Context-Sensitive Modeling of Metaphoric Potential with Large Language Models

MSc Thesis *(Afstudeerscriptie)*

written by

Shihui Li

under the supervision of **Dr Jelke Bloem** and **Dr Xiaojuan Tan**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
August 28, 2026	Dr Petter Törnberg
	Dr Marieke Schouwstra
	Dr Zheng Wei Lim



Institute for Logic, Language and Computation

Abstract

Psycholinguistic variables such as concreteness, imageability, familiarity, and iconicity are commonly considered to influence a word's metaphorical usages. However, existing studies usually rely on static lemma-level norms, although metaphorical meaning is strongly context-dependent. Recent work suggests that Large Language Models (LLMs) can approximate human psycholinguistic ratings, making them a possible tool for generating context-sensitive estimates at scale.

In this thesis, we construct a human-annotated dataset of in-context ratings for the four variables using metaphorical and literal sentences from the VU Amsterdam Metaphor Corpus. We then evaluate four LLMs on both isolated and contextual ratings, and use lexical and LLM-generated contextual features to model metaphoric potential. We find that GPT-4o achieves the strongest agreement with human judgments, while performance is higher for concreteness and imageability than for familiarity and iconicity. Model-human agreement also improves for items with more consistent human ratings.

In metaphor modelling, contextual and lexical ratings provide complementary information. Contextual ratings generally outperform lexical ratings alone, while combining both yields the best overall performance and generalization. Contextual ratings capture how psycholinguistic properties are realized in a particular sentence: odds ratio analyses show that concreteness decreases with a shift towards metaphorical meaning, whereas imageability exhibits the opposite pattern. In stepwise regression, familiarity emerges as the strongest predictor.

Overall, our dataset provides a resource for studying contextual variation in psycholinguistic properties and evaluating LLMs' ability to approximate human lexical judgments. The thesis also provides corpus-based evidence for cognitive accounts of metaphor and contributes a contextual psycholinguistic dataset at the intersection of NLP and psycholinguistics.

Contents

1 Introduction **4**

2 Related work **7**

2.1 Metaphoricity, metaphor potential and context 7

2.1.1 Metaphoricity and metaphor potential 7

2.1.2 Metaphor modelling 8

2.1.3 Context in metaphor modelling 8

2.2 Psycholinguistics norming 10

2.2.1 Traditional isolated psycholinguistic norms 10

2.2.2 Contextualized psycholinguistic ratings 11

2.2.3 Application in computational linguistics 11

2.3 Lexical variables and metaphoric potential 12

2.3.1 Psycholinguistic Variables 12

2.3.2 Other lexical factors 14

2.4 LLMs as predictors of psycholinguistic properties 15

2.5 Research gap and present study 16

3 Methodology **18**

3.1 Dataset development 18

3.1.1 Sentence selection 18

3.1.2 Annotation instructions 19

3.1.3 Questionnaire design 20

3.1.4 Ethical considerations 20

3.2 Post-processing 21

3.3 Model evaluation 21

3.3.1 Model selection 21

3.3.2 Prompts 22

3.4 Multifactorial analysis 23

4 Study 1: In-context dataset **25**

4.1 Dataset consistency and reliability 25

4.2 Descriptive statistics and homogeneity 25

4.3 Comparison with isolated ratings 26

4.4 Possible factors influencing data consistency 26

4.4.1 Questionnaire design and rating density 26

4.4.2 High-disagreement analysis . . . 27

5 Study 2: LLMs rating of psycholinguistics variables 31
5.1 Correlation analysis . . . 31
5.1.1 Isolated rating results . . . 31
5.1.2 In-context rating results . . . 31
5.1.3 In-context vs isolated ratings . . . 32
5.2 Influence of part of speech and metaphorical usage . . . 33
5.3 Error analysis . . . 34

6 Study 3: Modelling metaphoric potential with psycholinguistic variables 36
6.1 VUAMC subset extraction . . . 36
6.2 Cross-correlation matrix . . . 38
6.3 Binomial logistic regression . . . 40
6.4 Stepwise regression . . . 42
6.5 Multifactorial analysis . . . 43

7 Discussion and limitations 47
7.1 Further discussion of LLM prediction . . . 47
7.2 Complementary effects of lexical and contextual ratings . . . 47
7.3 Effects of each lexical variable in metaphor modelling . . . 48
7.3.1 Concreteness . . . 48
7.3.2 Imageability . . . 48
7.3.3 Familiarity and frequency . . . 49
7.3.4 Iconicity . . . 49
7.3.5 Part of speech . . . 50
7.4 Limitations . . . 50

8 Conclusion 52

Bibliography 54

A Annotation Instructions 62
A.1 Concreteness . . . 62
A.2 Iconicity . . . 63
A.3 Imageability . . . 64
A.4 Familiarity . . . 65

B Model rating prompts 66
B.1 Concreteness . . . 66
B.2 Iconicity . . . 67
B.3 Imageability . . . 67
B.4 Familiarity . . . 68

C Human rating highest disagreement 69

D Model rating full results **71**
D.1 Concreteness . . . 71
D.1.1 Isolated results . . . 71
D.1.2 In-context results . . . 71
D.2 Iconicity . . . 72
D.2.1 Isolated results . . . 72
D.2.2 In-context results . . . 73
D.3 Imageability . . . 74
D.3.1 Isolated results . . . 74
D.3.2 In-context results . . . 75
D.4 Familiarity . . . 76
D.4.1 Isolated results . . . 76
D.4.2 In-context results . . . 76

E Item level comparison lists **78**
E.1 Concreteness . . . 78
E.2 Imageability . . . 80
E.3 Iconicity . . . 81
E.4 Familiarity . . . 83

Chapter 1

Introduction

Metaphor is a fundamental mechanism through which abstract concepts are understood and expressed. In Conceptual Metaphor Theory, abstract target domains are commonly structured through mappings from more concrete source domains grounded in bodily and perceptual experience [LJ80]. From a linguistic perspective, however, metaphor is not determined by the word alone. The same lexical item may be used literally in one context and metaphorically in another. Metaphoric potential, understood as the possibility that a word is used or interpreted metaphorically, therefore depends on both relatively stable lexical properties and their realization in a particular context.

Psycholinguistic variables provide a way to quantify some of the properties associated with metaphoric potential. Concreteness measures the extent to which a meaning is grounded in sensory or bodily experience, while imageability reflects how easily it evokes a mental image. Familiarity concerns language users' experience with a meaning, and iconicity captures the perceived resemblance between linguistic form and meaning. Alongside frequency and part of speech, these variables have been associated with metaphorical usage in cognitive and corpus-based research. In particular, concrete and imageable meanings may provide perceptually accessible source concepts for metaphorical mappings, while familiarity and iconicity may affect the accessibility and salience of these meanings.

Large psycholinguistic norm datasets make these subjective properties available for quantitative analysis. Resources such as the MRC Psycholinguistic Database, the Brysbaert concreteness norms, the Glasgow Norms, and the English iconicity norms assign numerical ratings to thousands of words [Col81; BWK14; Sco+19; Win+24]. These ratings have been widely used as experimental controls, computational features, and evaluation benchmarks. They also allow researchers to combine metaphor-annotated corpora with psycholinguistic information and examine how different lexical properties predict metaphorical usage.

However, traditional norms normally assign one aggregate rating to each word. Such lemma-level ratings capture broad lexical tendencies, but cannot represent variation among different senses and contextual uses. For example, the psycholinguistic properties of a word may differ when it refers to a physical object and when the same word expresses an abstract or metaphorical meaning. Contextual norming studies provide evidence for this distinction. The CONcreTEXT dataset shows that concreteness ratings vary systematically across sentence contexts while remaining related to isolated ratings [Mon+23]. Similar context effects have been reported for sensorimotor properties and metaphor-related judgments. Nevertheless, contextual resources remain limited, particularly for imageability, familiarity, and iconicity.

This limitation is especially important for metaphor modelling. MIPVU identifies metaphor by

comparing the contextual meaning of a lexical unit with a more basic meaning [Ste+10]. Assigning the same psycholinguistic score to all occurrences of a word removes precisely this distinction. Existing corpus-based analyses have shown that lexical psycholinguistic variables contribute to the prediction of metaphorical usage, but they generally operationalize these variables as static lemma-level properties. Their coverage is also restricted to words included in existing human norm datasets.

Large Language Models (LLMs) offer a possible way to extend psycholinguistic ratings. Recent studies show that LLM-generated estimates can correlate strongly with human ratings of concreteness, imageability, familiarity, and iconicity [MMG+25; Con+25b; BMR25; Mar+25]. LLMs can generate ratings for lexical items not covered by existing norms and, importantly, can receive complete sentences as input. They may therefore provide token-level estimates that are sensitive to the meaning instantiated in a particular context.

Despite this potential, most evaluations of LLM-generated psycholinguistic ratings have focused on isolated words or multi-word expressions. It remains unclear how accurately LLMs reproduce human judgments for individual words in naturally occurring contexts, particularly when the data contain both literal and metaphorical usages. It is also unclear whether these contextual ratings contribute information beyond traditional lexical norms when modelling metaphoric potential. Psycholinguistic norming, LLM-based rating prediction, and corpus-based metaphor modelling have thus rarely been integrated.

In this thesis, we construct a human-annotated dataset of in-context psycholinguistic ratings, evaluate whether LLMs can reproduce both isolated and contextual judgments, and use lexical and contextual ratings to model metaphorical usage in the VU Amsterdam Metaphor Corpus. We focus on four variables: concreteness, imageability, familiarity, and iconicity. Our research questions are as follows:

RQ1: How do humans rate psycholinguistic variables at the token level in sentential context? (Chapter 4)

RQ2: How well can LLMs predict human psycholinguistic ratings both in isolation and in context? (Chapter 5)

RQ3: How do lexical and contextual psycholinguistic ratings contribute to the modelling of metaphoric potential? (Chapter 6)

Our contributions are as follows:

- We construct a human-annotated, token-level dataset of in-context concreteness, imageability, familiarity, and iconicity ratings using naturally occurring sentences from the VU Amsterdam Metaphor Corpus;
- We systematically evaluate LLM predictions of four psycholinguistic variables at both the lexical and contextual levels on three models, comparing several open-source and proprietary models with human judgments;
- We compare lexical, contextual, and combined representations in multifactorial models of metaphorical usage, allowing us to test whether contextual ratings provide information beyond conventional word-level norms.

The remainder of this thesis is organized as follows. Chapter 2 reviews related work on metaphoric potential and context, psycholinguistic norming, lexical variables associated with metaphor, and LLMs as predictors of psycholinguistic properties. Chapter 3 describes the dataset construction, human annotation, LLM evaluation, and multifactorial analysis. Chapter 4 presents the human in-context rating dataset and evaluates its consistency and reliability. Chapter 5 compares human and LLM ratings in isolation and in context. Chapter 6 examines how lexical and contextual psycholinguistic variables contribute to the modelling of metaphorical usage. Chapter 7 discusses the main findings and limitations, and Chapter 8 concludes the thesis.

Chapter 2

Related work

2.1 Metaphoricity, metaphor potential and context

2.1.1 Metaphoricity and metaphor potential

Metaphoric potential and metaphoricity are related but distinct concepts. Research on metaphoricity generally follows two branches. First, the static and gradient branch treats metaphoricity as a property of conceptual structures, mappings, concepts, words, or verbal expressions. From this perspective, metaphoricity is influenced by factors such as conventionality, semantic distance, metaphor type, and linguistic form. The dynamic approach, however, understands metaphoricity as the activation or awareness of metaphor in actual language use. Overall, the term metaphoricity may refer to both the static property of a metaphor and activation of a cross-domain mapping, and may also refer to immediately felt "experiential affordances" or behavioral processes.

The notion of metaphoric potential refers to the possibility of metaphorical activation, which is different from its actual realization. Cameron describes *potential metaphoricity* as some relatively stable conditions that make a metaphorical interpretation possible, while distinguishing them from "actual evidence of metaphoricity" [Cam97, p. 24]; [Cam18, pp. 20–24]. In actual discourse, the degree of metaphoricity is dynamic and depends on "contextual factors," such as the type and development of discourse, as well as the receiver's "background knowledge" and "processing demands" [Cam97, pp. 30, 210]; [Cam10, p. 79]. Potential metaphoricity therefore does not necessarily mean that metaphorical meaning is activated in a particular instance.

