT-2 large performed poorly on both subtasks: In a zero-shot scenario, the model gave 0.27 accuracy on the metaphor-to-paraphrase task, and 0.67 accuracy on the paraphrase-to-metaphor task.

The metaphor-literal pairs in the NewsMet dataset was created with the help of LLMs. Each news headline has a verb considered as the focus word. They first passed the headlines with the focus words masked to ALBERT (Lan et al.,

	Metaphor		Correct		Distractor	
	n	Length	n	Type	n	Type
MPEC	192	9 ± 4	218	$s \rightarrow s$	526	mixed
NewsMet	791	12 ± 3	791	$w \rightarrow w$	0	-
IMPLI	913	16 ± 10	1032	$w \rightarrow p$	281	context change
FLUTE	1500	11 ± 5	1500	$p \rightarrow p$	1500	opposite meaning
MiQA	150	8 ± 2	150	$s \rightarrow s$	150	context change
Fig-QA	10256	9 ± 3	10256	$s \rightarrow s$	10256	opposite meaning
MUNCH	2953	26 ± 15	10261	$w \rightarrow w$	1492	paraphrase

Table A.1: Differences between MUNCH and previous datasets that provide paraphrases for metaphors: MPEC (Bizzoni and Lappin, 2018; github.com/yuri-bizzoni/Metaphor-Paraphrase), NewsMet (Joseph et al., 2023; https://github.com/AxleBlaze3/NewsMet_Metaphor_Dataset/tree/main), IMPLI (Stowe et al., 2022; github.com/UKPLab/acl2022-impli), FLUTE (Chakrabarty et al., 2022b; https://github.com/tuhinjubcse/model-in-the-loop-fig-lang), MiQA (Comșa et al., 2022), and Fig-QA(Liu et al., 2022b; https://github.com/nightingal3/Fig-QA/tree/master). We present their differences regarding number of metaphor samples, mean ± SD length of the metaphor samples (measured by number of words), number of correct paraphrases, the part of a metaphor sample that is replaced to create correct paraphrases (s=sentence, p=phrase, w=word), number of distractors, and distractor type. The numbers are calculated from the datasets available on GitHub. Note that our dataset is much more extensive than the previous ones.

2020) to obtain the first 200 words that can replace the focus word. These 200 words were then passed to a metaphor detector to obtain the top-6 metaphorical and top-6 literal candidates. Human annotators then identified the best literal counterparts for metaphorical focus words and the best metaphorical counterpart for literal focus words.

In the IMPLI example (2), the correct paraphrase (2-a) uses a phrase, *paid for*, to explain the metaphorically used word *absorbed* in the original sentence. The distractor, on the other hand, is based on the literal meaning of *absorbed*. Fine-tuned RoBERTa base and RoBERTa large achieved high accuracies (> 0.8) on labelling these metaphor-paraphrase and metaphor-distractor sentence pairs.

(2) he absorbed the costs for the accident
  a. he paid for the costs for the accident
  b. ¿ he absorbed the sunlight after the accident

Example (3) is extracted from the FLUTE dataset; included in the parentheses are explanations for the paraphrase and the contradict respectively. Contrary to the MUNCH dataset, the authors aimed at paraphrases that use more than one word to replace a metaphorically used word. Note that sentence (3-b) is more of a direct contradiction of the reference metaphor than the paraphrase, as it preserves the metaphorically used word *louder*. The difference between the contradict and the reference metaphor may thus be easier to detect as compared to a contradict that is more similar to the paraphrase (e.g., *Actions are not more important than words*).

(3) Actions speak louder than words.
  a. Actions are more important than words. (This phrase is used to say that what someone does is more important than what they say.)
  b. ¿ Actions are not louder than words. (The metaphor suggests that deeds or actions are more important than words, while the contradiction suggests that words are more important than deeds or actions.)

As example (4) shows, each metaphorical premise in the MiQA dataset is paired with a literal premise exemplifying literal use of the metaphorical word; the dataset also includes implications (the text in parenthesis) of the metaphorical and literal premises:

(4)
  a. I <u>see</u> what you mean (I understand you)
  b. I <u>see</u> what you are pointing at (My eyes are working well)

Comșa et al. (2022) set up 2 binary-choice tasks using the MiQA dataset: *1)* Given a metaphorical premise, select the correct implication; *2)* given an implication, select the corresponding premise. They also set up a generative task: Given a metaphorical premise, answer whether it implies the literal conclusion. LLMs performed well on these tasks.

The Fig-QA dataset provides similes of opposite meanings as well as their implications (given in parentheses):

(5)
  a. The meteor was as bright as New York City (The meteor was very bright)
  b. The meteor was as bright as coal (The meteor was not bright at all)

The binary-choice task is similar to MiQA: Given a metaphorical premise, select the correct implication. They also develop a generative task which prompt models to generate implications freely. Liu et al. (2022b) found these tasks challenging for LLMs in zero-shot settings.

1	The summer's sprawl begins to be oppressive at this stage in the year and trigger fingers are itching to snip back overgrown mallows, clear out the mildewing foliage of golden rod and reduce the overpowering bulk of bullyboy ground cover.
2	The red and green of the Aztec necklace links it compositionally with the indigenous plants to the "south" of the painting, the pink colonial-style dress tonally blending with the skyscrapers to the "north".
3	Nine out of 10 are routine calls, many of which could be carried out by mini cabs.
4	This example assumes that a sympathy for motorists with overwhelm any tendency to logical analysis.
5	There were, in fact, about a score.
6	Mrs Bottomley is convinced the Tory victory provides the opportunity to entrench the reforms — and to give doctors, nurses and managers the confidence to make them work.
7	Thus, as with biological theories, crime is seen as pathological (a disease), as something to be looked at from the medical point of view.
8	"So you've decided to put in an appearance?"
9	He was in there twice, at a Wimpole Street number and again at an address in Mill Hill: Rufus H. Fletcher, MB, MRCP.
10	Once again he backtracks and assumes a larger unity in which conflict takes place.
11	no I'm alright Ann, I mean, feel a bit ba ah I mean I'm sorry I do have to buy a feel a bit of, I feel a bit dizzy you know as if I
12	Mick said to me last night, he said to me you can never fit not used to it, but
13	Now if he doesn't get the economy right he's gon na end up with egg on his face and
14	That take me nearly all the er
15	As this is been shared by lines int it?
16	Well seven nines, well ee er, it takes you so long
17	Take what you want and leave the rest, your mother'll get rid of it.

Table A.2: Sentences that did not receive single-word substitutions in the crowdsourcing task.

A.2 Crowdsourcing task

The participant information sheet, which was presented to the crowd workers prior to the consent form, has a section dedicated to potential disadvantages and risks involved in participating in the study—

> The sentences you will paraphrase were from a wide range of sources, including newspapers, fiction, and dialogues. You may occasionally encounter violence or taboo topics (e.g., war, crime, sex), as well as potentially disturbing opinions.
>
> If you are concerned, you do not have to give consent; you can also withdraw anytime during the experiment.

The information sheet also explains how data collected from the study will be used. The workers were informed that their participation would remain confidential, that their response would be anonymized, and that the data would be made open access at the end of the study.

The annotation guidelines are shown in Figure A.1. The trial sentence is provided in Example (6), where *introduce* is the metaphorical word to be interpreted. Our final list of acceptable answers includes: *address*, *advance*, *clarify*, *convey*, *cover*, *define*, *describe*, *discuss*, *elucidate*, *establish*, *explain*, *mention*, *present*, *propose*, *reveal*, *share*, *show*, *state*, *submit*, *suggest*, *teach*, *unveil*.

(6) I shall now introduce the concept of an elementary charge, 1.6 × 10 -19 C, carried by an elementary particle called the electron.

Table A.2 presents the 17 sentences for which none of the crowd workers were able to provide single-word substitutions for the metaphorical words. These are mainly highly conventionalized metaphors, for which it is usually difficult to find an alternative expression. There are also cases where the target word is part of a multi-part word (e.g., *carry out*, *point of view*) or a phrase (e.g., *put in an appearance*, *get rid of*). These seem to be an oversight of the VUA corpus: According to the MIPVU method the corpus employs, a multi-part word or phrase should be considered as a single annotation unit. Nonetheless, as there were no suitable way to filter them out automatically, such cases were included in the crowdsourcing task.

A.3 Guidelines for inapt paraphrase annotation

Thank you for taking part in this annotation task. I will send you 200 sentences that need your annotation, split into 20 surveys (10 sentences in each), so that it will be easier for you to navigate.

In this document I will explain the two annotation steps, namely identifying the more basic meaning and selecting good-enough inapt paraphrases. I use multiple choice questions to prompt your annotations; there is also a comment box for each sentence (at the end of all the questions for that sentence), in case you have anything that needs to be expressed about the sentence or your annotation. If you have any questions along the way, please feel free to contact me.

A.3.1 Identify the more basic meaning

Each sentence has a highlighted word, which we call the target word. The first question provides you with all the senses of the target word extracted from WordNet; your job is to see whether you could find one or more senses that are more basic than the word's contextual sense. In essence, you are asked to perform the contextual sense and basic sense identification steps of the MIPVU procedure, but with WordNet in the place of the Macmillan Dictionary.

You can choose more than one sense, as WordNet employs fine-grained sense distinctions, and multiple senses may qualify as more basic.

If none of the listed senses are more basic (for example, when you believe the target word is used non-metaphorically), please select "None of the above".

If you find a sentence too difficult to comprehend, or the target word's contextual meaning unclear without further context, you can say so in the comment box and skip the sentence.

A.3.2 Select inapt paraphrases

When a sense is selected, you will see a list of words related to that sense, each word being followed by a candidate paraphrase, which uses the related word to replace the target word in the original sentence. Please read through each sentence and select the ones that are good-enough *inapt* paraphrases of the original sentence. If multiple senses are selected in the first step, please go through the options for each selected sense; if no additional question appears when you select a sense, it means this sense does not have related words in WordNet, and you are done with the annotation of this sense.

A good-enough inapt paraphrase should meet the following requirements:

1. It is different from the original sentence in meaning.

2. It indicates that a more basic sense is mistaken as the contextual sense of the target word.

3. It is grammatically acceptable.

I further explain these requirements below.

Requirement 1: Different meanings. Consider the original sentence (7-a) and a candidate paraphrase (7-b). While Sentence (7-a) clearly refers to Paula's emotions, Sentence (7-b) presents some different images: Either Paula was protected (by sandbags or metaphorical sandbags) while repeating something dangerous, or what she repeated irritated someone and that person hit her hard with a sandbag. Since the two sentences invoke different images, Sentence (7-b) meets the first requirement and can be further considered for inapt paraphrase annotation (in fact, it also meets the other two requirements and should be marked out as a good-enough inapt paraphrase).

(7) a. Paula repeated, stunned.
b. Paula repeated, sandbagged.

Sentence (7-b) is also ambiguous and can be interpreted in different ways. Such ambiguous sentences are always considered inapt paraphrases, even if one of the interpretations does correspond to the original sentence–they do not necessarily convey the meaning of the original sentence.

Requirement 2: Wrong sense. A good-enough inapt paraphrase tells us that a more basic sense might have been assigned to the target word (by using a word related to the more basic sense to replace the target word), instead of the contextual sense. *Sandbag* is related to the hitting sense of *stun*; the resulting sentence (7-b) is thus an inapt paraphrase of (7-a). In the example below, however, *communication* is not necessarily related to the physical sense of *sign*. Sentence (8-b) thus should *not* be selected as a good-enough inapt paraphrase of Sentence (8-a), although it meets the first requirement of conveying a different meaning.

(8) a. The one thing they do not do is to re-examine the original for the tell-tale signs of forgery.
b. The one thing they do not do is to re-examine the original for the tell-tale communications of forgery.

Requirement 3: Grammar. We focus on semantic differences in this study, so ungrammatical candidate paraphrases should be ruled out. Sentence (9-b) is ungrammatical as *cover* is a transitive verb and should not be followed by a preposition. It thus should *not* be selected as a good-enough inapt paraphrase for (9-a).

(9) a. But the most striking thing about Bagehot's essay on Peel, in the light of the last full week of this election campaign, is that it simply does not apply to Major at all.
b. But the most striking thing about Bagehot's essay on Peel, in the light of the last full week of this election campaign, is that it simply

does not cover to Major at all.

Overall Please keep in mind that the candidate paraphrases were generated automatically by replacing a target word with a word related to a random sense of the former. They are not provided by humans with an attempt to paraphrase the original sentences, so please do not try to justify them (that is, to find a reason why an English speaker would paraphrase the original sentence like that).

A.4 Model evaluation details

We accessed the LLaMA models through Hugging Face; the queries used ~880 GPU hours. Our GPT-3.5 queries through the OpenAI API cost ~255 USD.

We provide all the prompts used in this study: three prompts for each condition of the paraphrase judgement tasks, including word judgement (Table A.3) and sentence judgement (Table A.4); and three prompts for the paraphrase generation task (Figure A.2).

A.5 Error analysis details

Table A.5 summarizes the novelty scores of the metaphor samples that receive correct versus incorrect answers from the models in the two paraphrase tasks. Table A.6 and A.7 show model accuracies in different genres and for different POS of the metaphorical word respectively. The statistics are based on the best performance of each model. For the paraphrase judgement task, this means the *Metaphor-Word* condition of *Word-judgement* for LLaMA-13B, using the third prompt (see Table A.3); the *Metaphor-Word* condition of *Sentence-judgement* for LLaMA-30B, using the second prompt (see Table A.4); the *Implicit* condition of *Word-judgement* for GPT-3.5, using the third prompt. In paraphrase generation, the LLaMA models achieve their respective best performance when given the first prompt (see Table A.2); for GPT-3.5, it is the second prompt.

Instructions

Each trial gives you a sentence with a target word, for example:

- The artist **captured** her perfectly.

Your task is to paraphrase the given sentence by substituting the target word with another word. We will provide you with the original sentence with the target word removed, so you will just need to fill in the blank:

- The artist _____ her perfectly.

Some trials may provide (much) longer or shorter sentences, but there will always be only one target word in each sentence.