Müller makes a similar distinction between *Sleeping* and *Waking metaphors*. Sleeping metaphors have "metaphoric potential" and may be "realized in an utterance or written text," but they lack empirical indications of activated metaphoricity and therefore show a relatively low degree of activation [Mül08, pp. 11, 196–198, 209]. Waking metaphors, by contrast, display actually realized metaphoricity and are accompanied by "metaphoricity indicators" in co-occurring modalities, such as gesture or pictures [Mül08, p. 198]. Her account also separates the presence of metaphorical potential from observable evidence that this potential has been activated.

A comparable distinction is found in Zlatev and colleagues' Motivation and Sedimentation Model (e.g.[BZ21], [DZ20]). Embodied experience and sedimented linguistic conventions may provide the basis for potential metaphoricity. However, "real metaphoricity" or "actual metaphor use" occurs at the *Situated level*, where language use is "tightly dependent on" and interacts with contextual factors. Metaphoric potential may therefore exist as an embodied or conventional resource, whereas actual

metaphoricity emerges only in situated language use.

Overall, metaphoric potential refers to the possibility that metaphorical meaning can be activated, whereas metaphoricity refers either to a static, gradable property or to the degree of metaphorical activation realized in discourse. An expression may therefore possess metaphoric potential without showing clear evidence of activated metaphoricity in a particular context. In this thesis, we mainly discuss a word's metaphoric potential.

2.1.2 Metaphor modelling

Metaphor is modelled contextually in both cognitive linguistics and computational linguistics.

Traditionally, Conceptual Metaphor Theory models metaphor as mappings between source and target domains Lakoff and Johnson [LJ80]. A primary tenet of this theory is that metaphors are matter of thought and not merely of language. Therefore, a source domain, which is typically a more physical concept, is used by metaphor to describe a target domain, which is typically more abstract. For example, in "life is a journey", *life* is the target domain, described by the source domain *journey*. The concept of a journey therefore provides the mental organization of the more abstract concept of life.

In computational linguistics, MIPVU [Ste+10] is the most widely used procedure to identify metaphor in text. The main procedures are:

- Identify the lexical units in the text.
- Determine the contextual meaning of the target lexical unit.
- Establish whether the lexical unit has a more basic contemporary meaning in other contexts, using dictionary evidence where necessary.
- Compare the contextual and basic meanings.
- Mark the lexical unit as metaphor-related when the meanings contrast but the contextual meaning can be understood through comparison with the basic meaning.

The procedure is based on the contextual nature of metaphor, as metaphor identification depends on the comparison between the contextual meaning and a more basic meaning. MIPVU has been used to create large metaphor datasets for downstream NLP tasks. It was used to construct the VU Amsterdam Metaphor Corpus, which contains 200,000 manual token-level annotations across academic texts, news, fiction, and conversation [KS17]. The corpus subsequently formed the basis of the VUA metaphor detection shared tasks and has become a standard benchmark for computational metaphor identification [LBS18b; Leo+20b]. The procedure has also been adapted to construct metaphor corpora in other languages, including the Spanish CoMeta corpus [SA22]. These resources have been used to train and evaluate contextualized neural models and, more recently, to test whether generative LLMs can reproduce individual steps of MIPVU-based reasoning [RS25].

2.1.3 Context in metaphor modelling

The contextual nature of metaphor has long been recognized in both theoretical and computational research. Context plays a central role in the activation and This view has a long theoretical history,

from Richards' "context theory of meaning" and Black's interaction theory [Ric36; Bla62] to later accounts of contextual interpretation, including Bosch's context-dependent account of metaphor, Grice's conversational implicature, Sperber and Wilson's relevance theory, Auer and Di Luzio's account of contextualization, and Giora's graded salience hypothesis [Bos85; Gri75; SW86; AD92; Gio97]. These approaches differ in focus, but they share the assumption that metaphorical meaning cannot be fully explained independently of the conditions in which an expression is interpreted.

More recent dynamic approaches develop this claim by treating metaphoricity as relative to particular users and situations. As Littlemore argues, an expression is rarely metaphorical "in any abstract sense," but rather for "a particular user in a given context" [Lit15]. Metaphoricity may therefore vary across speakers according to their familiarity, background knowledge, and participation in a specific situation. In this sense, context does not merely help the receiver identify a pre-existing metaphorical meaning; it affects whether metaphoricity becomes noticeable or activated, and to what degree.

As Julich-Warpakowski and Jensen [JJ23] mentioned, a useful distinction should be made between *co-text* and *context*. Corpus-based studies often focus on the immediate linguistic environment of a word, including its collocations, grammatical pattern, frequency, conventionality, and semantic relation to nearby items. For instance, *arrive* can be identified as metaphorical in *arrive at a conclusion* because its complement does not refer to a physical destination. In this type of analysis, metaphoricity is mainly treated as a lexical, colligational, or co-textual property. The surrounding concordance window provides useful evidence, but it captures a narrower linguistic environment than discourse or conversational data.

In dynamic metaphor research, *context* is understood more broadly. It may include the developing discourse, the communicative situation, the participants and their background knowledge, previous interaction, and wider cultural, political, social, or physical conditions. From this perspective, metaphoricity cannot be located only in the word itself or in the contrast between Topic and Vehicle. Rather, it emerges through language use and may vary with the speaker, the situation, and the development of the discourse. Metaphoricity is therefore viewed as a dynamic feature of communication rather than a fixed property of the linguistic system.

The difference between co-textual and broader contextual approaches also affects research design. Corpus-based methods make it possible to define the object of analysis clearly and to operationalize metaphoricity through variables such as frequency, conventionality, collocation, and domain concreteness. However, they usually capture less of the interactional and situational processes involved in online meaning construction. Dynamic approaches can account for these processes more directly, but they are also harder to quantify because they involve a larger number of linguistic and non-linguistic factors. The two approaches are therefore best seen as complementary.

The present study examines metaphoricity at the sentence, or co-textual, level. By showing systematic differences in psycholinguistic properties across sentence contexts, it provides corpus-based and computational evidence that metaphorical interpretation is sensitive to linguistic context. Although this does not capture the full range of cultural, political, and interactional context considered in dynamic metaphor theory, it offers quantitative support for the broader claim that metaphorical meaning changes with language use and should not be treated as entirely fixed.

2.2 Psycholinguistics norming

2.2.1 Traditional isolated psycholinguistic norms

Psycholinguistic norming aims to quantify lexical properties that cannot be obtained directly from dictionaries or corpus frequencies, such as concreteness, imageability, familiarity, age of acquisition, and iconicity. In the traditional norming paradigm, words are presented individually and participants rate each item on a numerical scale. Responses are then averaged to assign a single value to each lexical item. This approach has produced widely used resources for stimulus selection, lexical processing research, and computational modeling.

An influential early resource is the MRC Psycholinguistic Database [Col81]. It contains 150,837 lexical records, corresponding to 115,331 distinct orthographic strings. It was assembled by combining several smaller resources that were more difficult to obtain. Its core word list was drawn from the Edinburgh Associative Thesaurus and the machine-readable dictionary of Dolby et al., with additional proper names from Mitton's machine-usable dictionary [Kis+73; DRM63; Mit86].

The MRC familiarity, concreteness, and imageability ratings were derived by merging three sets of norms: an unpublished expansion of the Paivio norms [PYM68], the Colorado norms of Toglia and Battig [TB78], and the norms of Gilhooly and Logie [GL80b]. Before ratings for overlapping words were averaged, the means and standard deviations of the three datasets were adjusted to place them on comparable scales. The database also includes variables such as written and spoken word frequency, part of speech, meaningfulness, and age of acquisition. Coverage varies substantially across variables. For example, familiarity, concreteness, and imageability ratings are available for only 9,392, 8,228, and 9,240 records, respectively.

Later norming studies substantially increased both the coverage and methodological consistency of individual variables. Brysbaert, Warriner, and Kuperman [BWK14], for example, collected concreteness ratings for 37,058 English words and 2,896 two-word expressions from more than 4,000 participants. The single words are rated on the lemma level. Participants rated isolated items on a five-point scale from abstract to concrete, with concreteness defined in relation to direct sensory and motor experience. The resulting norms provide much broader lexical coverage than the concreteness ratings available in earlier databases and have become a standard resource in psycholinguistics and natural language processing.

Winter, Lupyan, Perry, Dingemanse, and Perlman [Win+24] collected iconicity ratings for more than 14,000 English words, which was the first systematic database of this property. Participants evaluated the perceived resemblance between aspects of a word's linguistic form and aspects of its meaning on a seven-point scale. This resource made it possible to treat iconicity as a gradient lexical variable rather than restricting it to a small class of clearly onomatopoeic words. It also provides sufficient coverage for corpus-based and computational analyses of the relationship between iconicity and lexical behavior.

Several other important norm datasets are available. The Glasgow Norms provide ratings for approximately 5,500 English words on nine dimensions, including concreteness, imageability, familiarity, age of acquisition, valence, arousal, and dominance [Sco+19]. Their main advantage is that several related properties were rated for the same vocabulary using a relatively consistent procedure. Other resources focus on affective variables, sensorimotor strength, age of acquisition, word prevalence, or particular classes of expressions. More recent work has also expanded norming from individual words

to multiword expressions, acknowledging that the properties of an expression cannot always be derived from those of its component words.

In this thesis, we use the MRC database for familiarity and imageability ratings. And we choose the concreteness norms by Brysbaert, Warriner, and Kuperman [BWK14] since it is the largest concreteness dataset currently, and for iconicity norms, we choose the work by Winter, Lupyan, Perry, Dingemanse, and Perlman [Win+24].

2.2.2 Contextualized psycholinguistic ratings

Most psycholinguistic norms assign a single rating to each lexical item. This approach is useful for capturing broad lexical tendencies, but it does not account for variation across different uses of the same word. Recent norming studies have therefore begun to collect ratings for words and expressions presented in specific linguistic contexts.

The clearest example is the CONcreTEXT dataset introduced by Montefinese, Gregori, Ravelli, Varvara, and Radicioni [Mon+23]. The dataset provides concreteness ratings for English and Italian target words embedded in sentences. The ratings show high inter-rater reliability and correlate strongly with ratings collected in isolation, while also varying systematically across contexts. This suggests that isolated concreteness ratings capture a general lexical tendency, whereas contextual ratings reflect the meaning instantiated by a particular occurrence.

A similar approach has been applied to sensorimotor properties. Trott and Bergen [TB22] collected ratings for 112 ambiguous English words, each presented in four sentence contexts, resulting in 448 contextualized items. Participants rated the association of each target word with six sensory modalities and five action effectors. The resulting profiles differed across senses of the same word, showing that context can affect not only the overall degree of perceptual grounding, but also the specific sensory and motor dimensions associated with a lexical meaning.

Contextual norming has also begun to appear in metaphor research. Pissani and Almeida [PA26] provide familiarity, aptness, concreteness, metaphoricity, and structure norms for 300 two-word metaphorical expressions presented both in isolation and in sentence context. Their study offers a direct comparison between isolated and contextual judgments, although the ratings are assigned to complete expressions rather than to individual target words.

Comparable resources remain limited for imageability, familiarity, and iconicity. Existing norms for these variables are still largely based on isolated words, and no broad dataset currently provides token-level iconicity ratings across different contextual meanings. This leaves an important gap for human-validated ratings that capture how psycholinguistic properties vary with the meaning of a word in context.

2.2.3 Application in computational linguistics

Psycholinguistic ratings turn subjective properties such as concreteness, imageability, familiarity, and iconicity into measurable variables. This has made them central to psycholinguistic research, where they are used both to control experimental stimuli and to explain variation in lexical processing, memory, and semantic access [BWK14; Sco+19; Win+24]. Large-scale norms allow researchers to test how abstract and concrete words differ systematically across broad vocabularies rather than only in small, manually selected stimulus sets.

In computational linguistics, these ratings provide a human-grounded interface between cognitive theory and model representations. They have been used as lexical features, prediction targets, and evaluation benchmarks. For instance, Kiela, Wang, and Cho [KWC18] used concreteness ratings to show that multimodal models rely more strongly on visual representations for concrete words, while Rabinovich, Sznajder, Spector, Shnayderman, Aharonov, Konopnicki, and Slonim [Rab+18] evaluated automatically inferred lexical abstractness against human judgments. In metaphor research, Do Dinh, Wieland, and Gurevych [DWG18] examined the relationship between lexical concreteness and metaphor novelty, and Wu and Bruch [WB25] used MRC imageability ratings to probe psycholinguistic information encoded in embedding spaces. More recently, Dönmez, Maurer, Lapesa, and Faleńska [Dön+25] incorporated concreteness and iconicity norms into an interpretable comparison of human- and LLM-generated language.

Together, these studies show that psycholinguistic norms serve as a common scale for comparing human judgments with computational behavior. However, this interface remains largely static: one aggregate rating is typically assigned to each lexical item, even when the relevant word is processed or annotated within a specific context.

2.3 Lexical variables and metaphoric potential

Metaphoric potential is influenced by a range of lexical and psycholinguistic variables. Previous studies have identified concreteness, imageability, familiarity, iconicity, word frequency, and part of speech as important factors associated with metaphor use. The following sections review how these variables contribute to metaphorical language from both psycholinguistic and computational perspectives.

2.3.1 Psycholinguistic Variables

Concreteness

Concreteness refers to the extent to which a meaning can be directly experienced through perception or bodily interaction [BWK14]. Concrete meanings are generally processed and remembered more easily than abstract meanings, an effect often explained by their richer perceptual and sensorimotor representations [Pai13].

Concreteness has long been linked to metaphor. Conceptual Metaphor Theory proposes that abstract concepts are commonly understood through mappings from more concrete source domains grounded in bodily experience [LJ80]. Similarly, MIPVU identifies metaphor by comparing a contextual meaning with a more basic meaning, which is often more concrete and closely related to physical experience [Ste+10]. Concreteness is therefore relevant both to metaphorical meaning construction and to metaphor identification.

Corpus and NLP studies provide empirical support for this relation. Lexical concreteness has been shown to distinguish metaphorical from literal uses in annotated corpora, and to improve automatic metaphor detection when combined with contextual representations [GRB14; BLF15; APC20]. Large-scale corpus analyses further suggest that metaphorical uses often co-occur with words with greater abstractness than literal uses [AST24]. These studies indicate that concreteness provides useful information for modelling metaphor, although its role depends on the linguistic context rather than on the target word alone.

However, almost all previous work relies on isolated lexical norms, assigning the same concreteness value to every occurrence of a lemma. This assumption is problematic for polysemous words. Concreteness ratings are influenced by the senses available to raters [Rei+19], and contextual norms show that the same word can receive different ratings across sentences [Mon+23]. The present study therefore measures concreteness for the target meaning in context instead of treating it as a fixed lexical property.

Imageability

Imageability refers to how easy a word evokes a mental image or perceptual representation [TB78; SD06]. Although closely related to concreteness, it reflects the ease of constructing a mental representation, not concerning the extent to which that meaning is grounded in direct sensory or bodily experience.

Compared with less imageable words, highly imageable words are generally recalled more accurately and remembered better [RF86; CKH10]. In cognitive science, imageability has long been associated with the processing of concrete meanings. Schwanenflugel et al. further showed that imagery contributes to the processing advantage of concrete meanings, as the memory advantage of concrete words is largely observed when participants make use of imagery during encoding [SS89].

Like concreteness, imageability has also been linked to metaphor. According to Conceptual Metaphor Theory, abstract concepts are understood through mappings from concrete domains rooted in bodily experience [LJ80]. The imageability of the source meaning may help maintain a vivid mental representation during this mapping, thereby providing perceptual support for interpreting an abstract target meaning [GIB06]. Julich-Warpakowski and Jensen further argue that concrete and imageable source concepts make metaphorical mappings more noticeable, although imageability itself does not determine whether an expression is perceived as metaphorical [JJ23].

The relationship between imageability and metaphoricity has also been explored in NLP. Corpus-based studies have incorporated imageability together with concreteness as lexical features for metaphor identification [GRB14]. Later work combined imageability with semantic and distributional features in supervised metaphor detection and showed that it provides complementary information beyond lexical semantics alone [Tsv+14; TÖS15]. More recently, large-scale corpus analyses found that metaphorically used verbs tend to occur with less imageable direct objects than literal verbs [AST24]. Together, these studies suggest that imageability is a useful cue for modelling metaphor, particularly when combined with semantic and contextual information.

Iconicity

Iconicity refers to the perceived resemblance between aspects of a word's form and aspects of its meaning [Win+24]. This resemblance is not limited to sound imitation. It can also involve movement, texture, intensity, or visual events, as in words such as *twirl*, *mush*, and *flash*. Iconicity is a graded lexical property and is related to several other psycholinguistic variables. More iconic words tend to be acquired earlier and are often more concrete and imageable than less iconic words [LW18; MP17; Per+18; Win+24]. These associations suggest that iconic words are often closely tied to embodied and perceptual experience.

This makes iconicity relevant to metaphor theory. Metaphorical meanings are often understood through concrete source domains grounded in bodily experience [LJ80]. Therefore, a recent unpublished corpus study proposes that, if the form of a source word already reinforces some perceptual aspect of its

basic meaning, that source meaning may remain more salient when the word is extended metaphorically. For example, in *sizzling romance*, the form and sound associations of *sizzling* reinforce the sensory source domain of heat and sound. Iconicity may therefore strengthen the perceptual availability of the source meaning and make the source–target relation more noticeable.

This proposal also fits the logic of MIPVU. MIPVU identifies metaphor by comparing a contextual meaning with a more basic meaning [Ste+10]. If the basic meaning is highly iconic and perceptually salient, the contrast between the basic and contextual meanings may be easier to recognise. Iconicity may therefore increase metaphor noticeability and, in turn, the likelihood that a use is identified as metaphorical.

Compared with concreteness and imageability, iconicity has received little attention in computational metaphor research. Existing NLP studies have instead focused on broader perceptual or sensorimotor features, which have been shown to improve metaphor detection when combined with contextual representations [WX20]. Whether lexical iconicity provides additional information beyond these perceptual features remains an open question.

Familiarity

Familiarity reflects how familiar a word meaning is to language users based on their previous language experience [SD06]. Although closely related to frequency and age of acquisition, familiarity is a subjective measure and depends on experience with a particular meaning rather than the word form alone. Sense-level studies further show that different meanings of the same word can differ substantially in familiarity, suggesting that a single lexical familiarity score may not adequately represent contextual meaning [Azu96].

Rather than determining whether an expression is metaphorical, familiarity mainly influences how noticeable a metaphor is. As a metaphorical meaning becomes conventional through repeated use, its contextual meaning becomes more familiar while its original source meaning becomes less salient. As a result, the metaphorical mapping becomes less noticeable, even though the expression remains metaphorical under MIPVU [Goa97; JJ23]. Familiarity therefore reflects the relative accessibility of the basic and contextual meanings during metaphor comprehension.The contribution of familiarity to metaphor mainly concerns the accessibility of metaphorical meanings.

Experimental studies also show that familiar metaphors are processed more efficiently than unfamiliar ones. Using a series of cross-modal priming experiments, Blasko and Connine [BC93] found that familiar metaphors activate their figurative meanings more readily than unfamiliar metaphors. Eye-tracking studies further demonstrate that familiarity facilitates metaphor comprehension even after controlling for metaphor conventionality and aptness [GV16]. These findings suggest that familiarity influences the accessibility of contextual meanings during metaphor processing.

Corpus evidence is consistent with this view. Aono, Sasano, and Takeda [AST24] found that metaphorically used verbs tend to occur with less familiar direct objects than literal verbs, suggesting that familiarity also characterises the semantic context in which metaphor occurs.

2.3.2 Other lexical factors

Besides psycholinguistic dimensions, lexical properties such as frequency and part of speech have also been associated with metaphor use.

Frequency

Word frequency reflects how often a lexical item occurs in language use. High-frequency words are generally recognised and processed more efficiently than low-frequency words, and frequency is strongly correlated with subjective familiarity [BWK14; SD06]. However, the two variables are not equivalent. Using behavioural familiarity ratings and corpus counts, Schock et al. [SCK12] showed that familiarity is predicted by frequency but also reflects language users' accumulated experience with individual meanings. Tanaka-Ishii and Terada [TT11] further compared familiarity ratings with several large corpora and found that although highly frequent words are almost always familiar, many familiar words are not frequent in a particular corpus. Frequency therefore reflects language distribution, whereas familiarity better captures the accessibility of a meaning. Because metaphor is interpreted at the level of contextual meaning rather than word form, contextual familiarity is expected to be more informative than lexical frequency for modelling metaphorical potential.

Part of speech

Part of speech influences both the distribution and the form of metaphor. Different lexical categories realise metaphor in different ways. Nouns typically evoke concrete entities and therefore provide explicit source-domain concepts, making nominal metaphors relatively easy to recognise [Cam03; JJ23]. Adjectives usually transfer perceptual or evaluative properties to abstract nouns, such as *bright idea* or *cold response*. By contrast, verbs primarily express metaphor through relational extensions between predicates and their arguments, where physical actions are mapped onto abstract events or states. Prepositions are another productive source of metaphor because they systematically extend spatial relations to temporal and abstract domains, such as *under pressure* or *in trouble*, reflecting the central role of spatial source domains in conceptual metaphor [LJ80].

Corpus studies support these differences. Analysing the newspaper section of the VU Amsterdam Metaphor Corpus, Krennmayr [Kre11] found that prepositions showed the highest within-category proportion of metaphorical uses, followed by verbs, whereas determiners and conjunctions were only rarely metaphorical. The high proportion of prepositional metaphors is consistent with the pervasive extension of spatial relations to abstract concepts in conceptual metaphor theory. Nouns accounted for many metaphorical tokens in absolute numbers, but their metaphor rate was considerably lower because nouns are also the most frequent content words. Similar differences are reflected in NLP research. The VUA metaphor shared tasks evaluate both verb-only and all-part-of-speech metaphor identification, recognising that metaphorical verbs rely heavily on predicate–argument relations, whereas nominal, adjectival, and prepositional metaphors involve different semantic mechanisms [LBS18a; Leo+20a]. Together, these findings suggest that part of speech affects not only how frequently metaphor occurs, but also how metaphorical meaning is realised and identified.

2.4 LLMs as predictors of psycholinguistic properties

Recent work has begun to evaluate whether Large Language Models can serve as a viable alternative to human norming studies for psycholinguistic ratings. This line of research is motivated by a practical constraint: traditional psycholinguistic norms are costly to collect and necessarily limited in vocabulary coverage, since they rely on large-scale human judgments collected through behavioral experiments.

Several studies have tested LLMs on the psycholinguistic variables relevant to metaphoricity. Martínez, Molero, González, et al. [MMG+25] show that LLM-generated concreteness ratings correlate strongly with human norms, extending this validation from single words to multi-word expressions. Conde, González, Grandury, Martínez, Reviriego, and Brysbaert [Con+25a] report similar findings for concreteness, familiarity, and imageability within a broader evaluation framework. For familiarity, Brysbaert, Martínez, and Reviriego [BMR25] further show that AI-generated familiarity estimates can supplement traditional frequency-based measures. Beyond these more established dimensions, Marklová, Milicka, Ryvkin, Lacková Bennet, and Kormaníková [Mar+25] extend this line of inquiry to iconicity, a property that has historically been more difficult to operationalize computationally, and show that LLMs can approximate human iconicity judgments with reasonable accuracy.

These findings suggest that LLMs encode psycholinguistic knowledge that aligns meaningfully with human judgments across a range of properties, including concreteness, familiarity, imageability, and iconicity. This makes them a promising tool for generating ratings at a scale that is difficult to achieve through human norming alone.

At the same time, recent work also shows that psycholinguistic ratings should not always be treated as fixed lemma-level properties. The CONcreTEXT norms provide concreteness ratings for Italian and English words presented in sentence contexts, showing that concreteness can be rated reliably in context and may be influenced by sentence-level information [Mon+23]. Similarly, research on metaphor norms has begun to compare ratings in context and in isolation. Pissani and Almeida [PA26] provide familiarity, aptness, concreteness, metaphoricity, and structure norms for 300 two-word metaphorical expressions, presented both in isolation and in sentence context. These studies suggest that context can play an important role in psycholinguistic and metaphor-related judgments.