What basic rules should I follow?

Your paraphrase should always be apt: You should be able to use your paraphrase in real life to express the meaning of the original sentence. For the sentence above, we consider the following apt paraphrases:

- The artist **depicted** her perfectly.
- The artist **portrayed** her perfectly.

As you can see, **the substitution should be a single word**: There should be no whitespace in your substitution.

Please also use the correct word form: The target word *captured* should be replaced by a verb in its past tense. If you replace *captured* with *depicts* instead of *depicted*, for example, your paraphrase will be describing a present instead of a past event.

- The artist **depicts** her perfectly. (The event being described is shifted to the present.)
- The artist **depict** her perfectly. (Ungrammatical paraphrases are always inapt.)

Can I use a dictionary?

Yes, you can use dictionaries, thesauruses, or any other resources to help finish the task.

Do I simply look for synonyms?

It depends; please always read through your paraphrase to check whether your synonym fits the context.

Synonyms could render inapt paraphrases as well. For the above example, a thesaurus would list *imprison* as a synonym of *capture*, but substituting *captured* with *imprisoned* would change the sentence's meaning:

- The artist **imprisoned** her perfectly.

Describe seems to be the right synonym, but to use it in your paraphrase, you would need to add more context, which is **not allowed** in this task:

- The artist **described** her perfectly. (The artist talked about her?)
- The artist **described** her perfectly in the picture.

Can I use the same substitution for the same target word?

You may encounter the same target word multiple times; we encourage you to find the most suitable paraphrase for each case. You can, of course, reuse a substitution if you believe that is the best option.

What if I can't find an apt paraphrase?

There is a comment box at the end of each trial. Please use the space to provide your reasons when you could not find an apt paraphrase. A *very short* explanation will do, for example:

> *Original sentence*: It's the first time in his career he hasn't come out on **top**.
> *Your explanation*: You'd need to remove "on" as well, i.e. "he hasn't come out as the best".

Please therefore do not feel pressured to fill in a blank—with the target word itself, a random word, "N/A", etc.—when you believe the target word is impossible to paraphrase given our requirements.

You can also leave comments in those boxes when you have found an apt paraphrase, but this is entirely optional.

Figure A.1: Instructions for the paraphrasing task.

Implicit	
1	Choose the word(s) that can replace the highlighted word in the given sentence without changing the meaning of the sentence. Sentence: `{metaphor_sample}` Option A: `{candidate_substitution_word_1}` Option A: `{candidate_substitution_word_2}` Option C: Both Option A and Option B Option D: Neither Option A nor Option B Correct answer: Option
2	Select words that can replace the highlighted word in the given sentence without altering the sentence's meaning. [. . .]
3	Which of the given options can replace the highlighted word in the given sentence without altering the sentence's meaning? [. . .]
Metaphor-Sent	
1	Choose the word(s) that can replace the highlighted word in the given metaphorical sentence without changing the meaning of the sentence. [. . .]
2	Select words that can replace the highlighted word in the given metaphorical sentence without altering the sentence's meaning. [. . .]
3	Which of the given options can replace the highlighted word in the given metaphorical sentence without altering the sentence's meaning? [. . .]
Metaphor-Word	
1	Choose the word(s) that can replace the highlighted metaphorically used word in the given sentence without changing the meaning of the sentence. [. . .]
2	Select words that can replace the highlighted metaphorically used word in the given sentence without altering the sentence's meaning. [. . .]
3	Which of the given options can replace the highlighted metaphorically used word in the given sentence without altering the sentence's meaning? [. . .]

Table A.3: Prompts for the word judgement task.

Implicit	
1	Choose the correct paraphrase(s) for the given sentence. Sentence: `{metaphor_sample}` Option A: `{candidate_paraphrase_1}` Option A: `{candidate_paraphrase_2}` Option C: Both Option A and Option B Option D: Neither Option A nor Option B Correct answer: Option
2	Select sentences that paraphrase the given sentence. [...]
3	Select sentences that are semantically equivalent to the following sentence. [...]
Metaphor-Sent	
1	Choose the correct paraphrase(s) for the given metaphorical sentence. [...]
2	Select sentences that paraphrase the given metaphorical sentence. [...]
3	Select sentences that are semantically equivalent to the following metaphorical sentence. [...]
Metaphor-Word	
1	You are given a sentence where the highlighted word is metaphorically used. Choose the correct paraphrase(s) for the given sentence. [...]
2	Given a sentence where the highlighted word is metaphorically used, select sentences that paraphrase this sentence. [...]
3	Given a sentence where the highlighted word is metaphorically used, select sentences that are semantically equivalent to this sentence. [...]

Table A.4: Prompts for the sentence judgement task.

1 Paraphrase the given sentence by substituting the highlighted word with another word. The substitution should be a single word.
Sentence: No golden light *bathed* the red brick of the house.

For LLaMA	For GPT
Paraphrase: No golden light *[blank]* the red brick of the house. [blank] should be “___	Paraphrase: No golden light *___* the red brick of the house.

2 Use a single word to replace the highlighted word in the given sentence, so that the new sentence and the given sentence mean the same thing.
Sentence: [...]
New sentence: [...]

3 Given a sentence with a highlighted word, replace this word with a different word to make a paraphrase.
Sentence: [...]
Paraphrase: [...]

Figure A.2: Prompts for the paraphrase generation task. The underscores (_) denote the place where models are asked to provide their answers: The LLaMA models append answer after the left quotation mark (“) while GPT-3.5 inserts answer between the two asterisks ($*$). The underscores themselves are not part of the prompts.

	Judgement	**Generation**
LLaMA-13B	0.07 / 0.07	0.07 / 0.06
LLaMA-30B	0.05 / 0.08	0.06 / 0.06
GPT-3.5	0.06 / 0.08	0.04 / 0.07

Table A.5: Mean novelty scores of metaphor samples for which each model gives correct/incorrect answers when it achieves its respective highest performance in the paraphrase judgement and paraphrase generation tasks. All standard deviations are 0.20 ± 0.01. Differences that are statistically significant are highlighted.

	ACPROSE	NEWS	FICTION	CONVRSN
Judgement				
LLaMA-13B	0.44	0.47	0.47	-
LLaMA-30B	0.37	0.33	0.24	-
GPT-3.5	0.34	0.36	0.32	-
Generation				
LLaMA-13B	0.15	0.17	0.21	0.13
LLaMA-30B	0.34	0.37	0.37	0.32
GPT-3.5	0.45	0.41	0.40	0.40

Table A.6: Model accuracy in different genres when the models achieve their best performance in the paraphrase judgement and paraphrase generation tasks. The metaphor samples for the paraphrase judgement task do not cover the conversation genre. Statistically significant difference between the highest and lowest accuracies on the same row is highlighted.

	N	**V**	**A**	**R**
Judgement				
LLaMA-13B	0.44	0.47	0.46	0.70
LLaMA-30B	0.34	0.32	0.30	0.30
GPT-3.5	0.38	0.29	0.31	0.30
Generation				
LLaMA-13B	0.18	0.16	0.15	0.13
LLaMA-30B	0.37	0.36	0.32	0.37
GPT-3.5	0.44	0.41	0.40	0.52

Table A.7: Model accuracy per POS of the metaphorical word (**N**oun, **V**erb, **A**djective, and adve**R**b) when each model achieves its best performance in the paraphrase judgement and paraphrase generation tasks. Statistically significant difference between the highest and lowest accuracies on the same row is highlighted. Accuracies for adverb metaphors in the paraphrase judgement task are disregarded as the task only includes 10 adverb samples.

Appendix B

Appendix to Chapter 4

B.1 The annotation guidelines

In this task, you are asked to annotate the intentions behind direct and indirect metaphors. For each sentence you are presented with, please annotate the text delimited by and . For instance, in the sentence "Usually the slightest whisper travelled like jungle drums through the world of fashion" you should annotate the word "drums", following the steps that are detailed below.

- **Step 1**: decide if the metaphoric expression could be avoided.
  If there are (literal) paraphrases that would convey roughly the same message in the given context, please continue the annotation and proceed with Step 2. If you cannot think of any paraphrase that avoids the metaphor and would work just fine, then mark the metaphor as *Lexicalized metaphor* and skip Step 2.

- **Step 2**: select categories from the taxonomy of intentions.
  In this step, you are asked to select a possible intention behind the metaphor you are analyzing. The list of categories that you should use is the following one: *Artistic metaphor*, *Visualization*, *Persuasiveness*, *Explanation*, *Argumentative metaphor*, *Social interaction*, *Humour*, *Heuristic reasoning.* If you think that more intentions might play a role, feel free to select multiple categories–up to a maximum of 3.

B.1.1 Explanation

Lexicalized metaphors. To discriminate between lexicalized metaphors and other metaphors, try to think about the subject matter (the Topic) of the metaphor. If the metaphor is just the most common way to talk about the Topic, then mark it as *Lexicalized metaphor*. On the other hand, if the metaphor could be avoided, and the intended message could be expressed in a different way, then the metaphor is not lexicalized. Consider the following examples:

(1) a. Do you follow?
b. Usually the slightest whisper travelled like jungle drums through the world of fashion.

(1-a) is an example of a lexicalized metaphor. The speaker is asking the hearer if they are "following" (most likely) their words. This simply reflects the way in which we generally conceptualize discourse, namely in spatial terms (e.g., as a path).

On the other hand, the metaphor in (1-b) is not lexicalized. The noun "drum" is not commonly used to talk about fashion. One could express the intended message through the following paraphrase "Usually the slightest whisper spread very fast and loud though the world of fashion".

Intention categories. For Step 2, try to think of which communicative goals the metaphor might accomplish better than its paraphrases. To decide which intention(s) to select, refer to the following overview of the taxonomic categories. Each item is provided with its description and some paradigmatic examples.

1. **Artistic metaphor**: These metaphors are used to predicate at once a whole set of features of the Topic. These features need not to be all clearly determined in advance. Ultimately, the intention is to stimulate the receiver's creative interpretation.

   - To her, the long summer days had stretched ahead, world without end.
   - Amaldi dodged the American invitation, perhaps because (with Rome liberated) Fermi's mantle in physics had fallen on his young shoulders and there were younger minds to teach.
   - The summer's sprawl begins to be oppressive at this stage in the year and trigger fingers are itching to snip back overgrown mallows, clear out the mildewing foliage of golden rod and reduce the overpowering bulk of bullyboy ground cover.

2. **Visualization**: The utterer might resort to a metaphor whose Vehicle (i.e., the conventional referent) is easier to visualize than the Topic (the contextual referent). Typically, this happens when the latter belongs to an abstract domain or when the audience is not familiar with it. The intention is to help the receiver to form an intuitive representation of the Topic.

   - Relief surged through her like a physical infusion of new blood.
   - And beyond, green grass and geraniums like splashes of blood.

- The results are terse and sharply etched, like the best line drawings.

3. **Persuasiveness**: Using the metaphor to refer to the Topic, the author gives it a non-neutral connotation, which is not motivated on explicit grounds. The intention is for the audience to adopt the utterer's positive or negative attitude towards the Topic.

    - The ramshackle Whitley Council negotiating machinery is the other reason why the ambulance workers have lost out.
    - America may have changed Presidents a year ago, but the fiscal ticket remains as impenetrable as ever.
    - An atmosphere poisoned by mistrust.

4. **Explanation**: These metaphors are used for didactic purposes. The intention is to explain a new or already familiar concept to the addressee.

    - Canals within the algae stand out as rods in this kind of preservation, which is common in Ordovician rocks.
    - Thus one can and must say, that each fight is the singularisation of all the circumstances of the social whole in movement and that by this singularisation, it incarnates the enveloping totalization which the historical process is.
    - The ego-identity of that person is shaped by these choices.

5. **Argumentative metaphor**: These metaphors are part of explicit arguments intended by the author to convince the audience of a certain claim. The intention is to support the argument, to make it more compelling for the addressee.

    - The effect is rather like an extended advertisement for Marlboro Lights.
    - There was already a rather perfunctory air to the Queen's visit three years ago, as if it were just a required coda to her tour of China.
    - But the villages are dying, becoming suburbs or dormitories where few people work but many sleep.

6. **Social interaction**: These metaphors focus on interpersonal relations, group or cultural conventions and the like. The intention is to create or strengthen some bond between producer and receiver.

    - But I'm starting to think that everything's a turn-off for you, doll .

- Smoking heroin ("chasing the dragon") was one feature of the upsurge.
- Political correctness, just as we suspected, will be perfectly grey .

7. **Humour**: The intention is to entertain the addressee, to be funny. Metaphoric language is exploited for its divertive effects, which would go missing in literal paraphrases.
   - Not sure of the music policy, but the name sounds like the ingredients of a takeaway from a less salubrious Chinese.
   - From there, like a buzzard in its eyrie, he would make forays round the US and abroad in spite of his advanced age.
   - It 's my life which is about to go down the plughole.
8. **Heuristic reasoning**: The intention is to provide an interpretative model for a scientific theory, a work of art, etc. The metaphoric expression is used to organize the addressee's conceptualization of the Topic, based on their prior knowledge about another domain.
   - It is her body as the canvas her appearance as art.
   - It is as if it is walking through a minefield.
   - At the moment, history is made without being known (l'histoire se fait sans se connaître); history constitutes, we might say today, a political unconscious.

B.1.2 Example

Here below is one example annotated following the guidelines.

(2) Allan Ahlberg says: "In the past, a lot of children's books seemed to be the work of talented illustrators whose pictures looked brilliant framed in a gallery, but when you tried to read the book, there was nothing there, because the words started as a coat-hanger to hang pictures on." (VUA)

Step 1. This sentence from a news fragment is about old children's books. The author highlights the characteristic of these books of focusing more on the quality of the illustrations, rather than on the narration. The words that make up the story are metaphorically compared to coat-hangers. The utterer invites us to think of the relation between the illustrations and the words as the one existing between a coat and a coat-hanger. The latter is just instrumental, it has no purpose or value in itself which is independent of the former. Through the metaphor, the author predicates these features of the words in the children's books. The same message could have been conveyed in a literal way, along the following lines: "the words had no value in themselves, they were just instrumental for the illustrations". Thus, the output of Step 1 is that the metaphor is *not lexicalized* and we may move on to Step 2.