However, LLM-generated ratings can also show systematic biases. Trott [Tro24] found that GPT-4's errors in semantic similarity judgments vary with word concreteness and semantic relations. Snelder, Huang, and Bloem [SHB26] also found that LLMs tend to confuse semantic similarity with relatedness, and that their predictions are sensitive to prompting methods. Some studies found that LLMs may have more difficulty capturing human sensory experiences associated with words. Conde, González, Grandury, Martínez, Reviriego, and Brysbaert [Con+25a] compared LLM predictions with human ratings across 13 psycholinguistic properties and found better agreement for the Glasgow norms than for the Lancaster sensorimotor norms. Hagihara and Miyazawa [HM26] also reported relatively low correlations for iconicity compared to concreteness.

Although LLMs can be useful for generating psycholinguistic ratings on a larger scale, most existing evaluations focus on isolated words or multi-word expressions, rather than individual words in sentence context. As contextual norming studies have shown, the same word can receive different ratings depending on its meaning in a particular sentence. It therefore remains unclear whether LLMs can capture these differences and produce ratings that agree with human contextual judgments.

2.5 Research gap and present study

Previous research has established that psycholinguistic variables are closely related to metaphor use, that many of these properties vary across contexts, and that LLMs can predict several psycholinguistic properties with high agreement with human judgments. However, these research directions have rarely been integrated. Most studies still rely on static lemma-level ratings when modelling metaphor,

while contextualized norm datasets remain scarce. Moreover, existing evaluations of LLM-generated psycholinguistic ratings have focused primarily on isolated words rather than individual word tokens in context. It therefore remains unclear whether context-sensitive psycholinguistic ratings, particularly those generated by LLMs, can be used to model metaphoric potential.

This thesis addresses these gaps through three studies:

- We construct a human-annotated in-context dataset of concreteness, imageability, familiarity, and iconicity using sentence contexts sampled from the VU Amsterdam Metaphor Corpus.
- We evaluate whether LLMs can predict both isolated and in-context psycholinguistic ratings by comparing model predictions with human judgments.
- We examine whether human- and LLM-generated psycholinguistic ratings can be used to model metaphoric potential in a metaphor-annotated corpus.

Chapter 3

Methodology

3.1 Dataset development

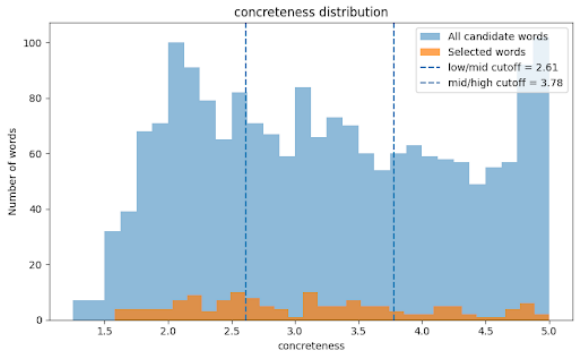
3.1.1 Sentence selection

Sentences from VUAMC were selected using the following strategy. We defined a word identity as the combination of lemma and POS. In other words, two tokens were treated as the same word only when both their lemma and POS matched. We first selected all word identities that had a human annotation in the chosen human dataset.

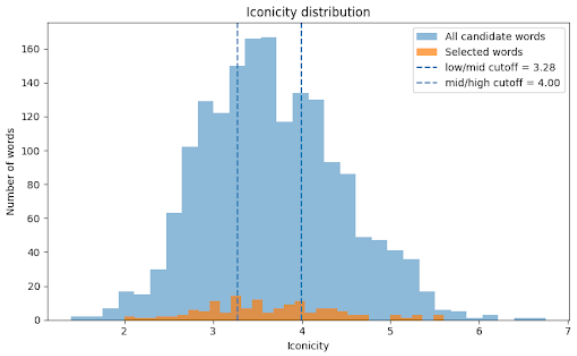
For each psycholinguistic feature, we selected lemmas with a balanced distribution of low, medium, and high ratings. To achieve this, we divided the words into three groups based on the tertiles of their feature ratings. We then randomly sampled an equal number of words from each group.

For each selected word, we selected three to five sentences from VUAMC. These sentences included both metaphorical and literal uses. We only included sentences with fewer than 25 words, so that they would be easier to annotate. We also ensured that each psycholinguistic subset contained an equal number of nouns, adjectives, and verbs.

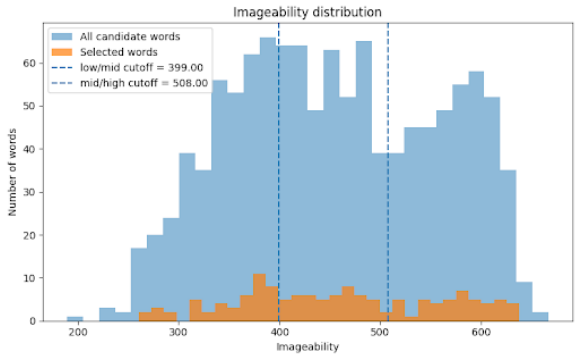
In total, the dataset includes four subsets: concreteness, iconicity, imageability, and familiarity. Each subset contains around 600 sentence-level ratings for 120 target words, with 40 nouns, 40 adjectives, and 40 verbs. The annotation distributions are presented in Figure 3.1.



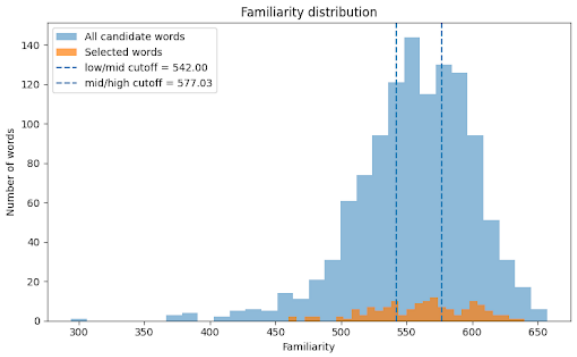
(a) Concreteness



(b) Iconicity



(c) Imageability



(d) Familiarity

Figure 3.1: Human annotation distribution across the four subsets. For each of the four psycholinguistic variables, the selected lemmas are distributed across the full range of ratings and follow a distribution similar to that of the full VUAMC dataset.

3.1.2 Annotation instructions

We collected the annotations through a questionnaire. Before annotation, participants read a general instruction about the overall research task, followed by a feature-specific instruction for the psycholinguistic variable they were going to rate.

For the instructions for each psycholinguistic variable, we adapted the rating instructions from the original human annotation datasets. These original instructions were designed for isolated word ratings, so we added two in-context examples below each instruction: one with a high rating and one with a low rating. Specifically, we used the instructions from Brysbaert, Warriner, and Kuperman [BWK14] for concreteness ratings and those from Winter, Lupyan, Perry, Dingemanse, and Perlman [Win+24] for iconicity ratings. For imageability and familiarity, we used the instructions from the Bristol Norms [SD06], which are based on earlier instructions for imageability [PYM68] and familiarity [GL80a].

The in-context examples were selected from the online Cambridge Dictionary. We made sure that all examples used the target word literally. This was done to avoid guiding participants to base their ratings on whether a word was used metaphorically or literally. For example, we did not want participants to systematically assign lower concreteness ratings to metaphorical uses. The full annotation instructions are listed in Appendix A.

3.1.3 Questionnaire design

For each psycholinguistic variable, namely concreteness, iconicity, imageability, and familiarity, we divided the sentences into six lists. Each list contained around 100 sentences. Each participant was randomly assigned to one psycholinguistic variable and rated one randomly selected list for that variable. To reduce possible order effects, we used the loop and merge function in Qualtrics to shuffle the order of the sentences within each list.

For each sentence, we collected at least 10 annotations. The target word was highlighted in bold in the sentence. Participants rated the highlighted word on a 7-point scale. For example, for concreteness, the prompt was: "How concrete is the highlighted word in this sentence?" The scale ranged from 1 = extremely abstract to 7 = extremely concrete. The word concrete was replaced by the corresponding variable for the other three ratings. Participants were asked to provide integer ratings from 1 to 7. The rating questions are presented as in Figure 3.2.

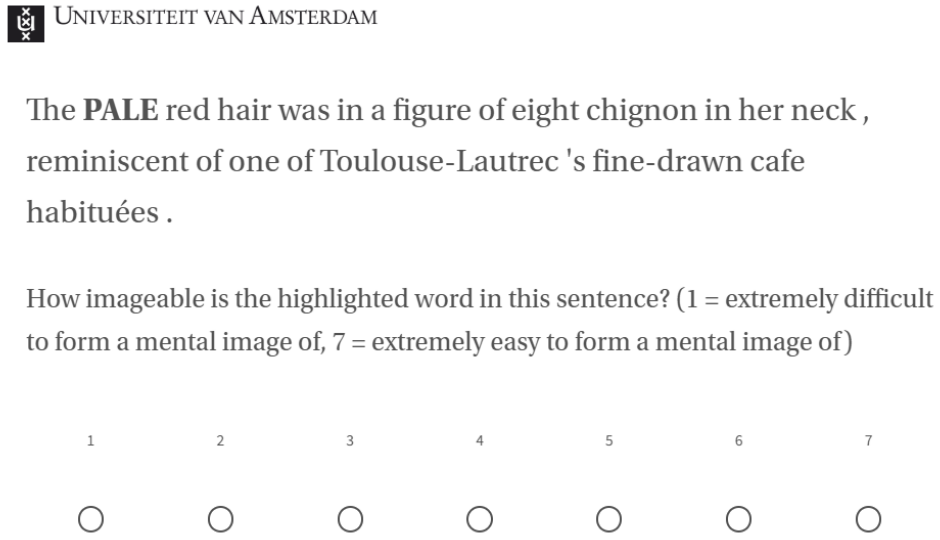


Figure 3.2: Screenshot of an imageability rating question in the questionnaire on Qualtrics.

Before the annotation task, participants answered several screening questions. First, we asked whether they were at least 18 years old. To include only native English speakers, we asked two questions: (1) whether English was one of their first languages, and (2) in which country or region they grew up speaking English. We also asked participants whether they had ever been diagnosed with a language disorder, such as dyslexia, aphasia, or Specific Language Impairment.

We created the survey in Qualtrics[1] and distributed it to native English speakers through personal contacts and colleagues in the UK, the US, and Australia.

3.1.4 Ethical considerations

Before starting the annotation task, participants were clearly informed about the research project and provided informed consent. Participation was voluntary, and participants were free to stop taking part at any stage without consequences or penalty. If they decided to withdraw their consent, all information gathered up to that point would be permanently deleted.

[1] https://www.qualtrics.com/

Participants were informed that the research had no known risks. The collected information would be used for academic purposes only, and fully anonymized data may be made available online together with scientific publications. Participants were guaranteed to remain anonymous under all circumstances. The study followed EU privacy law, including the GDPR, and provided contact information for further questions or ethics-related complaints. The study has passed the thesis fast-track ethics check by the Ethics Committee at the Faculty of Humanities, University of Amsterdam. The reference number is FGW-9426.

3.2 Post-processing

The dataset was processed in several steps before calculating the final item-level norms. First, participants were screened according to the inclusion criteria. Participants were excluded if they reported that they are under 18 year old, English was not (one of) their first language(s) or if they reported having a language disorder. This ensured that the ratings were based on responses from the target participant group.

Second, participant-level response quality was checked within each rating block. Participants were removed if their response patterns suggested low-effort responding. Specifically, participants were excluded if they gave the same rating to all items in a block, or if 90% or more of their responses consisted of the same rating value. We removed 2, 1, 0, 12 annotators from concreteness, iconicity, imageability and familiarity ratings in this process, respectively.

Third, to further identify annotators whose ratings were systematically different from the rest of the group, for each rating dimension, each annotator's responses were compared with a leave-one-annotator-out group consensus. We regard each psycholinguistics variable as a rating block. The consensus was calculated as the item-level mean rating of all remaining annotators in the same rating block. Then, the deviation of each annotator from this consensus was quantified using root mean squared error (RMSE). Annotators were excluded from a rating block if their RMSE was unusually high, defined as exceeding the block-specific median RMSE by more than 2.5 median absolute deviations. Using this criterion, we removed annotators who showed large overall deviation from the group pattern while avoiding overly aggressive exclusion. Specifically, 4, 4, 2, and 5 annotators were identified as low-quality annotators and excluded from the concreteness, iconicity, imageability, and familiarity rating blocks, respectively.