Step 2. The metaphoric expression is used in this case to explain the way in which illustrations and words are related in old children's books. The author invites the addressees to understand this relation in terms of the more familiar and concrete relation between coats and hangers. For this reason, the metaphor can be annotated as *Explanation*. It should be noted, however, that also other intentions seem to play a role. For instance, one might read a negative judgment of value in the author's remark. Thus, the annotation could also be *Persuasiveness* or *Argumentative metaphor*, depending on whether some rational justification is given by the utterer to support their judgment.

B.2 Inter-annotator agreement

Our annotation task consists of a multi-label classification with multiple annotators—individual instances can be associated with multiple, non-exclusive intentions. After a brief survey of the available options (Artstein and Poesio, 2008), we opted for a variant of Krippendorff's α as an indicator of the inter-annotator agreement. In particular, we adopted the MASI distance, which is suitable for set-valued labelling tasks such as ours[1]. Out of the 360 MRWs included in the reliability study, 59 distinct items were judged as cases to be excluded by either or both of the two coders. Inter-annotator agreement was computed on the remaining 301 metaphors, where at least one intention was assigned by each annotator. The inter-annotator agreement score was 0.77.

While in his seminal work Krippendorff (1980) sets 0.8 as the minimal requirement for reliable annotation schemes, we believe that 0.77 is a satisfactory result in our specific case for various reasons. First, we can refer to other paradigms in the literature that confirm our value reflects high agreement beyond chance (Green, 1997). Second, the task of inferring communicative intentions behind metaphoric expressions is complex, even for humans, requiring advanced semantic and pragmatic reasoning capacities. Such tasks tend to exhibit lower inter-annotator agreement than many other annotation tasks (e.g., those related to syntax). Third, as detailed in Section 5, in most cases only one intention was assigned per metaphor. Our metric to compute the IAA score is sensible to each element in the set of intentions assigned to metaphorical items. For the cases

[1]The metric has been applied by Passonneau and colleagues to the annotation of co-reference chains (Passonneau, 2004) and Summary Content Units (Passonneau, 2006).

that are currently a full disagreement between the annotators, adding more intentions would increase the probability of marginal agreement, leading eventually to a higher IAA score.

Overall, unlike other classification tasks such as POS tagging, there may be no gold standard for our task: different annotators can indeed interpret the same metaphor in different, yet equally acceptable ways. However, this does not mean that any annotation would be acceptable. What we hope to track with our annotation scheme are the intentions most likely perceived by humans. In other words, there is individual variation in the interpretation of metaphors that we should not expect to erase entirely with our scheme. While this variation does not invalidate the annotation effort, it does make the objective of a near-perfect agreement score unrealistic.

B.3 Corpus analysis: Type

Proponents of DMT maintain that direct metaphors constitute principled examples of deliberate metaphors. Since direct metaphors overtly introduce a referent from a source domain from which a conceptual mapping has to be made (Steen, 2011), they would require the intentional use of metaphor *as* metaphor. On the contrary, given the availability of a contextually relevant non-basic meaning, indirect metaphors would be non-deliberate–though ambiguous cases are possible (Steen, 2023). Thus, information on the type of linguistic metaphor would help to identify deliberate uses in communication. In Table 4.1, we outline the distribution of metaphors in our dataset across the intention categories for all metaphor types.

The results partially align with the claim that direct and indirect metaphors show different tendencies when it comes to their perceived intentions. While all meaningful metaphors are uttered with the minimal intention to communicate, direct metaphors generally correlate with other discourse goals, too. The categories mostly associated to direct metaphors are Visualization, Artistic metaphor, Heuristic reasoning. Indirect metaphors, especially the most conventional ones, are instead judged as lexicalized metaphors.

B.4 Model details

The GPT-4 model is accessed through the OpenAI API, and the two Llama2-Chat models Hugging Face. We employ greedy search for all 3 models. For the two Llama2-Chat models, this is done by setting `do_sample=False` and `num_beams=1`; for the GPT-4 model, `temperature` is set to 0.

Our GPT-4 queries cost $\sim$ 60 USD. Our Llama2-Chat queries used $\sim$ 460 GPU hours (58946:35 SBU).

B.5 Prompts

The prompts for zero-shot and five-shot experiments are presented in Figure B.1 and B.2 respectively. In the zero-shot experiments, the GPT-4 model always starts its answer with the intention category it predicts for the given metaphor. The Llama2-Chat models, on the other hand, need to generate some text (for example, *Based on the provided sentence, I would select the category of . . .*) before providing its prediction. We thus provide the Llama2-Chat models the text they tend to generate at the start of their assistant messages (as part of the prompts), so that the first few new tokens they generate will be the intention category they predict.

Such assistant prompts are determined in the following way: We first take a prompt (system message and user message) that works for GPT-4 and apply it directly to a Llama2-Chat model (the 13B model for the first 2 prompts, and the 70B model for the last one). We do this for 3 different input sentences to obtain the text the model is most likely to produce before providing its prediction. This text is then used as the assistant prompt for both Llama2-Chat models. As shown in Figure B.1, the 3 prompts contain different assistant messages, as we follow the messages that the Llama2-Chat models naturally produce when provided with different system prompts.

B.6 Model performance

As reported in Table 4.2, Llama2-13b-Chat outperformed Llama2-70b-Chat in most few-shot-learning settings. We decided to carry out a more fine-grained analysis of the performance across intention categories to shed some light on this surprising result. Figures B.3a and B.3b show the three models' performance (F_1 scores) in the 5-shot settings with regard to each intention category. Figures B.3c and B.3d show analogous results for the 9-shot settings.

The standard deviation across prompts (indicating model robustness) as well as the F1 score show significant variation across intention categories. For instance, Llama2-70b consistently outperforms Llama2-13b in recognizing Visualization, Persuasiveness, Humour, and Heuristic reasoning, while it surprisingly shows difficulty with Lexicalized metaphors in few-shot settings. The tentative conclusion we can draw is that different models have implicitly learned different aspects of metaphor use. A more detailed analysis of why this is the case–whether it depends on the model training and/or the experimental setup–will be investigated in future work.

```
[SYSTEM]
You are a linguist. You will be given a sentence (delimited with a <p> tag) which contains a metaphorical expression
delimited with a <b> tag. Please annotate the intention behind the metaphor in <b> tag. Your answer should be one of the
following intention categories:

- Argumentative metaphor: The intention is to support an explicit argument, to make it more compelling.
- Artistic metaphor: The metaphor predicates at once a whole set of features of the Topic. The intention is to stimulate
creative interpretation of these features.
- Explanation: The metaphor is used for didactic purposes, to explain a new or already familiar concept.
- Heuristic reasoning: The intention is to provide an interpretative model for a scientific theory, a work of art, etc..
The metaphor organizes the receiver's conceptualization of the Topic.
- Humour: The intention is to entertain, to be funny.
- Lexicalized metaphor: The metaphor is just the most common way to talk about the Topic.
- Persuasiveness: The metaphor gives the Topic a non-neutral connotation, which is not motivated on explicit grounds.
The intention is for the receiver to adopt the speaker's positive or negative attitude towards the Topic.
- Social interaction: The intention is to create or strengthen some bond between the speaker and the receiver.
- Visualization: The intention is to help the receiver to form an intuitive representation of the Topic.

[USER]
<p>But there is a <b>puff</b> of dust on the horizon .</p>

[ASSISTANT]
Based on the sentence you provided, the intention behind the metaphorical expression in the <b> tag is:

* Visualization
```

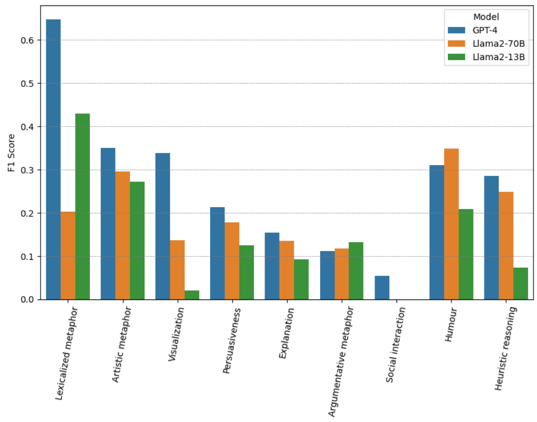
```
[SYSTEM]
You are an expert at metaphor analysis, and you are asked to annotate the intentions behind metaphors. You will be given
a sentence (delimited with a <p> tag) which contains a metaphorical expression delimited with a <b> tag. Please annotate
the intention behind the metaphor in <b> tag by selecting from the following intention categories:

[...]

Please select one and only one category.

[USER]
[...]

[ASSISTANT]
Based on the provided sentence, I would select the category of "[...]
```

```
[SYSTEM]
You are a linguist. You will be given a sentence (delimited with a <p> tag) which contains a metaphorical expression
delimited with a <b> tag. Please annotate the intention behind the metaphor in <b> tag. Your answer should be one of the
following intention categories:

[...]

Please always start your answer with the intention category you choose.

[USER]
[...]

[ASSISTANT]
Category: [...]
```

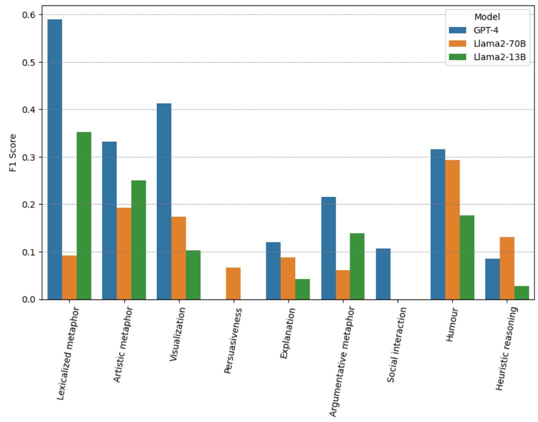
Figure B.1: Prompts for zero-shot experiments and example model output. The same explanations for the intention categories are used in all 3 prompts. Assistant messages in violet are provided to the Llama2-Chat models, so that model outputs always start with the predicted intention category.

```
[SYSTEM]
You are a linguist. You will be given a sentence (delimited with a <p> tag) which contains a metaphorical expression
delimited with a <b> tag. Please annotate the intention behind the metaphor in <b> tag. Your answer should be one of the
following intention categories:

- Argumentative metaphor: The intention is to support an explicit argument, to make it more compelling.
- Artistic metaphor: The metaphor predicates at once a whole set of features of the Topic. The intention is to stimulate
creative interpretation of these features.
- Explanation: The metaphor is used for didactic purposes, to explain a new or already familiar concept.
- Heuristic reasoning: The intention is to provide an interpretative model for a scientific theory, a work of art, etc..
The metaphor organizes the receiver's conceptualization of the Topic.
- Humour: The intention is to entertain, to be funny.
- Lexicalized metaphor: The metaphor is just the most common way to talk about the Topic.
- Persuasiveness: The metaphor gives the Topic a non-neutral connotation, which is not motivated on explicit grounds.
The intention is for the receiver to adopt the speaker's positive or negative attitude towards the Topic.
- Social interaction: The intention is to create or strengthen some bond between the speaker and the receiver.
- Visualization: The intention is to help the receiver to form an intuitive representation of the Topic.
[USER]
<p>Lewis got up and pulled the curtains across with a <b>pettish</b> jerk of his hands .</p>
[ASSISTANT]
Persuasiveness
[USER]
<p>‘ He told me you two were pakiteng , little <b>heifers</b> . ’</p>
[ASSISTANT]
Social interaction
[USER]
<p>This time the front door was open and a <b>swathe</b> of sunlight lay across the red-tiled floor .</p>
[ASSISTANT]
Visualization
[USER]
<p>Ruth could think of few worse fates , but her stomach <b>lurched</b> at the thought of leaving Millfield .</p>
[ASSISTANT]
Visualization
[USER]
<p>The storm was abating rapidly , the evening sky clearing in the west with the golden rays of the setting sun adding a
dying colour to the <b>sullen</b> slate blue water .</p>
[ASSISTANT]
Visualization
[USER]
<p>But there is a <b>puff</b> of dust on the horizon .</p>
[ASSISTANT]
Visualization
```

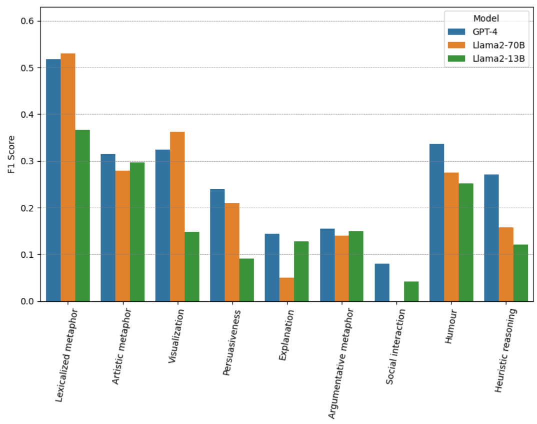
Figure B.2: Prompts for five-shot experiments and example model output. The explanations for intention categories are removed in the 5-shot-short setting.



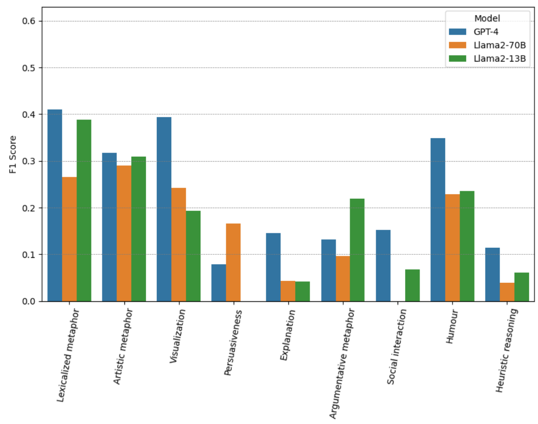
(a) 5-shot



(b) 5-shot-short



(c) 9-shot



(d) 9-shot-short

Figure B.3: Model F_1 score in few shot-settings, averaged across three prompts. Figures (a), (b) show the F_1 score for the 5-shot experiments with and without explanations respectively. Similarly, Figures (c), (d) present the F_1 score for the 9-shot experiments, again comparing results with and without explanations.

Appendix C

Appendix to Chapter 5

C.1 Annotation guidelines

C.1.1 Objective

You will annotate New Yorker cartoons and their captions in terms of humorous metaphor use.

C.1.2 Labelling steps

1. Look at the cartoon image alone (without referring to the caption). Find any incongruities in the image. You can write these down in the `incongruity` field, but this is optional.

2. Copy and paste the caption to `text_hl`.

3. How does the caption resolve the incongruities in the image? If you can answer this question, then you understand the humor of the image-and-caption combination. You can write down your understanding in the `resolution` field, but this is optional.

   (a) If you understand the humor, continue with the annotation;

   (b) Otherwise, select "Discard" for the `is_met` field, assign "1" to `complete` and submit your annotation for this item.

4. Based on your understanding of the humor, consider whether metaphor use (as detailed in "What counts as metaphor") is involved in creating that humorous effect. Any kind of involvement counts. In essence, for each metaphor use you find in the image-and-caption combination, consider whether the humor will still be there when the metaphor use is removed.

   (a) If the answer is an absolute yes, assign "Yes" to `is_met` and continue with the annotation;
   (b) If the answer is an absolute no, assign "No" to `is_met`, "1" to `complete` and submit your annotation;
   (c) If you hesitate, assign "WIDLII" to `is_met` and continue with the annotation.

5. Specify the conceptual metaphor(s) in "X is Y" format. "X" should be the target domain; "Y" should be the source domain. Separate multiple conceptual metaphors with "; " (a semicolon followed by a whitespace).
   (a) To decide how general/specific the conceptual metaphor should be, try to relate the instantiation of the metaphor in the image-and-caption combination to other instantiations of the same metaphor (in any modality), and then find a term that can cover them all.
   (b) You can draw inspiration from this (non-exhaustive) list of conceptual metaphors.
   (c) Finishing the `image_hl` and `text_hl` annotation first (Step 6) could be helpful if you had difficulty pinpointing the two domains.

6. Annotate the metaphor-related image areas (`image_hl`) and text fragments (`text_hl`):
   (a) To annotate metaphor-related image areas, try to assign each object/element of the image to either the target or the source domain of the conceptual metaphors you have annotated. Typically, the majority of the image should belong to the same domain (e.g., the target domain); that one object/element that belongs to the other domain (e.g., the source domain) is the metaphor-related image area. Draw a `image_hl` bounding box for each such "odd" object/element (e.g., if the minority domain is CAT and there are 3 cats in the image, draw 3 bounding boxes, one box for each cat);
   (b) Do the same for the caption. In the `text_hl` field, where you already have the caption copied and pasted, surround each word/phrase/clause that belongs to the "odd" domain with a pair of <i></i> tag. If the entire caption really belongs to a single domain, surround the entire caption in a pair of <i></i> tag. Note that `image_hl` and `text_hl` do *not* have to belong to the same domain.

C.1.3 Mandatory fields

1. Every item should have a `is_met` value.

2. If `is_met` value is "Yes" or "WIDLII", the following fields should also be filled:

    (a) `image_hl`,

    (b) `text_hl` (with <i>highlighted text</i>), and

    (c) `conceptual_metaphor`.

C.1.4 What counts as metaphor

Following the Conceptual Metaphor Theory, we define metaphors as conceptual mappings between two difference domains. Therefore:

1. Personification and zoomorphism are considered metaphor use. Personification is metaphor use with HUMAN as the source domain; zoomorphism is metaphor use with ANIMAL as the source domain.

2. Metonymy within the same conceptual domain does not count as metaphor use.

3. Idioms are not metaphors, unless the underlying cross-domain mapping is strictly required to make sense of the humor.

4. Puns can indicate metaphor use. Whether they are metaphorical depends on whether the two meanings can be attributed to some sort of cross-domain mapping.

C.1.5 Non-metaphorical examples

See Figure C.1.

C.1.6 Metaphorical examples

See Figure C.2.

C.1.7 Bounding box examples

The bounding box should fit tightly around the object of interest—for example, the bear in Figure C.3a. Make sure the entire object is included; Figure C.3b is therefore incorrect. But also avoid including excessive empty space around the object, such as Figure C.3c.

Model Output	Ground Truth	Score
Dogs are people	Animals are humans	0.83
Temperature is music	Temperature is pitch	0.73
Thoughts are bathwater	Psychotherapy is a bath	0.63
Parenting is egg incubation	Human baby is an egg	0.60
Humor is a drug	Preaching is a joke	0.55
Faith is a wedding	Alcohol is god	0.45
Work is a tool	Psychotherapy is a bath	0.35
Slaying a dragon is a task	Modern man is knight	0.26
Ending a relationship is falling off a cliff	Social media is physical world	0.17

Table C.1: Sample model outputs in the Naming benchmark task and their cosine similarity scores compared with ground truth.

C.2 Additional prompts and model outputs

Table C.1, Table C.2, and Figure C.4 provide example model outputs in the ImageLabel, Naming, and Explanation benchmark tasks respectively.

The top-3 alternative Classification prompts (Section 5.5.2) are provided in Table C.3.

The prompt used in Section 5.5.3 to compare model and human performance is provided in Figure C.5.

Model Output	**Ground Truth**	**Score**
Gun	Pistol	0.86
Teddy-bear	Toy bear	0.75
Meteor	Asteroid	0.69
Tree stump	Trunk	0.62
Ball	Egg	0.57
Hat	Cellphone	0.47
Alligator	Customer	0.40
Laptop	Asteroid	0.32
Beer	Human-shaped hole	0.21
Broom	Alien spaceship	0.10

Table C.2: Sample model outputs in the ImageLabel benchmark task and their cosine similarity scores compared with ground truth.

`{image}`
Caption: `{caption}`

Does the humor of the given image-and-caption combination involve metaphor use?

⊞ Answer the question with Yes (i.e., metaphor use is involved) or No (i.e., metaphor use is not involved).

⊟ Answer the question with No or Yes.

⊞ Choose from the following options:
A. The humor of the given image-and-caption combination involves metaphor use.
B. The humor of the given image-and-caption combination does not involve metaphor use.
Answer the question with A or B.

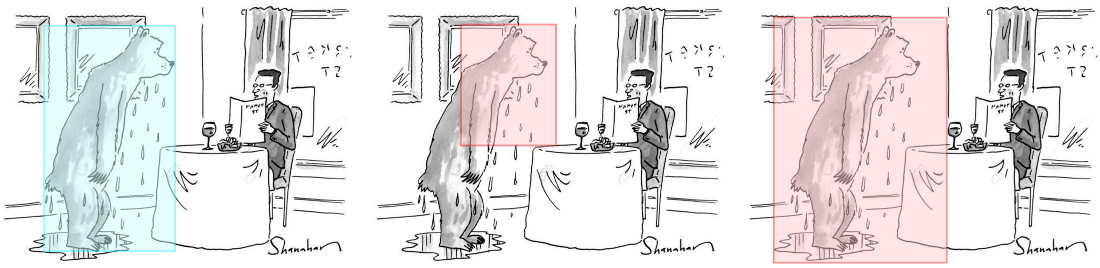
Table C.3: The top-3 prompts in the prompt engineering study for the Classification task. A prompt always starts with an image, a caption, and the question, followed by an instruction about what the model should output: ⊞, ⊟, or ⊞. The order of the two options in ⊞ is randomized for each query.

Image-Caption Pair	**Annotation**
 *"'I'll catch it', was the last I heard."*	? Giant beach ball ! Whoever that was confident to catch the ball is now under it > It's literally a giant ball; someone's literally under it in the pool.
 *"We all deal with divorce in our own way."*	? Man going to a bar with boxer shorts and duck on his head ! He just got a divorce; it's his coping strategy. > The man is literally doing this because he doesn't feel well.
 *"I hope you don't mind. My couch is brand new."*	? Bathing in someone's living room ! The man taking a bath is a guest. The host insists that he does it, otherwise he is not allowed to sit on the brand new couch. > The host literally can't stand it otherwise. It's an exaggeration, but no metaphor use is involved.

Figure C.1: Non-metaphorical examples and annotation of (?) incongruity, (!) resolution, and (>) why they are non-metaphorical.

Humor Understanding & Metaphoricity	**Detailed Metaphor Annotation**
 *"Thanks for checking on the salmon. How's the lamb?"*	
? Bear in a restaurant ! The bear is a waiter > Restaurant waiters are human beings	"Thanks for checking on the <i>salmon</i>. How's the lamb?" ANIMALS ARE HUMAN WORKERS
 *"Three yea's, six ney's, and Anderson is still up in the air on this one."*	
? Someone's defying gravity ! Pun "up in the air" > The pun is based on a metaphor	"Three yea's, six ney's, and Anderson is still <i>up in the air</i> on this one." MAKING A DECISION IS SETTLING ON THE GROUND
 *"Never mind. It looks like they fixed the copier."*	
? Identical men ! The are copied by a copier > Men cannot be copied by a copier	"Never mind. It looks like they fixed the <i>copier</i>." PEOPLE ARE PAPERS

Figure C.2: Metaphorical examples and annotations. The left side shows the image-caption pairs and annotation of (?) incongruity, (!) resolution, and (>) why they are metaphorical. The right side shows annotation of the image, the caption, and underlying conceptual metaphors.



(a) Correct. (b) Too small. (c) Too big.

Figure C.3: Good and bad examples for bounding box annotation.

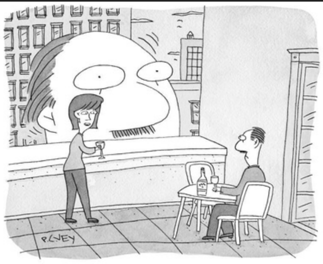
Image-Caption Pair	Ground Truth	Model Output	Rouge Scores
 *"Oh honey look, it's your inflated ego"*	A giant head of the man floats like a balloon outside building, and the woman refers to it as the man's inflated ego. The joke is thus comparing someone's ego to a huge floating balloon.	[LLaVA-NeXT-110B]: The metaphor of an inflated ego as a giant balloon adds humor by exaggerating the concept and making it visually absurd, as the woman points out the comically oversized representation of the man's ego.	0.47/0.229
 *"What we have here is a reptile dysfuntion."*	The joke is based on a metaphorical comparison between a man's penis (erectile dysfunction) and a snake ("reptile dysfunction").	[LLaVA-NeXT-110B]: The metaphor of a snake wrapped around a person's body to represent erectile dysfunction adds humor by applying a familiar concept in an unexpected and absurd context.	0.38/0.087
 *"Let's just say he'll have what you're having."*	The waiter addresses the alligator as if he is addressing a person, and he indicates that the alligator will have a meal, which is very likely to be the customer. The joke thus compares the alligator to a person, and a person to a meal.	[LLaVA-NeXT-8B]: The metaphor suggests the man will have a meal with an alligator, implying a dangerous or absurd situation, which is humorous due to the unexpected twist on a typical dining scenario.	0.34/0.108
 *"So this is your idea of treating me like a queen?"*	The joke uses metaphors that compare a man to a king, and a woman to a queen. It is humorous that the man's idea of treating his wife like a queen is to treat himself like a king.	[LLaVA-NeXT-8B]: The metaphor of a "king" and "queen" in a mundane setting creates a humorous contrast between grandeur and reality, implying a lack of actual royal treatment.	0.31/0.063
 *"Is that a new outfit?"*	The man assumes that the woman is wearing a protective suit for sex appeal. The joke thus uses a metaphor that compares the protective suit to a sexy roleplay costume.	[Qwen2-VL-7B]: The robot's human-like appearance and actions create a humorous contrast, suggesting it's not just a machine but a personified one.	0.15/0.000

Figure C.4: Sample model outputs in the Explanation benchmark task and their ROUGE-1/ROUGE-2 scores as compared to ground truth.

```
[SYSTEM]
You are a linguist with extensive knowledge and experience in metaphor study. You are an expert in the
Conceptual Metaphor Theory and in using the Metaphor Identification Procedure VU University Amsterdam to
manually identify metaphor-related words in language use.

[USER]
{image}
Caption: {caption}

You are an annotator of **humorous metaphor use in New Yorker cartoons**. Each item consists of an image (the
cartoon) and a caption. Follow the steps below carefully and produce structured annotations.

## Steps

1. **Look at the cartoon image (without the caption).**

 * Identify any incongruities in the image. (Optional: put them in the `incongruity` field.)

2. **Read the caption.**

 * Ask: *How does the caption resolve the incongruities in the image?*
 * If you understand the humor, proceed. (Optional: describe it in the `resolution` field.)
 * If you cannot understand the humor, set `is_met = "Discard"` and stop.

3. **Decide if metaphor use is part of the humor.**

 * Ask: *Would the humor still exist if the metaphor use were removed?*

 * If **yes**, set `is_met = "Yes"` and continue.
 * If **no**, set `is_met = "No"` and stop.
 * If unsure, set `is_met = "WIDLII"` and continue.

4. **Identify the conceptual metaphor(s).**

 * Use **"X is Y"** format (target domain = X, source domain = Y).
 * Separate multiple metaphors with `; `.
 * Example: "ARGUMENT is WAR".
 * Choose a level of generality that connects this instance to other possible instances of the same metaphor.

5. **Highlight metaphor-related parts of the image and caption.**

 * `image_hl`: mark bounding boxes around objects/elements that belong to the *minority* domain (the one that
stands out as metaphorical). If multiple, draw multiple boxes. Output should be normalized bounding box
coordinates in **"[top, left, height, width]"** format.
 * `text_hl`: wrap the metaphor-related words/phrases in `<i> ... </i>`. If the entire caption belongs to one
domain, wrap the whole caption.

## Output Fields

* `is_met`: One of `"Yes"`, `"No"`, `"WIDLII"`, `"Discard"`.
* `conceptual_metaphor`: List of metaphors in "X is Y" format (if `is_met` is `"Yes"` or `"WIDLII"`).
* `image_hl`: Normalized bounding box coordinates in "[top, left, height, width]" format for metaphor-related
areas (if `is_met` is `"Yes"` or `"WIDLII"`).
* `text_hl`: Caption with `<i> ... </i>` around metaphor-related text (if `is_met` is `"Yes"` or `"WIDLII"`).
* `incongruity` (optional).
* `resolution` (optional).

## What counts as metaphor

* **Metaphor definition:** a conceptual mapping between two domains.
* **Personification (X is HUMAN)** and **zoomorphism (X is ANIMAL)** count as metaphor.
* **Metonymy** within the same domain does *not* count.
* **Idioms** are not metaphors unless cross-domain mapping is essential to the humor.
* **Puns** may indicate metaphor; they only count if based on cross-domain mapping.
```

Figure C.5: Prompt for comparing model and human performance.