After this post-processing, the final dataset contained 51 annotators and 593 items for concreteness, 57 annotators and 594 items for iconicity, 58 annotators and 594 items for imageability, and 42 annotators and 597 items for familiarity. The mean number of ratings per item was 8.50 for concreteness, 9.50 for iconicity, 9.57 for imageability, and 7.00 for familiarity. The final item-level norms were then calculated by computing the mean and standard deviation of the remaining ratings for each item in each rating dimension.

3.3 Model evaluation

3.3.1 Model selection

We selected four models, with two instruction-tuned open-sourced models and two API-based proprietary models. But Llama did not generate any results in some settings, so we mainly only report the rest three model results:

- Llama-3.1-8B-Instruct. Meta's 8B-parameter instruction-tuned autoregressive Transformer.
- Qwen-2.5-7B-Instruct. Alibaba's 7.6B-parameter instruction-tuned decoder-only Transformer.
- GPT-4o-mini. OpenAI's lightweight proprietary model, released in July 2024. It supports a 128K-token context window and was designed as a smaller, more efficient version of GPT-4o.
- **GPT-4o**: OpenAI's proprietary multimodal model, released in May 2024. It also supports text and image inputs with a 128K-token context window.

The open-source models were accessed through Hugging Face[2]. The GPT models were accessed through the OpenAI API[3]. For the experiments in Chapter 6, we used the Batch API to process requests asynchronously, allowing large numbers of prompts to be submitted efficiently within the API rate limits.

3.3.2 Prompts

The specific rating prompts are listed in Appendix B.

For isolated ratings, we first tested a simple zero-shot prompt (Appendix B.1, Concreteness isolated rating prompt 1). We then adopted the prompt used by Brysbaert, Martínez, and Reviriego [BMR25] for concreteness, changing only the rating scale from 1–5 to 1–7 (Appendix B.1, Concreteness isolated rating prompt 2). Testing on the concreteness dataset showed that the second prompt achieved substantially higher correlation with the human ratings, so we adopted it for the remaining three variables. For iconicity, imageability, and familiarity, we only replaced the example words with those receiving the highest and lowest ratings in the corresponding human annotation datasets used in this thesis. The four isolated-rating prompts therefore follow the same template.

For in-context ratings, we first adapted the isolated prompt 2 by replacing the isolated examples with three high-rated and three low-rated {word, sentence} examples ((Appendix B.1, Concreteness in-context rating prompt 1)). We then tried another prompt, using a one-sentence definition of the variable taken from our human annotation instructions (Appendix A) and the same in-context example as in the human annotation ((Appendix B.1, Concreteness in-context rating prompt 2)). The second version achieved higher correlation with the human ratings. Therefore, we use isolated prompt 2 and in-context prompt 2 in all subsequent experiments and analyses.

Prompt template for isolated ratings

Could you please rate the {*concreteness / iconicity / imageability / familiarity*} of the following word on a scale from 1 to 7, where 1 means {*very abstract / the sound of the word does not resemble its meaning at all / very difficult to imagine / very unfamiliar*} and 7 means {*very concrete / the sound strongly resembles its meaning / very easy to imagine / very familiar*}?
Examples of words that would get a rating of 1 are {*low-rated example 1*}, {*low-rated example 2*} and {*low-rated example 3*}.
Examples of words that would get a rating of 7 are {*high-rated example 1*}, {*high-rated example 2*} and {*high-rated example 3*}.
The word is: {*word*}.
Only answer a number from 1 to 7. Please limit your answer to numbers.

[2] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`
[3] `https://platform.openai.com`

Prompt template for in-context ratings

Could you please rate the {*concreteness / iconicity / imageability / familiarity*} of the word in the sentence on a scale from 1 to 7, where:

1 = {*very abstract / not iconic at all / very difficult to imagine / very unfamiliar*}
7 = {*very concrete / extremely iconic / very easy to imagine / very familiar*}

{*A concrete word refers to things or actions in reality. / Words with high iconicity have forms that resemble their meanings. / An imageable word easily evokes a mental image or sensory experience. / A familiar word is one that is frequently encountered and easily recognized.*}
Please evaluate the word based on the context of the sentence you are reading.
For example, in the sentence "{*high-rating example sentence*}", the word {*target word*} may be evaluated {*as 7 / very high*}. In contrast, in the sentence "{*low-rating example sentence*}", the word {*target word*} may receive a lower rating.
Now rate the word: {*word*} in the sentence: {*sentence*}.
Only answer with a number from 1 to 7. It is important that you try to use the entire range of the scale (from 1 to 7). Please limit your answer to numbers.

3.4 Multifactorial analysis

To examine how psycholinguistic variables contribute to metaphoric potential, we model token-level metaphorical usage using binomial logistic regression. The dependent variable is whether a token is metaphorical ($Y = 1$) or literal ($Y = 0$). The independent variables are psycholinguistic ratings together with lemma frequency, part of speech, and semantic class.

Three models are considered.

Lexical model The lexical model uses conventional isolated psycholinguistic norms:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 C_{\text{lex}} + \beta_2 I_{\text{lex}} + \beta_3 F_{\text{lex}} + \beta_4 K_{\text{lex}} + \mathbf{Z}\boldsymbol{\gamma},$$

where C, I, F, and K denote concreteness, imageability, familiarity, and iconicity, respectively, and $\mathbf{Z}$ represents the control variables (lemma frequency and part of speech).

Contextual model The contextual model replaces the isolated norms with GPT-predicted in-context ratings:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 C_{\text{ctx}} + \beta_2 I_{\text{ctx}} + \beta_3 F_{\text{ctx}} + \beta_4 K_{\text{ctx}} + \mathbf{Z}\boldsymbol{\gamma}.$$

Combined model The combined model includes both lexical and contextual representations:

$$\text{logit}(P(Y = 1)) = \beta_0 + \sum_j \beta_j X_{\text{lex},j} + \sum_j \delta_j X_{\text{ctx},j} + \mathbf{Z}\boldsymbol{\gamma},$$

where X_j denotes the four psycholinguistic variables.

Model fit is compared using log-likelihood, Akaike Information Criterion (AIC), and the area under the ROC curve (AUC). To evaluate generalization beyond observed lexical items, we additionally perform a lemma-disjoint train–test evaluation. Finally, forward stepwise regression is used to examine

the relative contribution of different predictors, and full multifactorial logistic regression is used to estimate the independent effect of each variable while controlling for all remaining predictors. The experiment details and results are described in Chapter 6.

Chapter 4

Study 1: In-context dataset

4.1 Dataset consistency and reliability

We first evaluate the consistency and reliability of the human in-context ratings. First, we used Cronbach's alpha to examine the internal consistency of the dataset. For data stability, we calculated the generalizability coefficient (G-coefficient), bootstrap item-mean stability and Spearman-Brown corrected split-half reliability. G-coefficient decomposes rating variability into between-item variance and residual within-item variance. It was calculated as

$$G = \frac{\sigma^2_{\text{item}}}{\sigma^2_{\text{item}} + \frac{\sigma^2_{\text{error}}}{k}}, \tag{4.1}$$

where σ^2_{item} denotes the variance between different rating items, σ^2_{error} denotes the residual within-item variance, and k is the number of ratings per item.

The results are shown in Table 4.1. Overall, the dataset shows high internal consistency (Cronbach's alpha), relatively high stability of mean ratings, and moderate to acceptable generalizability, while item-level split-half reliability is comparatively lower. Familiarity was less reliable than the other dimensions.

4.2 Descriptive statistics and homogeneity

Table 4.2 presents the descriptive statistics of the item-level mean ratings for the four dimensions. The distributions of the ratings are shown in Figure 4.2.

The distributions of concreteness, iconicity, and familiarity ratings deviated significantly from a normal distribution. Imageability ratings did not significantly deviate from normality. Concreteness, imageability, and iconicity were approximately symmetric. Familiarity showed a clearer negative skew (-0.697), with the concentration of ratings towards the upper end of the scale.

A bias towards the high range of the scale was observed for concreteness and imageability, since 63.24% and 58.18% of the items, respectively, received a mean rating above the scale midpoint of 4. Familiarity showed a particularly strong bias towards the high end of the scale, with 99.33% of the items rated above the midpoint.

Figure 4.1 plots the homogeneity of the item-level mean ratings against their corresponding standard deviations. A quadratic model was fitted separately for each dimension:

Dimension	α	G-coeff.	Bootstrap	Split-half
Concreteness	.991	.764	.872	.698
Imageability	.991	.741	.859	.618
Iconicity	.984	.733	.858	.536
Familiarity	.981	.675	.832	.403

Table 4.1: Reliability and consistency of human in-context ratings across four psycholinguistic dimensions. Bootstrap values represent median item-mean correlations over 1,000 resamples; split-half values are Spearman-Brown corrected.

Table 4.2: Descriptive statistics of the item-level mean ratings for the four psycholinguistic dimensions. Kurtosis refers to excess kurtosis.

Dimension	N	Mean	SD	Median	IQR	Min	Max	Skewness	Kurtosis	Shapiro W	Above midpoint (%)
Concreteness	593	4.393	0.923	4.400	1.375	2.000	6.750	-0.004	-0.638	0.991	63.24
Imageability	593	4.278	0.946	4.300	1.300	1.667	6.667	-0.027	-0.271	0.997	58.18
Iconicity	594	3.502	0.861	3.500	1.100	1.250	5.600	-0.030	-0.336	0.994	26.43
Familiarity	597	5.985	0.623	6.000	0.833	3.667	7.000	-0.697	0.383	0.962	99.33

$$SD_i = \beta_0 + \beta_1 Mean_i + \beta_2 Mean_i^2 + \epsilon_i. \tag{4.2}$$

4.3 Comparison with isolated ratings

To assess the relationship between contextual and isolated ratings, we compared the mean contextual rating of each target word with its corresponding isolated rating from existing psycholinguistic norms: the concreteness norms of Brysbaert, Warriner, and Kuperman [BWK14], the iconicity norms of Winter, Lupyan, Perry, Dingemanse, and Perlman [Win+24], and the imageability and familiarity norms from the MRC Psycholinguistic Database [Col81]. As shown in Figure 4.3, contextual and isolated ratings were significantly correlated for concreteness ($r = 0.65$), imageability ($r = 0.67$), and iconicity ($r = 0.55$). Familiarity showed almost no association between the two measures ($r = 0.08$).

4.4 Possible factors influencing data consistency

4.4.1 Questionnaire design and rating density

Several design factors may explain the relatively low rating agreement. First, we collected 7-10 ratings per item in our final dataset. The number of ratings per item was lower than in most established norming datasets. Each item in the present study received approximately 7–10 ratings, compared with at least 25 in the Brysbaert concreteness norms, 20 for most items in the Bristol norms, and an average of 62.9 in the English CONcreTEXT dataset. The only comparable dataset was the Winter et al. iconicity norms, which contained approximately 10.9 ratings per word. However, our in-context rating task is more difficult for participants, so including more raters could help to increase the annotation agreement.

The questionnaire was also relatively demanding. Each participant rated approximately 100 sentences, which may have caused fatigue and reduced attention in later trials. In addition, the instructions had to explain both the psycholinguistic variable and the distinction between lexical and

Table 4.3: Quadratic regression models predicting item-level rating standard deviations from item-level mean ratings.

Dimension	Intercept	Mean	Mean^2	R^2
Concreteness	0.286	0.658	-0.080	0.070
Imageability	-0.333	1.229	-0.163	0.417
Iconicity	-1.259	1.603	-0.208	0.410
Familiarity	-2.941	2.138	-0.242	0.698

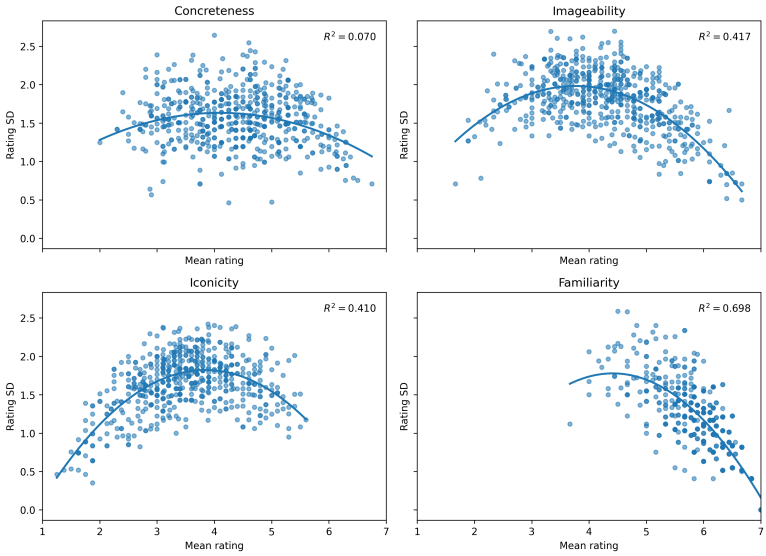


Figure 4.1: Homogeneity of the ratings. Item-level standard deviations plotted against their corresponding mean ratings for concreteness, imageability, iconicity, and familiarity. The solid lines represent quadratic regression fits, with the corresponding R^2 values shown in each panel.

contextual meaning. This increased the cognitive load of the task and may have led some participants to rely on the general meaning of the word rather than the meaning expressed in the sentence.

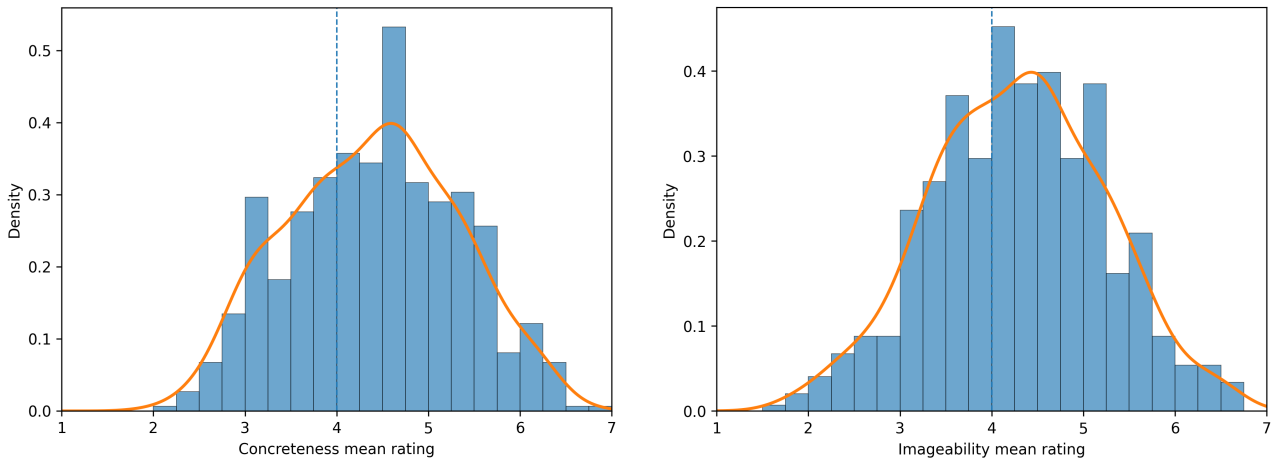
Apart from these design factors, we next examine the linguistic properties of the high-disagreement items.

4.4.2 High-disagreement analysis

Within each dimension, we now define items with an SD above the third quartile (Q_3) as high-disagreement items. The thresholds were 1.764 for concreteness, 2.058 for imageability, 2.000 for iconicity, and 1.246 for familiarity. This resulted in 148 items for concreteness, imageability, and iconicity, and 149 items for familiarity. Analysis of these items produced the following main findings:

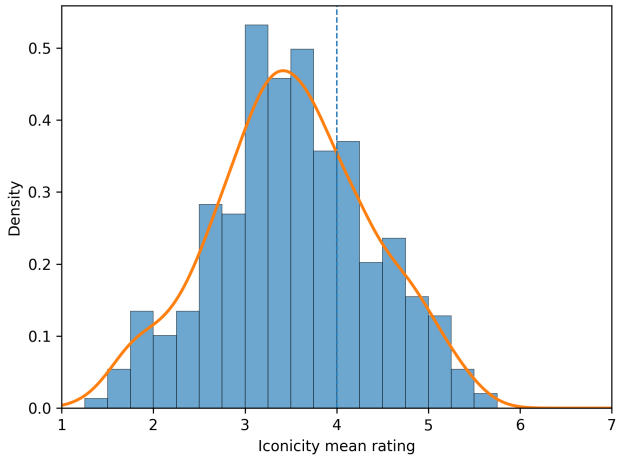
- Items near the middle of the scale showed greater disagreement. Spearman correlations between SD and absolute distance from the midpoint were $\rho = -.514$ for concreteness, $\rho = -.665$ for imageability, $\rho = -.412$ for iconicity, and $\rho = -.838$ for familiarity. Items that were clearly concrete, clearly non-imageable, or highly familiar were easier to judge, whereas intermediate items allowed different interpretations of the rating criteria. This effect was strongest for familiarity because most words in the dataset were familiar to native speakers, while the few intermediate items depended more on individual linguistic experience.

- High disagreement was related to metaphor type, although the pattern differed across variables.

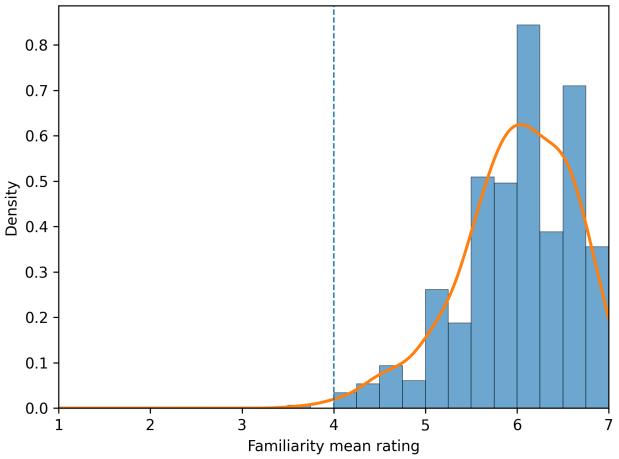


(a) Concreteness

(b) Imageability



(c) Iconicity



(d) Familiarity

Figure 4.2: Distributions of the item-level mean ratings for the four psycholinguistic dimensions. Histograms show the observed distributions of item means, the solid lines represent kernel density estimates, and the dashed vertical lines indicate the midpoint (4) of the seven-point rating scale.

As shown in Table 4.4, only 20.4–27.9% of literal items had high SD. For imageability, WIDLII items had the highest rate (40.0%), while for familiarity, PP items had the highest rate (40.0%). Metaphor status was significantly associated with high disagreement for imageability ($p = .010$) and familiarity ($p = .049$), but not for concreteness ($p = .942$) or iconicity ($p = .096$). This demonstrates that our rating task is inherently more challenging than conventional isolated rating tasks, since metaphorical meanings require participants to distinguish the contextual meaning from the more basic meaning.

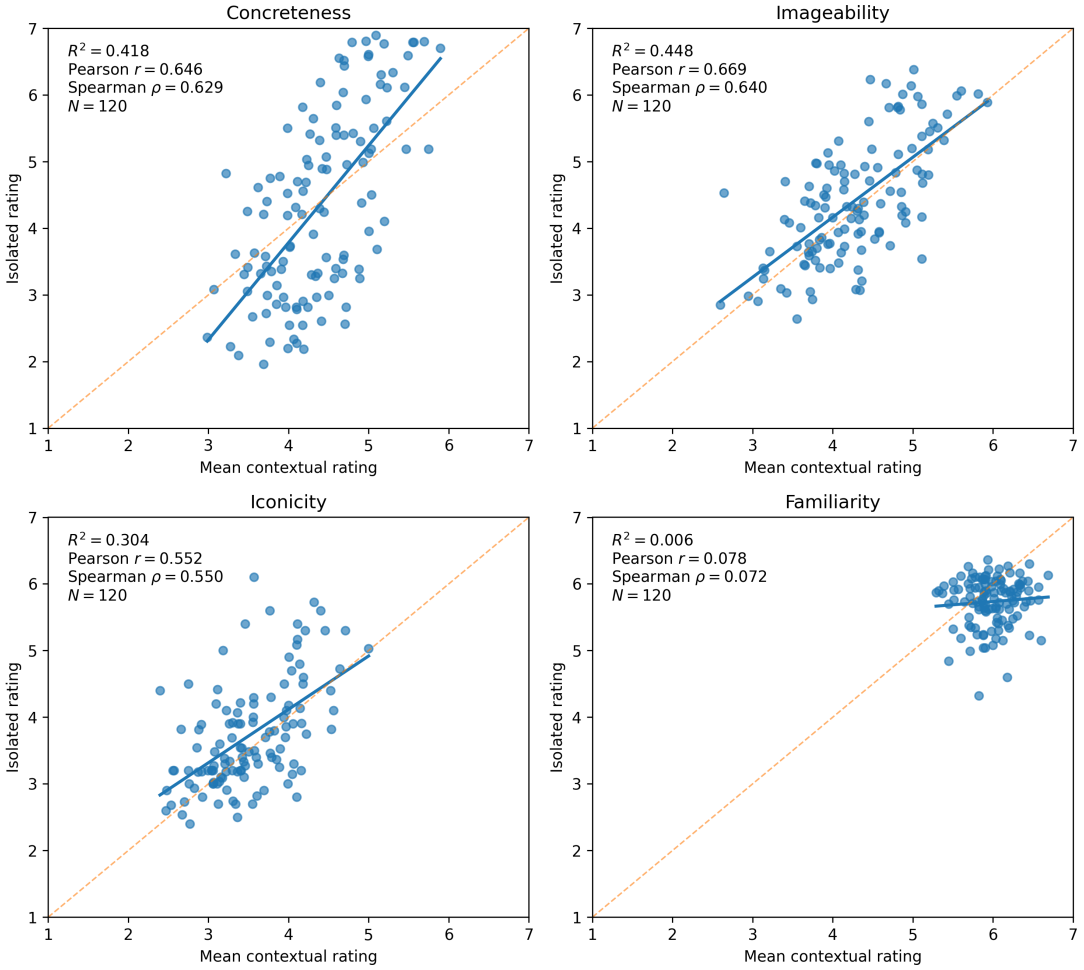


Figure 4.3: Comparison between contextual and isolated ratings for the four psycholinguistic dimensions. Each point represents one target lemma (averaged across all sentence contexts). The solid line shows the linear regression fit, and the dashed line indicates the identity line ($y = x$). Pearson's r, Spearman's ρ, and the corresponding R^2 values are shown in each panel.

Table 4.4: High-disagreement items by VUAMC metaphor type. Values show high-SD items over the total number of items in each category, with percentages in parentheses. The p-values were obtained from 2×2 chi-square tests with Yates' continuity correction, comparing the proportions of high-SD and non-high-SD items between literal and metaphorical groups.

Metaphor type	Concreteness	Imageability	Iconicity	Familiarity
Literal/unannotated	80/317 (25.2%)	64/313 (20.4%)	87/312 (27.9%)	71/328 (21.6%)
`mrw/met`	60/225 (26.7%)	65/218 (29.8%)	47/229 (20.5%)	57/208 (27.4%)
`mrw/met/PP`	1/22 (4.5%)	7/28 (25.0%)	6/26 (23.1%)	8/20 (40.0%)
`mrw/met/WIDLII`	7/27 (25.9%)	12/30 (40.0%)	7/26 (26.9%)	12/38 (31.6%)
`mrw/met/double`	0/1 (0.0%)	0/4 (0.0%)	1/1 (100.0%)	1/3 (33.3%)
χ^2 test p	.942	.010	.096	.049

- POS had no significant effect. The proportions of high-SD nouns, verbs, and adjectives were similar in all four dimensions, as shown in Table 4.5. Disagreement was therefore more closely related to the specific meaning and context than to grammatical category.

Table 4.5: High-disagreement items by part of speech.

Dimension	Noun	Verb	Adjective	p
Concreteness	47/200 (23.5%)	48/197 (24.4%)	53/196 (27.0%)	.698
Imageability	47/200 (23.5%)	48/200 (24.0%)	53/193 (27.5%)	.615
Iconicity	53/198 (26.8%)	44/196 (22.4%)	51/200 (25.5%)	.595
Familiarity	54/199 (27.1%)	52/198 (26.3%)	43/200 (21.5%)	.375

- Familiarity had the weakest agreement partly because the dataset contained few genuinely unfamiliar words. High-SD examples included *back pack*, *show myself very vicious*, and *mind of its own*. The words themselves are common, but familiarity with their particular senses, constructions, and registers differs across participants. Familiarity disagreement therefore mainly reflects sense- and construction-level exposure, rather than whether participants know the word.