Appendix D

Appendix to Chapter 6

D.1 Participant details

Sex. We had a balanced amount of female ($N = 649$) and male ($N = 631$) participants overall. As shown in Figure D.1, there were approximately the same number female and male participants from Mexican and the U.S. culture. However, female participants outnumbered male participants in Chinese culture, whereas the opposite was found in Polish culture.

Age. Our participants ranged from 18 to 83 years old. The mean age was 34.83 ($SD = 11.44$); around 75% of our participants were no more than 41 years old. As shown in Figure D.2, the U.S. culture contributed the most to the more senior samples. The average participants from Polish ($M = 29.64$, $SD = 7.82$) and Mexican (30.65, $SD = 6.96$) culture were much younger than those from the U.S. culture ($M = 44.96$, $SD = 12.82$). Chinese culture also provided more senior samples than Polish and Mexican culture, although it also had much larger number of younger participants ($M = 34.09$, $SD = 10.11$) than the U.S. culture.

Education. Prolific provided 8 categories regarding participants' highest level of education completed. All categories were found in our data. Figure D.3 shows the percentage of each category, with categories "No formal qualifications" and "Don't know / not applicable" combined into one, as very few participants were from these two categories ($N = 6$ and $N = 3$ respectively).

Overall, those who had received higher education constitute the majority of the recruited participants; 41.6% of our participants had obtained an undergraduate degree. A similar distribution in this regard was observed within the individual cultures. Nevertheless, a much higher percentage of participants from Polish (25.9%) and the U.S. (21.9%) culture had completed a high school diploma or A-levels as compared to Mexican (8.7%) and Chinese (5.6%). On the other

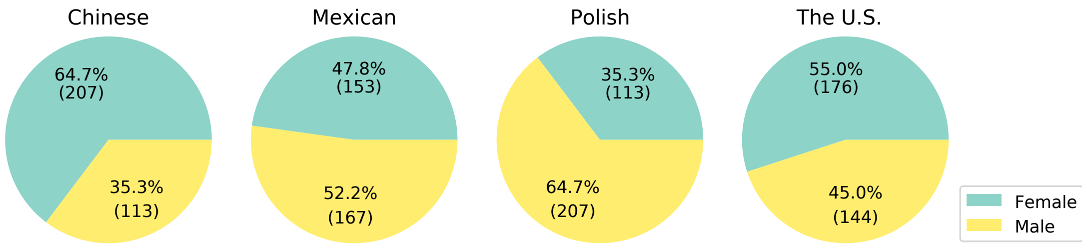
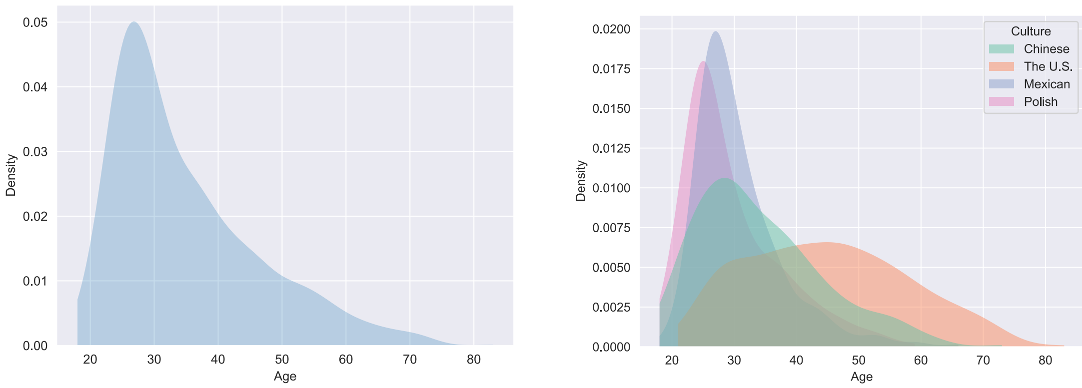


Figure D.1: Participants' sex distribution per culture.



(a) Overall (b) Per culture

Figure D.2: Participants' age distribution.

hand, Chinese culture had the highest percentage of participants with a graduate degree (40.6%), as well as a much higher percentage of participants with a doctorate degree (12.8%) than the other three cultures (lower than 3%).

D.2 Additional data analysis results

ANOVA summary for the mixed effects logistic regression model of each emotion category is provided in Table D.1. Estimated marginal means by culture is provided in Table D.2.

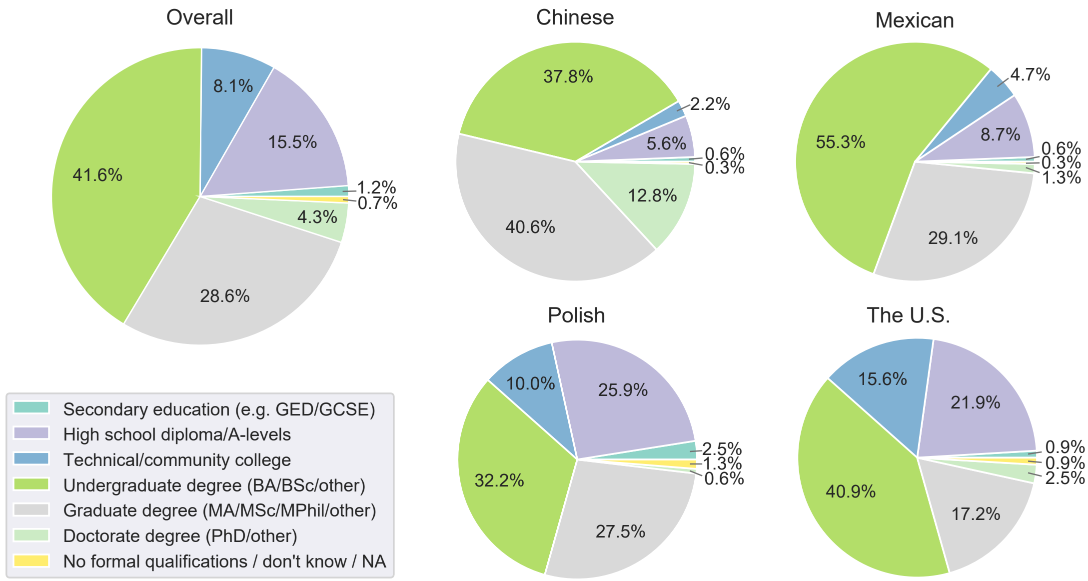


Figure D.3: Participants' highest level of education completed.

	Culture		Age		Sex		Education	
Emotion	χ^2	p	χ^2	p	χ^2	p	χ^2	p
Joy	68.28	< .001	4.66	.031	0.29	.590	13.36	.064
Sadness	61.43	< .001	2.71	.100	0.23	.636	13.94	.052
Fear	74.46	< .001	0.00	1.000	7.74	.005	2.75	.916
Disgust	18.66	< .001	0.90	.343	0.06	.801	10.25	.175
Surprise	19.34	< .001	0.81	.369	10.70	.001	8.35	.303
Anger	28.88	< .001	0.00	1.000	0.47	.491	0.00	1.000
Other	18.30	< .001	5.53	.019	12.12	< .001	7.43	.386
None	61.74	< .001	1.59	.207	0.03	.864	8.48	.292

Table D.1: ANOVA summary of mixed effects (culture, age, sex, highest level of education completed) from mixed effects logistic regression model of each emotion with random intercepts for participants and cartoons. Degrees of freedom for culture, age, sex, and education level are 3, 1, 1, and 7 respectively.

				95% CI	
Emotion	Culture	Estimate	SE	Lower	Upper
Joy	Chinese	0.385	0.034	0.320	0.454
	Mexican	0.483	0.037	0.412	0.555
	Polish	0.364	0.033	0.302	0.431
	The U.S.	0.261	0.029	0.208	0.321
Sadness	Chinese	0.040	0.009	0.025	0.063
	Mexican	0.024	0.006	0.015	0.038
	Polish	0.015	0.004	0.009	0.024
	The U.S.	0.013	0.003	0.008	0.022
Fear	Chinese	0.012	0.004	0.006	0.024
	Mexican	0.004	0.001	0.002	0.008
	Polish	0.003	9.807×10^{-4}	0.001	0.005
	The U.S.	0.002	7.422×10^{-4}	8.059×10^{-4}	0.004
Disgust	Chinese	0.008	0.233	2.220×10^{-16}	1.000
	Mexican	0.008	0.239	2.220×10^{-16}	1.000
	Polish	0.005	0.152	2.220×10^{-16}	1.000
	The U.S.	0.005	0.152	2.220×10^{-16}	1.000
Surprise	Chinese	0.111	0.018	0.081	0.150
	Mexican	0.164	0.025	0.122	0.219
	Polish	0.114	0.018	0.084	0.153
	The U.S.	0.105	0.017	0.076	0.144
Anger	Chinese	8.277×10^{-4}	2.031	2.220×10^{-16}	1.000
	Mexican	3.526×10^{-4}	0.865	2.220×10^{-16}	1.000
	Polish	2.476×10^{-4}	0.608	2.220×10^{-16}	1.000
	The U.S.	3.934×10^{-4}	0.965	2.220×10^{-16}	1.000
Other	Chinese	0.008	0.018	9.918×10^{-5}	0.390
	Mexican	0.006	0.014	8.025×10^{-5}	0.342
	Polish	0.005	0.012	6.566×10^{-5}	0.297
	The U.S.	0.013	0.028	1.589×10^{-4}	0.506
None	Chinese	0.155	0.026	0.111	0.213
	Mexican	0.131	0.023	0.092	0.182
	Polish	0.282	0.039	0.213	0.364
	The U.S.	0.267	0.039	0.198	0.350

Table D.2: Estimated marginal means of emotion response by culture, adjusted for age, sex, and highest level of education completed. Highest estimate per emotion category is highlighted.

Bibliography

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23201–23211, June 2023.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.671. URL https://aclanthology.org/2024.acl-long.671/.

Khalid Alnajjar, Mika Hämäläinen, and Shuo Zhang. Ring that bell: A corpus and method for multimodal metaphor detection in videos. In Debanjan Ghosh, Beata Beigman Klebanov, Smaranda Muresan, Anna Feldman, Soujanya Poria, and Tuhin Chakrabarty, editors, *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 24–33, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.flp-1.4. URL https://aclanthology.org/2022.flp-1.4/.

G. E. M. Anscombe. *Intention.* Cambridge, Mass.: Harvard University Press, 1957.

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://aclanthology.org/J08-4004.

Salvatore Attardo. *Linguistic Theories of Humor.* Mouton de Gruyter, 1994.

Salvatore Attardo. Humorous metaphors. In Geert Brône, Kurt Feyaerts, and Tony Veale, editors, *Cognitive Linguistics and Humor Research*, pages 91–110. De Gruyter Mouton, Berlin, München, Boston, 2015. ISBN 9783110346343. doi: doi:10.1515/9783110346343-005. URL https://doi.org/10.1515/9783110346343-005.

Salvatore Attardo, Władysław Chłopicki, and Giovannantonio Forabosco. *6 The Role of Incongruity in Humorous Texts*, pages 105–124. De Gruyter, Berlin, Boston, 2024. ISBN 9783110755770. doi: doi:10.1515/9783110755770-007. URL https://doi.org/10.1515/9783110755770-007.

Lisa Aziz-Zadeh and Antonio Damasio. Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology - Paris*, 102(1-3), 2008. doi: 10.1016/j.jphysparis.2008.03.012.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.45. URL https://aclanthology.org/2023.ijcnlp-main.45/.

Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660, 1999.

James Beattie. *Essays on Poetry and Music: As They Affect the Mind: On Laughter, and Ludicrous Composition: On the Usefulness of Classical Learning*. Printed for E. and C. Dilly and W. Creech, Edinburgh, M.DCC.LXXIX, London, 1779. Retrieved from the Library of Congress, https://www.loc.gov/item/20010861/.

Anke Beger and Olaf Jäkel. The cognitive role of metaphor in teaching science: Examples from physics, chemistry, biology, psychology and philosophy. *Philosophical Inquiries*, 3:89–112, 03 2015.

Beata Beigman Klebanov and Michael Flor. Argumentation-relevant metaphors in test-taker essays. In Ekaterina Shutova, Beata Beigman Klebanov, Joel Tetreault, and Zornitsa Kozareva, editors, *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-0902/.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. A corpus of non-native written English annotated for metaphor. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2014. URL https://aclanthology.org/N18-2014/.

Yuri Bizzoni and Mehdi Ghanimifard. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0911. URL https://www.aclweb.org/anthology/W18-0911.

Yuri Bizzoni and Shalom Lappin. Predicting human metaphor paraphrase judgments with deep neural networks. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors, *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0906. URL https://aclanthology.org/W18-0906/.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. Ets corpus of non-native written english ldc2014t06. https://catalog.ldc.upenn.edu/LDC2014T06, 2014. Philadelphia: Linguistic Data Consortium.

Isabelle Blanchette, Kevin Dunbar, John Hummel, and Richard Marsh. Analogy use in naturalistic settings: The influence of audience, emotion and goals. *Memory and Cognition*, 29(5), 2001. doi: 10.3758/BF03200475.

BNC Consortium. The British National Corpus, 2007. URL http://www.natcorp.ox.ac.uk/. BNC XML Edition.

Brian F. Bowdle and Dedre Gentner. The career of metaphor. *Psychological Review*, 112(1):193–216, 2005. doi: 10.1037/0033-295X.112.1.193.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. Using imageability and topic chaining to locate metaphors in linguistic corpora.

In Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 102–110, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37210-0.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2084.

Bram Bulté and Ayla Rigouts Terryn. Llms and cultural values: the impact of prompt language and explicit cultural framing, 2025. URL https://arxiv.org/abs/2511.03980.

L. Cameron. *Metaphor in Educational Discourse*. Advances in Applied Linguistics. Bloomsbury Publishing, 2003. ISBN 9781441175649. URL https://books.google.nl/books?id=hjvLVbA16r8C.

Noel Carroll. *Visual Metaphor*, chapter 6, pages 189–218. Springer Netherlands, Dordrecht, 1994. ISBN 978-94-015-8315-2. doi: 10.1007/978-94-015-8315-2_6. URL https://doi.org/10.1007/978-94-015-8315-2_6.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606, 2022a. doi: 10.1162/tacl_a_00478. URL https://aclanthology.org/2022.tacl-1.34/.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

7139–7159, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.481. URL https://aclanthology.org/2022.emnlp-main.481/.

Ke Chang, Hao Li, Junzhao Zhang, and Yunfang Wu. Nyk-ms: A well-annotated multi-modal metaphor and sarcasm understanding benchmark on cartoon-caption dataset, 2024. URL https://arxiv.org/abs/2409.01037.

Guo-Hai Chen and Rod A Martin. A comparison of humor styles, coping humor, and mental health between chinese and canadian university students. *Humor: International Journal of Humor Research*, 20(3), 2007. doi: 10.1515/HUMOR.2007.011.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mWVoBz4W0u.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.figlang-1.32.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.141. URL https://aclanthology.org/2021.naacl-main.141/.

A. Ciaramidaro, M. Adenzato, I. Enrici, S. Erk, L. Pia, B.G. Bara, and H. Walter. The intentional network: How the brain reads varieties of intentions. *Neuropsychologia*, 45(13):3105–3113, 2007. ISSN 0028-3932. doi: https://doi.org/10.1016/j.neuropsychologia.2007.05.011. URL https://www.sciencedirect.com/science/article/pii/S0028393207002035.

Alan Cienki and Cornelia Müller. Metaphor, gesture, and thought. In *The Cambridge handbook of metaphor and thought*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511816802.029. URL https://doi.org/10.1017/CBO9780511816802.029.

Francesca MM Citron and Adele E Goldberg. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, 26(11):2585–2595, 2014.

Michael Clark. Humour and incongruity. *Philosophy*, 45(171):20–32, 1970. ISSN 00318191, 1469817X. URL http://www.jstor.org/stable/3749521.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, second edition, 1988.

Ted Cohen. Metaphor and the cultivation of intimacy. *Critical Inquiry*, 5(1): 3–12, 1978. ISSN 00931896, 15397858. URL http://www.jstor.org/stable/1342974.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. MiQA: A benchmark for inference on metaphorical questions. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-short.46. URL https://aclanthology.org/2022.aacl-short.46/.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020. URL https://arxiv.org/abs/2006.13979.

Elizabeth Crawford. Conceptual metaphors of affect. *Emotion Review*, 1(2), 2009. doi: 10.1177/1754073908100438.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. Scientific and creative analogies in pretrained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.153. URL https://aclanthology.org/2022.findings-emnlp.153/.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1227. URL https://www.aclweb.org/anthology/D19-1227.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.figlang-1.31.

Sara Dellantonio, Claudio Mulatti, Luigi Pastore, and Remo Job. Measuring inconsistencies can lead you forward: Imageability and the x-ception theory. *Frontiers in Psychology*, Front. Psychol.(708):1–9, 07 2014. doi: 10.3389/fpsyg.2014.00708.

Rutvik H. Desai. Are metaphors embodied? the neural evidence. *Psychological Research*, 86(8):2417–2433, 2022. ISSN 1430-2772. doi: 10.1007/s00426-021-01604-4. URL https://doi.org/10.1007/s00426-021-01604-4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pages 1412–1424, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1171. URL https://aclanthology.org/D18-1171/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Marta Dynel. Creative metaphor is a birthday cake: Metaphor as the source of humour. *Metaphorik.de*, 17, 01 2009.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

Lynn Fainsilber and Andrew Ortony. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbolic Activity*, 2(4):239–250, 1987. doi: 10.1207/s15327868ms0204_2. URL https://doi.org/10.1207/s15327868ms0204_2.

Andrew H. Farkas, Rebekah L. Trotti, Elizabeth A. Edge, Ling-Yu Huang, Aviva Kasowski, Olivia F. Thomas, Eli Chlan, Maria P. Granros, Kajol K. Patel, and Dean Sabatinelli. Humor and emotion: Quantitative meta analyses of functional neuroimaging studies. *Cortex*, 139:60–72, 2021. ISSN 0010-9452. doi: https://doi.org/10.1016/j.cortex.2021.02.023. URL https://www.sciencedirect.com/science/article/pii/S0010945221000824.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

Charles Forceville. (a)symmetry in metaphor: The importance of extended context. *Poetics Today*, 16(4):677–708, 1995. ISSN 03335372, 15275507. URL http://www.jstor.org/stable/1773369.

Charles Forceville. The identification of target and source in pictorial metaphors. *Journal of Pragmatics*, 34(1):1–14, 2002. ISSN 0378-2166. doi: https://

doi.org/10.1016/S0378-2166(01)00007-8. URL https://www.sciencedirect.com/science/article/pii/S0378216601000078.

Charles Forceville. Visual and multimodal metaphor in film: charting the field. In *Embodied metaphors in film, television, and video games*, pages 17–32. Routledge, 2015.

Charles Forceville. Visual and multimodal metaphor in advertising: Cultural perspectives. *Styles of communication*, 9(2), 2017.

Sigmund Freud. *Jokes and Their Relation to the Unconscious*. Norton, 1960.

Susan R Fussell and Mallie M Moss. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141, 1998.

Susan R Fussell and Mallie M Moss. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press, 2014.

Vittorio Gallese and George Lakoff. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, 22(3-4): 455–479, 2005.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1060. URL https://www.aclweb.org/anthology/D18-1060.

Andrew Gargett and John Barnden. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 21–30, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1403. URL https://www.aclweb.org/anthology/W15-1403.

D. Gentner and A. B Markman. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56, 1997. URL https://doi.org/10.1037/0003-066X.52.1.45.

Raymond W. Gibbs. What do idioms really mean? *Journal of Memory and Language*, 31(4):485–506, 1992. ISSN 0749-596X. doi: https://doi.org/10.1016/0749-596X(92)90025-S. URL https://www.sciencedirect.com/science/article/pii/0749596X9290025S.

Raymond W. Jr Gibbs. *Intentions in the Experience of Meaning*. Cambridge University Press, 1999. ISBN 9780521576307. URL https://books.google.fr/books?id=FD1rVU4LtBgC.

Raymond W. Jr Gibbs. Are "deliberate" metaphors really deliberate? a question of human consciousness and action. *Metaphor and the Social World*, 1(1):26–52, 2011. doi: 10.1075/msw.1.1.03gib.

Sam Glucksberg and Boaz Keysar. Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97(1):3–18, 1990. doi: 10.1037/0033-295X.97.1.3.

Andrew Goatly. *The Language of Metaphors*. Routledge, first edition, 1997. doi: https://doi.org/10.4324/9780203210000.

Alberto Godioli and Władysław Chłopicki. *Humor and Figurative Language*, chapter 8, pages 145–162. De Gruyter, Berlin, Boston, 2024. ISBN 9783110755770. doi: doi:10.1515/9783110755770-009. URL https://doi.org/10.1515/9783110755770-009.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. IlliniMet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.figlang-1.21.

Joseph Grady, Todd Oakley, and Seana Coulson. Blending and metaphor. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 101–124, 1999.

Annette M. Green. Kappa statistics for multiple raters using categorical classifications. 1997. URL https://api.semanticscholar.org/CorpusID:770316.

H. Paul Grice. Meaning. *Philosophical Review*, 66(3):377–388, 1957. doi: 10.2307/2182440.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and Puranen B. World values survey wave 7 (2017-2022) cross-national data-set, 2022. Version: 4.0.0. World Values Survey Association.

Michael Haugh and Kasia M. Jaszczolt. Speaker intentions and intentionality. In Keith Allan and Kasia Jaszczolt, editors, *Cambridge Handbook of Pragmatics*, pages 87–112. Cambridge University Press, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dustin Hellberg. Funny in the bones: The neural interrelation of humor, irony, and metaphor as evolved mental states. *Interdisciplinary Literary Studies*, 20 (3):237–254, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

M. B. Hesse. *Models and Analogies in Science*. Newman history and philosophy of science series. Ind., 1966. ISBN 9780268001827. URL https://books.google.fr/books?id=mZM1AAAAIAAJ.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41/.

Neelam Arjan Hiranandani and Xiao Dong Yue. Humour styles, gelotophobia and self-esteem among chinese and indian university students. *Asian Journal of Social Psychology*, 17(4):319–324, 2014. doi: https://doi.org/10.1111/ajsp.12066. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajsp.12066.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and Organizations: Software of the Mind*. McGraw-Hill, third edition, 2010.

Keith J Holyoak and Dušan Stamenković. Metaphor comprehension: A critical review of theories and evidence. *Psychological bulletin*, 144(6):641, 2018.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110, 2017. doi: 10.1109/CVPR.2017.123.

Francis Hutcheson. *Reflections Upon Laughter, and Remarks on the Fable of the Bees*. R. Urie, 1750.

Arthur M. Jacobs and Annette Kinder. "The Brain Is the Prisoner of Thought": A machine-learning assisted quantitative narrative analysis of literary metaphors for use in neurocognitive poetics. *Metaphor and Symbol*, 32(3):139–160, 2017. doi: 10.1080/10926488.2017.1338015. URL https://doi.org/10.1080/10926488.2017.1338015.

Lalit Jain, Kevin Jamieson, Robert Mankoff, Robert Nowak, and Scott Sievert. The New Yorker cartoon caption contest dataset, 2020. URL https://nextml.github.io/caption-contest-data/.

JASP Team. JASP (Version 0.96.0)[Computer software], 2026. URL https://jasp-stats.org/.

Feng Jiang, Xiao Dong Yue, and Su Lu. Different attitudes toward humor between chinese and american students: Evidence from the implicit association test. *Psychological Reports*, 109(1):99–107, 2011. doi: 10.2466/09.17.21.PR0.109.4.99-107. URL https://doi.org/10.2466/09.17.21.PR0.109.4.99-107. PMID: 22049652.

Tonglin Jiang, Hao Li, and Yubo Hou. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10:123, 2019. doi: 10.3389/fpsyg.2019.00123.

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai. NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10090–10104, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.641. URL https://aclanthology.org/2023.findings-acl.641/.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.2. URL https://aclanthology.org/2025.emnlp-main.2/.

Morris Kalliny, Kevin W. Cruthirds, and Michael S. Minor. Differences between american, egyptian and lebanese humor styles: Implications for international management. *International Journal of Cross Cultural Management*, 6(1):121–134, 2006. doi: 10.1177/1470595806062354. URL https://doi.org/10.1177/1470595806062354.

Hermann Kappelhoff and Cornelia Müller. Embodied meaning construction: Multimodal metaphor and expressive movement in speech, gesture, and feature film. *Metaphor and the social world*, 1(2):121–153, 2011.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

Gillian A. Kirsh and Nicholas A. Kuiper. Positive and negative aspects of sense of humor: Associations with the constructs of individualism and relatedness. *Humor: International Journal of Humor Research*, 16(1), 2003.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101861. URL https://www.sciencedirect.com/science/article/pii/S156625352300177X.

Tetsuta Komatsubara. Framing and metaphor in media discourse: multi-layered metaphorical framings of the COVID-19 pandemic in newspaper articles. In *The Routledge Handbook of Language and Mind Engineering*, chapter 18, pages 274–292. Taylor & Francis, 2024. doi: 10.4324/9781003289746-23.

Zornitsa Kozareva. Multilingual affect polarity and valence prediction in metaphor-rich texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–691, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1067.

Klaus Krippendorff. *Content Analysis: An Introduction to Methodology.* Sage Publications, Inc., Beverly Hills, CA, 1980.

Lisa Kugler and Christof Kuhbandner. That's not funny! – but it should be: effects of humorous emotion regulation on emotional experience and memory. *Frontiers in Psychology*, 6, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.01296. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.01296.

Nicholas A. Kuiper, Melissa Grimshaw, and Catherine Leite. Humor is not always the best medicine: Specific components of sense of humor and psychological well-being. *Humor: International Journal of Humor Research*, 17(1/2), 2004. doi: 10.1515/humr.2004.002. URL https://doi.org/10.1515/humr.2004.002.

George Lakoff. Metaphor and war: The metaphor system used to justify war in the gulf. *Peace Research*, 23:25–32, 1991.

George Lakoff. Mapping the brain's metaphor circuitry: metaphorical thought in everyday reason. *Frontiers in Human Neuroscience*, 8, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00958. URL https://www.frontiersin.org/articles/10.3389/fnhum.2014.00958.

George Lakoff and Mark Johnson. Conceptual metaphor in everyday language. *The Journal of Philosophy*, 77(8):453–486, 1980a. ISSN 0022362X. URL http://www.jstor.org/stable/2025464.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago Press, 1980b.

George Lakoff and Elisabeth Wehling. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. Free Press, New York, 2012.

George Lakoff, Jane Espenson, and Alan Schwartz. Master metaphor list, 1991. URL https://araw.mede.uic.edu/~alansz/metaphor/METAPHORLIST.pdf. Second Draft Copy, University of California, Berkeley.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International conference on machine learning*, pages 1188–1196, 2014.