- Polysemy and idiomaticity were frequent among the ten highest-SD items (listed in Appendix C): at least 7 items for concreteness, 8 for imageability, 5 for iconicity, and 6 for familiarity involved a polysemous, idiomatic, metaphorical, or phrasal interpretation. Examples include *put the cat among the pigeons*, *throw us a drink*, *set off*, *on the other hand*, *caught a flight*, and *mind of its own*. Also, we used original VUAMC excerpts, which contained incomplete sentences, spoken-language disfluencies, and insufficient local context, as in *Well should have won* and *Start thinking, got nothing lose, sta start*. These factors also made the intended contextual meaning difficult to determine.

Chapter 5

Study 2: LLMs rating of psycholinguistics variables

5.1 Correlation analysis

We next compare LLM-generated ratings with human judgments for the four psycholinguistic variables. The correlations between the four models' isolated ratings and human ratings are summarized in Table 5.1.

5.1.1 Isolated rating results

Table 5.1 shows that GPT-4o achieves the highest agreement with human lexical norms for concreteness and imageability, reaching around 0.8 correlations. Performance is much lower for iconicity and familiarity. Llama performs slightly better than Qwen, and outperforms GPT models for isolated familiarity rating. Full correlation results of each model and human annotation are in Appendix tables D.1, D.8, D.14 and D.20.

The correlation matrices across parts of speech are shown in Table D.2, D.11, D.21 and D.15. For concreteness, imageability, and iconicity, nouns consistently yield the highest correlations across models, while familiarity does not exhibit a consistent part-of-speech effect.

The largest isolated prediction errors are concentrated on words with more abstract or flexible meanings. For example, GPT-4o substantially underestimates the imageability of *current*, *speak*, *tough*, and *link*, and similarly underestimates the concreteness of *world*, *dark*, *market*, and *call* (Tables E.5 and E.1). By contrast, words such as *door*, *bed*, and *ring* show almost identical isolated ratings from the model and human annotators.

5.1.2 In-context rating results

All models achieve moderate correlations with human ratings for concreteness and imageability, while the correlations for iconicity and familiarity are substantially lower. GPT-4o consistently achieves the highest correlation with human annotations across all four variables, followed by GPT-4o-mini, while Qwen-2.5-Instruct performs slightly worse than Llama-3.1-Instruct. Correlations with human ratings for concreteness and imageability, while correlations for iconicity and familiarity are substantially lower.

To investigate whether our annotation quality affects model performance, we further divided the dataset into different agreement levels based on human annotation consistency. As shown in

Table 5.1: Spearman correlations (ρ) between LLM predictions and human ratings for isolated and in-context words. Best results for each property and setting are in bold.

Setting	Model	Con.	Ima.	Ico.	Fam.
Isolated	GPT-4o	**.860**	**.786**	**.517**	.230
	GPT-4o-mini	.726	.718	.472	.287
	Llama	.697	.625	.337	**.579**
	Qwen	.680	.539	.074	-.065
In-context	GPT-4o	**.543**	**.494**	**.349**	**.316**
	GPT-4o-mini	.438	.466	.203	**.316**
	Llama	.428	.391	.207	.263
	Qwen	.397	.369	.179	.258

Table D.5, D.13, D.19 and D.25, model-human correlations generally increase for items with higher human agreement.

We further evaluated GPT-4o on two existing in-context psycholinguistic datasets using exactly the same prompting setup: the CONcreTEXT concreteness dataset Montefinese, Gregori, Ravelli, Varvara, and Radicioni [Mon+23] and Pisani et al.'s in-context two-word familiarity dataset Pissani and Almeida [PA26]. To our knowledge, these are the only publicly available human-annotated in-context datasets for the psycholinguistic variables investigated in this work. On CONcreTEXT, GPT-4o achieves a Spearman correlation of $\rho = 0.801$. On Pisani et al.'s familiarity dataset, GPT-4o reaches Pearson and Spearman correlations of $r = 0.777$ and $\rho = 0.813$, respectively. These results are substantially higher than those obtained on our dataset.

The lower correlations observed on our dataset are likely explained by the greater difficulty of the annotation task rather than the limitation of the prompting strategy itself. Existing datasets were designed specifically for a single psycholinguistic variable and generally contain carefully selected examples with relatively unambiguous meanings. In contrast, our dataset was collected from naturally occurring sentences in the VU Amsterdam Metaphor Corpus and contains a much wider range of lexical items, contextual variability, and both literal and metaphorical usages. The annotation therefore involves more ambiguous contexts and greater individual variation, making agreement between both humans and models more challenging.

The relatively lower performance on iconicity and familiarity is also consistent with previous findings. Recent work has shown that LLMs generally estimate iconicity less accurately than concreteness [Mar+25]. One possible explanation is that concreteness primarily depends on perceptual grounding that is extensively encoded in linguistic data, whereas iconicity additionally requires sensitivity to non-arbitrary form-meaning correspondences. Familiarity is likewise inherently subjective, reflecting differences in individual experience and language exposure rather than lexical semantics alone. These characteristics make both variables more difficult to infer consistently from text, for both human annotators and language models.

5.1.3 In-context vs isolated ratings

We further examine the relationship between isolated and context-dependent ratings at the item level. Tables E.1-E.16 present the largest and smallest isolated prediction errors, isolated-to-context shifts, and context rating variabilities for GPT-4o across the four psycholinguistic variables.

A clear difference emerges across variables. Imageability is affected most strongly by context, with

words such as *rise*, *ring*, *face*, and *build* showing the largest shifts between isolated ratings and the mean context ratings (Table E.6). In contrast, familiarity is comparatively stable, with most words remaining close to their isolated ratings even when used in different contexts (Table E.14). Concreteness and iconicity lie between these two extremes.

Another interesting observation is that GPT-4o and human annotators do not always identify the same words as being most sensitive to context. For concreteness, GPT-4o shows the largest variation across contexts for words such as *hand*, *room*, and *face*, whereas human ratings vary most for *come*, *house*, *job*, and *make* (Tables E.3 and E.4). Similar mismatches are also observed for the other three variables. Some words, such as *hand*, consistently appear among the most variable items for GPT-4o but not for human annotators, suggesting that the model tends to assign stronger context effects to these words than humans do.

The item-level analyses also suggest that a large isolated-to-context shift and high context variability reflect different patterns. For example, *rise*, *ring*, and *cut* show some of the largest shifts between isolated and context ratings. However, they do not appear among the words with the highest context standard deviations. Words such as *hand*, *room*, and *body* show much larger variation across different contexts. These two measures therefore capture different aspects of context effects.

5.2 Influence of part of speech and metaphorical usage

We further analyse performance across different parts of speech and metaphor types. As shown in Table 5.2, for concreteness and imageability, correlations are highest for nouns among the three POS categories. As for metaphor types, correlations are much lower for metaphorical than literal usages, especially for imageability (.170 vs. .624). However, this pattern is not observed for familiarity and iconicity. Full correlation matrices are reported in Appendix tables D.2, D.11, D.15, D.21 for POS, and Tables D.6, D.12, D.18, D.24 for metaphor types.

Part of speech. As shown in Table 5.2, GPT-4o aligns best with human ratings on nouns on concreteness and imageability, substantially higher than verbs and adjectives.

One possible explanation comes from the nature of the psycholinguistic variables themselves. Nouns typically denote concrete entities with relatively stable semantic representations, whereas verbs and adjectives are more dependent on events and contextual interpretation. Previous psycholinguistic norming studies have likewise found that ratings for nouns are generally more reliable and exhibit higher inter-annotator agreement than those for other parts of speech [PYM68; BWK14]. Consequently, both human annotators and language models can estimate noun properties more consistently. In contrast, the meanings of verbs and adjectives are often more context-sensitive, making their concreteness and related psycholinguistic properties more difficult to judge.

Literal and metaphorical usage. Table 5.2 shows that on concreteness and imageability ratings, GPT-4o correlates much better with human ratings on literal usages than on metaphorical usages. This result is expected because metaphorical expressions involve semantic extension beyond the prototypical meaning of a word. While literal usages generally preserve the lexical meaning associated with a word, metaphorical usages often shift attention from perceptual properties to more abstract conceptual mappings. As a result, concreteness judgments become substantially more variable, making prediction more difficult for both humans and language models. Similar observations have been reported in

Table 5.2: GPT-4o's performance across linguistic categories, measured by Spearman correlations (ρ) with human in-context ratings. For each dimension, we keep each POS and meta/lit type balanced.

Property	Category	Spearman ρ
Concreteness	Adjective	.515
	Noun	.663
	Verb	.392
	Literal	.638
	Metaphorical	.417
Imageability	Adjective	.471
	Noun	.570
	Verb	.423
	Literal	**.624**
	Metaphorical	**.170**
Iconicity	Adjective	.345
	Noun	.304
	Verb	.286
	Literal	.362
	Metaphorical	.345
Familiarity	Adjective	.273
	Noun	.302
	Verb	.423
	Literal	.251
	Metaphorical	.284

previous studies of contextual concreteness, which found that contextual meaning can substantially alter psycholinguistic ratings compared with isolated words [Mon+23; MMG+25].

5.3 Error analysis

We list the sentences with 10 largest errors between gpt-4o-mini and human annotations in Table D.7. For these items, the model assigns concreteness ratings around three points lower than the human mean. Most are annotated as indirect metaphors in VUAMC.

One possible reason is that human ratings are still influenced by the basic meaning of the target word, even though participants were asked to focus only on its meaning in context. This can be seen by comparing metaphorical and literal uses of the same lemma (Table 5.3).

For the literal uses, GPT-4o-mini ratings are generally close to the human means. For the metaphorical uses, however, the model gives consistently lower concreteness ratings, while human ratings remain relatively high. One explanation is that human ratings appear to retain more influence from the basic meaning, and consistent with the source–target structure of conceptual metaphor, the meaning is a more concrete source concept used to express a more abstract target concept [LJ80].

Another group of large discrepancies includes *worry*, *worried*, and *surprised*. In *Do not* ***worry***, the human mean is 5.57, while the model gives 2.00. Human ratings are also high for *Some parents were very* ***worried*** and *He is perhaps* ***surprised*** *by her composure*, at 6.00 and 6.43. These psychological states are abstract under the usual definition of concreteness, since they cannot be directly perceived through the senses.

One possible explanation is that high human ratings may reflect overlap with imageability, familiarity, or experiential salience. Emotional states are not physical, but they can be felt clearly and vividly.

Table 5.3: Human and GPT-4o-mini concreteness ratings for metaphorical and literal uses of the same target words. The target word is shown in bold in each sentence.

Target	Metaphorical sentence	Human	LLM	Literal sentence	Human	LLM
direction	But its future **direction** remains unknown and indeterminate.	5.71	2.00	Enlargement shows a section through the lenses to show the **direction** of the field of view.	5.25	3.99
hand	On the other **hand**, sometimes the drama of the auction and the competition between bidders can have the opposite effect.	5.14	2.00	He has in one **hand** a stick and is followed by a thin dog.	6.10	6.85
shake	Through it, Palestinians are struggling on a daily basis to **shake** off Israel's authority and to construct their own.	6.14	3.00	The back of Lawton's head **shook**, his radio-projected voice crackling.	4.75	4.91
scale	Furthermore, the **scales** at which population estimates are often required mean that even EDs are too coarse for risk assessments.	5.71	3.00	Assessing those effects of global change which will be large **scale** and cause major modifications to both renewable and non-renewable resources.	4.13	3.80

Participants may therefore treat this felt experience as a form of concreteness. The same issue may apply to *clean* in *I searched him and his luggage, and he was **clean***, where the contextual meaning is "not carrying anything prohibited" rather than physically clean. In-context concreteness ratings may therefore capture not only perceptual accessibility, but also the vividness of the experience described.

Chapter 6

Study 3: Modelling metaphoric potential with psycholinguistic variables

This chapter examines whether psycholinguistic variables can account for the metaphoric potential of lexical items. Using a subset of VUAMC, we first reproduce the analysis conducted on isolated lexical norms in an unpublished manuscript, and then replace these norms with the GPT-predicted in-context ratings. Finally, we combine the isolated and contextual ratings in the same model to test whether contextual information contributes beyond conventional word-level norms.