Jens Lemmens and Victor De Marez. Computational humor modeling: A survey on the state of the art. *ACM Comput. Surv.*, 58(7), January 2026. ISSN 0360-0300. doi: 10.1145/3778357. URL https://doi.org/10.1145/3778357.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. A report on the 2018 VUA metaphor detection shared task. In Beata

Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, editors, *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0907. URL https://aclanthology.org/W18-0907/.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. A report on the 2020 VUA and TOEFL metaphor detection shared task. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.3. URL https://aclanthology.org/2020.figlang-1.3/.

Jiahao Li, Greg Shakhnarovich, and Raymond A. Yeh. Adapting clip for phrase localization without further training, 2022. URL https://arxiv.org/abs/2204.03647.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. FrameBERT: Conceptual metaphor detection with frame embedding learning. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.114. URL https://aclanthology.org/2023.eacl-main.114/.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.476. URL https://aclanthology.org/2022.emnlp-main.476/.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.330. URL https://aclanthology.org/2022.naacl-main.330/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jackson G. Lu. Cultural differences in humor: A systematic review and critique. *Current Opinion in Psychology*, 53:101690, 2023. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2023.101690. URL https://www.sciencedirect.com/science/article/pii/S2352250X23001355.

Weicong Lyu, Yijun Cheng, Jiaying Xiao, He Ren, Ruoyi Zhu, Gongjun Xu, and Chun Wang. *VEMIRT: Variational Expectation Maximization for High-Dimensional IRT Models*, 2025. URL https://MAP-LAB-UW.github.io/VEMIRT. R package version 2.13.

Rui Mao, Chenghua Lin, and Frank Guerin. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1113. URL https://www.aclweb.org/anthology/P18-1113.

Rui Mao, Chenghua Lin, and Frank Guerin. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1378. URL https://www.aclweb.org/anthology/P19-1378.

Rod A. Martin. The Situational Humor Response Questionnaire (SHRQ) and Coping Humor Scale (CHS): A decade of research findings. *HUMOR*, 9(3-4): 251–272, 1996. doi: doi:10.1515/humr.1996.9.3-4.251. URL https://doi.org/10.1515/humr.1996.9.3-4.251.

Rod A. Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of Research in Personality*, 37(1):48–75, 2003. ISSN 0092-6566. doi: https://doi.org/10.1016/S0092-6566(02)00534-2. URL https://www.sciencedirect.com/science/article/pii/S0092656602005342.

Reem I. Masoud, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models using soft prompt tuning, 2025a. URL https://arxiv.org/abs/2503.16094.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE, January 2025b. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.567/.

Rowan Hall Maudslay and Simone Teufel. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.7/.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559, 2005. doi: 10.3758/BF03192726.

Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

Jessica Mesmer-Magnus, David J. Glew, and Chockalingam Viswesvaran. A meta-analysis of positive humor in the workplace. *Journal of Managerial Psychology*, 27(2):155–190, 02 2012. ISSN 0268-3946. doi: 10.1108/02683941211199554. URL https://doi.org/10.1108/02683941211199554.

Gianluca Michelli, Xiaoyu Tong, and Ekaterina Shutova. A framework for annotating and modelling intentions behind metaphor use, 2024. URL https://arxiv.org/abs/2407.03952.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. Metaphor as a medium for emotion: An empirical study. In Claire Gardent, Raffaella Bernardi, and Ivan Titov, editors, *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-2003. URL https://aclanthology.org/S16-2003/.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. Introducing the LCC metaphor datasets. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1668/.

John Morreall. *Taking laughter seriously.* State University of New York, 1983.

John Morreall. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Fall 2024 edition, 2024.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. Learning outside the box: Discourse-level features improve metaphor identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 596–601, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1059. URL https://www.aclweb.org/anthology/N19-1059.

Agnieszka Mykowiecka, Aleksander Wawer, and Malgorzata Marciniak. Detecting figurative word occurrences using recurrent neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 124–127, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0916. URL https://www.aclweb.org/anthology/W18-0916.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

E. Oring. *Engaging Humor*. University of Illinois Press, 2003. ISBN 9780252027864. URL https://books.google.fr/books?id=nGgA7SY1Uq4C.

Andrew Ortony. Why metaphors are necessary and not just nice1. *Educational Theory*, 25:45–53, 04 1975. doi: 10.1111/j.1741-5446.1975.tb00666.x.

Andrew Ortony. Beyond literal similarity. *Psychological Review*, 86(3):161–180, 1979.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76:1–25, 1968. doi: 10.1037/h0025327.

Rebecca Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth Interna-*

tional Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/636_pdf.pdf.

Rebecca J. Passonneau. Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/752.pdf.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Robert Plutchik. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980. ISBN 978-0-12-558701-3. doi: https://doi.org/10.1016/B978-0-12-558701-3.50007-7. URL https://www.sciencedirect.com/science/article/pii/B9780125587013500077.

Emese Pozdena. Cultural differences in humour usage in American and Hungarian TED talks. *International Journal of Humanities and Social Science Invention*, 9 (11):12–21, 2020. doi: 10.35629/7722-0911021221. URL https://www.ijhssi.org/papers/vol9(11)/Ser-2/B0911021221.pdf.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0908. URL https://www.aclweb.org/anthology/W18-0908.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is ChatGPT a general-purpose natural language processing task solver? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.85. URL https://aclanthology.org/2023.emnlp-main.85/.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 475–479, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1076/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Victor Raskin. *Survey of Humor Research*, pages 1–44. Springer Netherlands, Dordrecht, 1984. ISBN 978-94-009-6472-3. doi: 10.1007/978-94-009-6472-3_1. URL https://doi.org/10.1007/978-94-009-6472-3_1.

Carina Rasse, Alexander Onysko, and Francesca M. M. Citron. Conceptual metaphors in poetry interpretation: a psycholinguistic approach. *Language and Cognition*, 12(2):310–342, 2020. doi: 10.1017/langcog.2019.47.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1162. URL https://www.aclweb.org/anthology/D17-1162.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147, June 2018. ISSN 2509-9515. doi: 10.1007/s41701-017-0026-7. URL https://doi.org/10.1007/s41701-017-0026-7.

Paul Ricœur. *La métaphore vive.* Éditions du seuil, 27, rue Jacob, Paris VIe, 1975.

Richard M. Roberts and Roger J. Kreuz. Why do people use figurative language? *Psychological Science*, 5(3):159–163, 1994. doi: 10.1111/j.1467-9280.1994.tb00653.x.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. V-flute: Visual figurative language understanding with textual explanations, 2024. URL https://arxiv.org/abs/2405.01474.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Andrea C. Samson and James J. Gross. Humour as emotion regulation: The differential consequences of negative versus positive humour. *Cognition and Emotion*, 26(2):375–384, 2012. doi: 10.1080/02699931.2011.585069. URL https://doi.org/10.1080/02699931.2011.585069.

Julie Aitken Schermer, Radosław Rogoza, Maria Magdalena Kwiatkowska, Christopher Marcin Kowalski, Sibele Aquino, Rahkman Ardi, Henrietta Bolló, Marija Branković, Razieh Chegeni, Jan Crusius, Violeta Enea, Thi Khanh Ha Truong, Dzintra Iliško, Tomislav Jukić, Emira Kozarević, Gert Kruger, Adil Kurtić, Jens Lange, Kadi Liik, Sadia Malik, Samuel Lins, Agim Mamuti, Laura Martinez-Buelvas, Benjamin Mrkušić, Ginés Navarro-Carrillo, Oscar Oviedo-Trespalacios, Emrah Özsoy, Eva Papazova, Joonha Park, Natalia Pylat, Goran Riđić, Ognjen Riđić, Dženan Skelić, Chee-Seng Tan, Jorge Torres-Marín, Osman Uslu, Tatiana Volkodav, Anna Włodarczyk, and Georg Krammer. Humor styles across 28 countries. *Current Psychology*, 42:16304–16319, 2023.

John R. Searle. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press, 1983. doi: 10.1017/CBO9781139173452.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. Inside jokes: Identifying humorous cartoon captions. In *KDD*, 2015.

Ekaterina Shutova. Automatic metaphor interpretation as a paraphrasing task. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/N10-1147/.

Ekaterina Shutova. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623, 12 2015. ISSN 0891-2017. doi: 10.1162/COLI_a_00233. URL https://doi.org/10.1162/COLI_a_00233.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1020. URL https://www.aclweb.org/anthology/N16-1020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Pradeep Sopory and James Price Dillard. The persuasive effects of metaphor: A meta-analysis. *Human Communication Research*, 28(3):282–419, 2002. doi: 10.1111/j.1468-2958-2002.tb00813.x.

Herbert Spencer. *The physiology of laughter*, chapter 4, pages 194–209. D Appleton & Company, 1875. doi: 10.1037/12203-004.

Rand Spiro, Paul J. Feltovich, Richard Coulson, and Daniel Anderson. Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition. In S. Vosniadou and A. Ortony, editors, *Similarity and analogical reasoning*, pages 498–530. Cambridge University Press, 01 1989.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen

Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael

Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Harshvardhan Srivastava. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 102–107, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.cmcl-1.11. URL https://aclanthology.org/2022.cmcl-1.11/.

Gerard Steen. The contemporary theory of metaphor - now new and improved! *Review of Cognitive Linguistics*, 9(1):26–64, 2011. URL https://doi.org/10.1075/rcl.9.1.03ste.

Gerard Steen. Deliberate metaphor affords conscious metaphorical cognition. *Cognitive Semiotics*, 5(1-2):179–197, 2014. doi: doi:10.1515/cogsem.2013.5.12.179. URL https://doi.org/10.1515/cogsem.2013.5.12.179.

Gerard Steen. Thinking by metaphor, fast and slow: Deliberate metaphor theory offers a new model for metaphor and its comprehension. *Frontiers in Psychology*, 14, 09 2023. doi: 10.3389/fpsyg.2023.1242888.

Gerard J. Steen. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol*, 23(4):213–241, 2008. doi: 10.1080/10926480802426753.

Gerard J. Steen. Deliberate Metaphor Theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics*, 14(1):1–24, 2017. doi: 10.1515/ip-2017-0001.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, and Tina Krennmayr. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796, 2010a. doi: 10.1515/cogl.2010.024.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU.* John Benjamins, 2010b.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. IMPLI: Investigating NLI models' performance on figurative language. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.369. URL https://aclanthology.org/2022.acl-long.369/.

Tomek Strzalkowski, Samira Shaikh, Kit Cho, George Aaron Broadwell, Laurie Feldman, Sarah Taylor, Boris Yamrom, Ting Liu, Ignacio Cases, Yuliya

Peshkova, and Kyle Elliot. Computing affect in metaphors. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 42–51, Baltimore, MD, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2306. URL https://www.aclweb.org/anthology/W14-2306.

Chang Su, Shuman Huang, and Yijiang Chen. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311, 2017. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2016.09.030. URL http://www.sciencedirect.com/science/article/pii/S0925231216310475.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.figlang-1.4.

Jerry M. Suls. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey H. Goldstein and Paul E. McGhee, editors, *The Psychology of Humor*, pages 81–100. Academic Press, San Diego, 1972. ISBN 978-0-12-288950-9. doi: 10.1016/B978-0-12-288950-9.50010-9. URL https://www.sciencedirect.com/science/article/pii/B9780122889509500109.

Krishnkant Swarnkar and Anil Kumar Singh. Di-LSTM contrast : A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0914. URL https://www.aclweb.org/anthology/W18-0914.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. Learning to evaluate humor in memes based on the incongruity theory. In Xianchao Wu, Peiying Ruan, Sheng Li, and Yi Dong, editors, *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.cai-1.9/.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1404. URL https://www.aclweb.org/anthology/W15-1404.

The BNC Baby. The BNC Baby, version 2, 2005. URL http://www.natcorp.ox.ac.uk/. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Paul H. Thibodeau and Lera Boroditsky. Metaphors we think with: The role of metaphor in reasoning. *PLOS ONE*, 6(2):1–11, 02 2011. doi: 10.1371/journal.pone.0016782. URL https://doi.org/10.1371/journal.pone.0016782.

Paul H. Thibodeau and Lera Boroditsky. Natural language metaphors covertly influence reasoning. *PLOS ONE*, 8(1):1–7, 01 2013. doi: 10.1371/journal.pone.0052961. URL https://doi.org/10.1371/journal.pone.0052961.

Xiaoyu Tong. Metaphor paraphrasing and word-sense disambiguation: toward a new approach to automated metaphor processin. Master's thesis, Universitetit van Amsterdam, the Netherlands, 2021. URL https://scripties.uba.uva.nl/download?fid=681664.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.372. URL https://aclanthology.org/2021.naacl-main.372/.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. Metaphor understanding challenge dataset for LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.193. URL https://aclanthology.org/2024.acl-long.193/.

Roger Tourangeau and Lance Rips. Interpreting and evaluating metaphors. *Journal of Memory and Language*, 30(4):452–472, 1991. ISSN 0749-596X. doi: https://doi.org/10.1016/0749-596X(91)90016-D. URL http://www.sciencedirect.com/science/article/pii/0749596X9190016D.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. doi: 10.1109/CVPR.2018.00675.

Villy Tsakona. Language and image interaction in cartoons: Towards a multimodal theory of humor. *Journal of Pragmatics*, 41(6):1171–1188, 2009. ISSN 0378-2166. doi: 10.1016/j.pragma.2008.12.003. URL https://www.sciencedirect.com/science/article/pii/S0378216608003056.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1024. URL https://www.aclweb.org/anthology/P14-1024.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D11-1063.

L. van Poppel. The study of metaphor in argumentation theory. *Argumentation*, 35:177–208, 2021.

Stephanie K. van Stee. Meta-analysis of the persuasive effects of metaphorical vs. literal messages. *Communication Studies*, 69(5):545–566, 2018. doi: 10.1080/10510974.2018.1457553.

J. H. M. Wagemans. Analyzing metaphor in argumentative discourse. *Rivista Italiana Di Filosofia Del Linguaggio*, 2016.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. Metaphor detection with effective context denoising. In Andreas Vlachos and

Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.102. URL https://aclanthology.org/2023.eacl-main.102/.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models, 2023.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0913. URL https://www.aclweb.org/anthology/W18-0913.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2887–2899, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532019. URL https://doi.org/10.1145/3477495.3532019.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023.

Xiao Dong Yue. The chinese ambivalence to humor: views from undergraduates in hong kong and china. *Humor: International Journal of Humor Research*, 24 (4), 2011.

Xiao Dong Yue and Neelam Arjan Hiranandani. Perception of humorists: A cross-cultural study of undergraduates in hong kong, hangzhou, and vancouver1. *Comprehensive Psychology*, 3:07.17.CP.3.19, 2014. doi: 10.2466/07.17.CP.3.19. URL https://journals.sagepub.com/doi/abs/10.2466/07.17.CP.3.19.

Xiaodong Yue, Feng Jiang, Su Lu, and Neelam Hiranandani. To be or not to be humorous? cross cultural perspectives on humor. *Frontiers in Psychology*, 7, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.01495. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01495.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. MultiMET: A multimodal dataset for metaphor understanding. In Chengqing

Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.249. URL https://aclanthology.org/2021.acl-long.249/.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. MultiCMET: A novel Chinese benchmark for understanding multimodal metaphor. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6141–6154, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.409. URL https://aclanthology.org/2023.findings-emnlp.409/.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, Robert Mankoff, and Robert Nowak. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning, 2024. URL https://arxiv.org/abs/2406.10522.

Shenglong Zhang and Ying Liu. Metaphor detection via linguistics enhanced Siamese network. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.364/.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert, 2023.

Samenvatting

Deze scriptie onderzoekt de mogelijkheden van grote taalmodellen (LLM's) met betrekking tot de verwerking van metaforen en humor. Metaforen en humor zijn onmisbare onderdelen van menselijke cognitie en communicatie, maar ze kunnen uitdagingen vormen voor LLM's. Naarmate mensen over de hele wereld steeds vaker LLMs gebruiken, is het belangrijk te weten hoe goed de modellen metaforen en humor begrijpen en hoe ze verbeterd kunnen worden. Deze scriptie onderzoekt de verwerkingscapaciteiten van LLM's voor metaforen en humor in de volgende opzichten:

Parafrasering van taalkundige metaforen. Voortbouwend op eerder onderzoek naar automatische interpretatie van metaforen, beschouw ik het begrijpen van metaforen als een parafraseertaak. Ik selecteer zinnen met metaforen uit het VU Amsterdam Metaphor Corpus (VUA) en creëer een dataset met meer dan 10.000 handmatig gemaakte, passende parafrases voor deze metaforische zinnen. Daarnaast construeer ik handmatig 1.500 <referentiezin, parafrase 1, parafrase 2>-voorbeelden met ongeschikte parafrases; De paren van passende en ongepaste parafrases leggen de verschillen vast tussen een contextuele, metaforische interpretatie en een letterlijke interpretatie van de termen. Ik evalueer taalmodellen (LLM's) op twee taken: het genereren van parafrases (met behulp van alle passende parafrases in de dataset) en het beoordelen van parafrases (een meerkeuzetaak gebaseerd op de paren van passende en ongepaste parafrases). De experimenten tonen aan dat LLM's moeite hebben met het correct parafraseren van taalkundige metaforen.

Intenties achter metafoorgebruik. Ik ontwikkel samen met anderen een taxonomie met negen categorieën van mogelijke intenties achter metafoorgebruik. Op basis van deze taxonomie annoteer ik samen met anderen een dataset met intentie-annotaties voor 1000 metaforische zinnen uit VUA. Vervolgens gebruik ik deze dataset om de capaciteiten van LLM's te onderzoeken om de intenties achter taalkundige metaforen te voorspellen. Onze experimenten met zero- en few-shot voorbeelden laten zien dat het afleiden van de intenties achter taalkundige metaforen een uitdagende taak is voor de huidige LLM's.

Humoristisch multimodaal metafoorgebruik. Wat betreft het gebruik van multimodale metaforen, focus ik op de wisselwerking tussen metafoor en humor in multimodale communicatie: de twee fenomenen hebben gemeenschappelijke kenmerken en metafoor is een van de meest voorkomende humoristische mechanismen. Geïnspireerd door de incongruiteitstheorie van humor, de conceptuele metafoortheorie en het annotatieschema achter VUA, ontwikkel ik een nieuw annotatieschema voor humoristisch multimodaal metafoorgebruik in beeld-bijschriftparen. Ik annoteer 1000 beeld-bijschriftparen uit het corpus van de New Yorker Caption Contest. Op basis van deze dataset ontwerp ik een reeks taken om het vermogen van multimodale LLM's te testen om humoristisch multimodaal metafoorgebruik te detecteren en te begrijpen. De experimenten tonen aan dat huidige LLM's nog steeds moeite hebben met het verwerken van humoristische multimodale metaforen, met name wat betreft de integratie van visuele en tekstuele informatie.

Culturele verschillen in humorwaardering. Humor vertoont zowel universaliteit als culturele variatie. Het vermogen om aan te sluiten bij het 'gevoel voor humor' van individuele culturen is belangrijk in de interactie tussen mens en AI. Als eerste stap naar een raamwerk voor het evalueren van de culturele aansluiting van LLM's bij de verwerking van humor, beoogt deze studie menselijke basiswaarden vast te stellen die culturele verschillen in humorwaardering vertegenwoordigen. Specifiek onderzoek ik de associatie tussen humor, metafoor en emotie, en hoe deze verschilt per cultuur. Hiertoe rekruteer ik deelnemers uit de Chinese, Mexicaanse, Poolse en Amerikaanse cultuur en verzamel ik 25.600 beoordelingen van de grappigheid en annotaties van emotionele reacties voor 800 cartoons uit The New Yorker met bijschriften, waaronder 482 met gedetailleerde annotaties van humoristisch multimodale metaforen. Mijn kwantitatieve en kwalitatieve analyses onthullen zowel algemene patronen als de complexiteit van wat in verschillende culturen als humoristisch wordt beschouwd, hoe humorwaardering samenhangt met emotionele reacties en hoe metaforen de humorwaardering kunnen beïnvloeden, afhankelijk van de cultuur.

Abstract

This thesis investigates the capabilities of large language models (LLMs) with regard to the processing of metaphor and humor. Metaphor and humor are indispensable parts of human cognition and communication, yet they can pose challenges to LLMs. As LLMs enter the lives of people around the world, it is important to know how well the models understand metaphor and humor, and how they can be improved. This thesis studies LLMs' metaphor and humor processing capabilities in the following respects—

Paraphrasing of linguistic metaphors. Following prior research on automatic metaphor interpretation, I frame metaphor understanding as a paraphrasing task. I sample sentences containing metaphor use from the VU Amsterdam Metaphor Corpus (VUA) and create a dataset that includes over 10,000 manually created apt paraphrases for these metaphorical sentences. I also manually construct ~1,500 $\langle$ reference sentence, paraphrase 1, paraphrase 2 $\rangle$ instances that involve inapt paraphrases; the apt-inapt paraphrase pairs capture differences between a contextual, metaphorical interpretation and a literal interpretation of the vehicle terms. I evaluate LLMs on two tasks: paraphrase generation (using all apt paraphrases in the dataset) and paraphrase judgement (a multiple choice task based on the apt-inapt pairs). The experiments show that LLMs face challenges in correctly paraphrasing linguistic metaphors.

Intentions behind metaphor use. I co-develop a taxonomy that contains nine categories of possible intentions behind metaphor use. Based on the taxonomy, I co-annotate a dataset that provides intentions annotation for ~1,000 metaphorical sentences sampled from VUA. I then use the dataset to examine LLMs' capabilities to predict the intentions behind linguistic metaphors. Our zero- and few-shot experiments show that inferring the intentions behind linguistic metaphors is a challenging task for current LLMs.

Humorous multimodal metaphor use. With regard to multimodal metaphor use, I focus on the interplay between metaphor and humor in multimodal communication: The two phenomena share common grounds, and metaphor is one of the most common humorous mechanisms. Taking inspiration from the Incongruity Theory of humor, Conceptual Metaphor Theory and the annotation scheme behind VUA, I develop a novel annotation scheme for humorous multimodal metaphor use in image-caption pairs. I annotate 1,000 image-caption pairs sampled from the New Yorker Caption Contest corpus. Based on the dataset, I design a set of tasks to test multimodal LLMs' ability to detect and understand humorous multimodal metaphor use. The experiments show that current LLMs still struggle with processing humorous multimodal metaphors, particularly with regard to integrating visual and textual information.

Cultural differences in humor appreciation. Humor exhibits both universality and cultural variance. The ability to align with the "sense of humor" of individual cultures is important in human-AI interaction. As a first step towards a framework for evaluating LLMs' cultural alignment in humor processing, this study aims to establish human baselines representing cultural differences in humor appreciation. Specifically, I consider the association between humor, metaphor, and emotion, and how it differs across culture. To this end, I recruit participants from Chinese, Mexican, Polish, and the U.S. culture, and collect 25,600 funniness ratings and annotation of emotional reactions for 800 captioned New Yorker cartoons, including 482 with detailed annotation of humorous multimodal metaphor use. My quantitative and qualitative analyses reveal both general patterns and intricacies of what is considered humorous in different cultures, how humor appreciation is associated with emotional reactions, and how metaphor may affect humor appreciation depending on the culture.

Titles in the ILLC Dissertation Series:

ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism

ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions

ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners

ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories

ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction

ILLC DS-2021-14: **Álvaro Piedrafita**
On Span Programs and Quantum Algorithms

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions

ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes

ILLC DS-2022-05: **Lwenn Bussière-Carae**
No means No! Speech Acts in Conflict

ILLC DS-2022-06: **Emma Mojet**
Observing Disciplines: Data Practices In and Between Disciplines in the 19th and Early 20th Centuries

ILLC DS-2022-07: **Freek Gerrit Witteveen**
Quantum information theory and many-body physics

ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity

ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms

ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories

ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

ILLC DS-2023-05: **Dean McHugh**
Causation and Modality: Models and Meanings

ILLC DS-2023-06: **Jialiang Yan**
Monotonicity in Intensional Contexts: Weakening and: Pragmatic Effects under Modals and Attitudes

ILLC DS-2023-07: **Yiyan Wang**
Collective Agency: From Philosophical and Logical Perspectives

ILLC DS-2023-08: **Lei Li**
Games, Boards and Play: A Logical Perspective

ILLC DS-2023-09: **Simon Rey**
Variations on Participatory Budgeting

ILLC DS-2023-10: **Mario Giulianelli**
Neural Models of Language Use: Studies of Language Comprehension and Production in Context

ILLC DS-2023-11: **Guillermo Menéndez Turata**
Cyclic Proof Systems for Modal Fixpoint Logics

ILLC DS-2023-12: **Ned J.H. Wontner**
Views From a Peak: Generalisations and Descriptive Set Theory

ILLC DS-2024-01: **Jan Rooduijn**
Fragments and Frame Classes: Towards a Uniform Proof Theory for Modal Fixed Point Logics

ILLC DS-2024-02: **Bas Cornelissen**
Measuring musics: Notes on modes, motifs, and melodies

ILLC DS-2024-03: **Nicola De Cao**
Entity Centric Neural Models for Natural Language Processing

ILLC DS-2024-04: **Ece Takmaz**
Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective

ILLC DS-2024-05: **Fatemeh Seifan**
Coalgebraic fixpoint logic Expressivity and completeness result

ILLC DS-2024-06: **Jana Sotáková**
Isogenies and Cryptography

ILLC DS-2024-07: **Marco Degano**
Indefinites and their values

ILLC DS-2024-08: **Philip Verduyn Lunel**
Quantum Position Verification: Loss-tolerant Protocols and Fundamental Limits

ILLC DS-2024-09: **Rene Allerstorfer**
Position-based Quantum Cryptography: From Theory towards Practice

ILLC DS-2024-10: **Willem Feijen**
Fast, Right, or Best? Algorithms for Practical Optimization Problems

ILLC DS-2024-11: **Daira Pinto Prieto**
Combining Uncertain Evidence: Logic and Complexity

ILLC DS-2024-12: **Yanlin Chen**
On Quantum Algorithms and Limitations for Convex Optimization and Lattice Problems

ILLC DS-2024-13: **Jaap Jumelet**
Finding Structure in Language Models

ILLC DS-2025-01: **Julian Chingoma**
On Proportionality in Complex Domains

ILLC DS-2025-02: **Dmitry Grinko**
Mixed Schur-Weyl duality in quantum information

ILLC DS-2025-03: **Rochelle Choenni**
Multilinguality and Multiculturalism: Towards more Effective and Inclusive Neural Language Models

ILLC DS-2025-04: **Aleksi Anttila**
Not Nothing: Nonemptiness in Team Semantics

ILLC DS-2025-05: **Niels M. P. Neumann**
Adaptive Quantum Computers: decoding and state preparation

ILLC DS-2025-06: **Alina Leidinger**
Towards Language Models that benefit us all: Studies on stereotypes, robustness, and values

ILLC DS-2025-07: **Zhi Zhang**
Advancing Vision and Language Models through Commonsense Knowledge, Efficient Adaptation and Transparency

ILLC DS-2025-08: **Sophie Klumper**
The Gap and the Gain: Improving the Approximate Mechanism Design Frontier in Constrained Environments

ILLC DS-2026-01: **Bryan Eikema**
A Sampling-Based Exploration of Neural Text Generation Models

ILLC DS-2026-02: **Marten Folkertsma**
Empowering Quantum Computation with: Measurements, Catalysts, and Guiding States

ILLC DS-2026-03: **Valentin Richard**
Presuppositional and Dynamic Aspects of Questions

ILLC DS-2026-04: **Puyu Yang**
Bringing Science to the Public: The Role of Wikipedia in Scientific Communication

ILLC DS-2026-05: **Johannes Kloibhofer**
Cycles with Annotations: Non-Wellfounded Proof Theory of Modal Fixpoint Logics

ILLC DS-2026-06: **Danish Kashaev**
Approximation via duality in offline, online and strategic settings

ILLC DS-2026-07: **Oskar van der Wal**
Taking a Step Back: Measuring and Mitigating Bias in Language Models