The dependent variable is token-level metaphoricity in the VU Amsterdam Metaphor Corpus (VUAMC). Tokens annotated as `mrw` are treated as metaphorical, while tokens annotated as `-` are treated as non-metaphorical. The predictors of primary interest are concreteness, familiarity, imageability, and iconicity. Lemma frequency, part of speech, and semantic class are included as control variables where appropriate.

6.1 VUAMC subset extraction

We first integrated the four human annotation datasets we are using with the full VUAMC dataset. After annotating all the tokens with the four datasets, we removed tokens with missing values for any of the four psycholinguistic variables, semantic class, or lemma frequency. Concreteness was transformed from its original 1-5 scale to a 100-700 scale. Familiarity, imageability, and iconicity were represented on the same approximate scale, and lemma frequency was log-transformed. Part-of-speech categories were simplified into noun, verb, adjective, adverb, and function word.

After preprocessing, the complete dataset contained 134,928 token instances, of which 19,155 were metaphorical and 115,773 were non-metaphorical. Metaphorical tokens therefore accounted for 14.20% of the complete filtered dataset.

Running the LLM-based contextual rating procedure over the entire dataset would require four separate predictions for more than 130,000 token instances. We therefore extracted approximately 10,003 tokens for contextual rating, while keeping the original distribution of met/lit label and psycholinguistics ratings. Full sentences were recovered and validated from the official VUAMC release on Huggingface [1] using document ID, sentence ID, and token position.

We first compared the sampled subset with the complete filtered VUAMC data. Figure 6.1 presents

[1] https://huggingface.co/datasets/matejklemen/vuamc

the proportions of metaphorical and non-metaphorical tokens in the subset and full datasets. Figure 6.2 compares the part-of-speech distributions. The distributions of both met/non-met and POS are basically the same, except that the sampled data excluded all function words, focusing primarily on content words for which the relevant psycholinguistic norms were available.

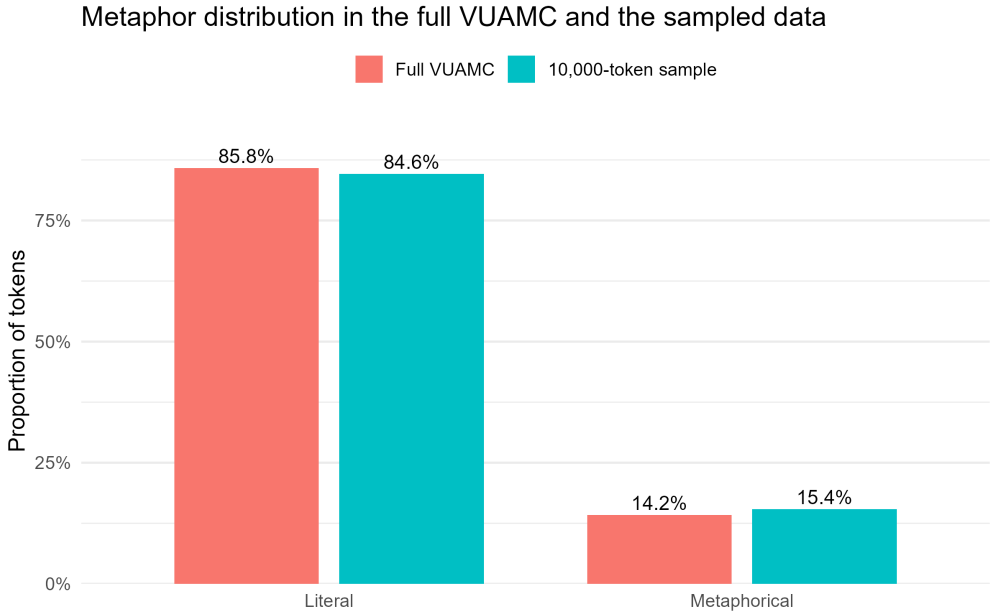


Figure 6.1: Proportions of metaphorical and non-metaphorical tokens in the complete filtered VUAMC data and the sampled subset.

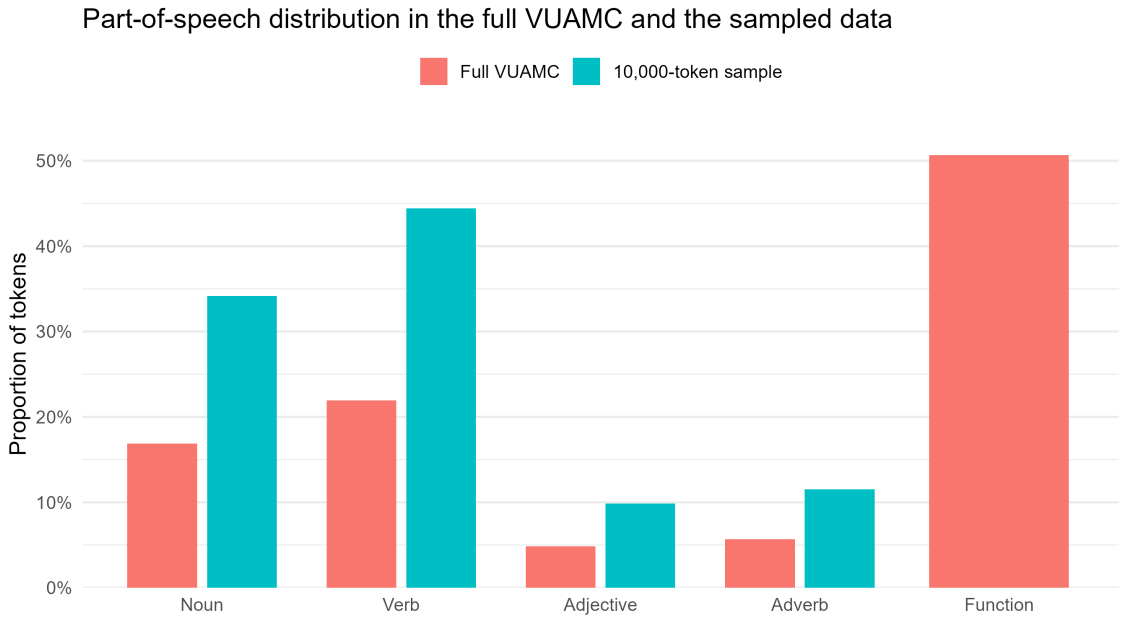


Figure 6.2: Part-of-speech distributions in the complete filtered VUAMC data and the sampled subset.

Figure 6.3 further compares the distributions of the four isolated psycholinguistic variables in our extracted dataset and full VUAMC dataset. Considering we mainly focus on content words, the sample's mean concreteness was higher than that of the full dataset (376.89 versus 328.61), as was its mean imageability (382.43 versus 329.35) and iconicity (338.04 versus 305.30). Despite these differences,

the sampled dataset is still a representative subset of the content words in the full natural VUAMC dataset, retaining a similar metaphor proportion, POS and the original isolated ratings distributions on each psycholinguistic dimension.

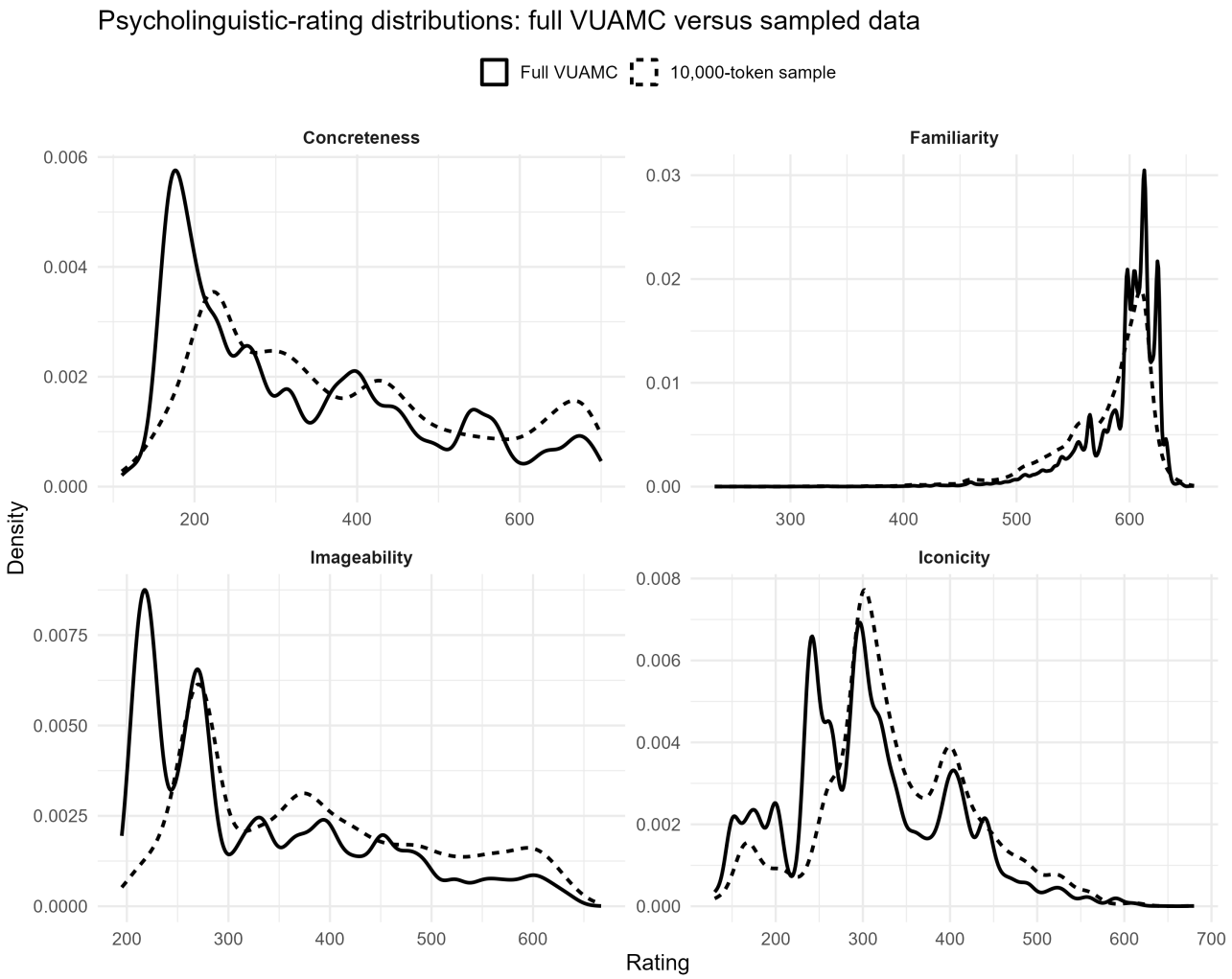


Figure 6.3: Distributions of isolated concreteness, familiarity, imageability, and iconicity in the complete filtered VUAMC data and the sampled subset.

6.2 Cross-correlation matrix

For every sampled token, we used GPT-4o to predict its in-context concreteness, imageability, familiarity, and iconicity on a 1-7 scale. We used the same prompt in Appendix A (in-context prompt 2). We calculated a probability-weighted score from the probabilities assigned to all seven possible ratings to get a continuous rating, which is more accurate and align with human annotations better than the dominant rating.

After getting the in-context ratings from GPT-4o, we first examined correlations among both the isolated and in-context ratings of all the four dimensions. Figure 6.5 shows the full Pearson correlation results on our sampled dataset, while Figure 6.4 shows the isolated ratings of the full VUAMC dataset, as in the previous unpublished manuscript.

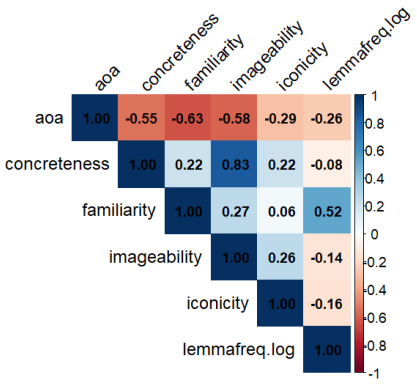


Figure 6.4: Pearson cross-correlation matrix of isolated lexical norms of the full VUAMC dataset.

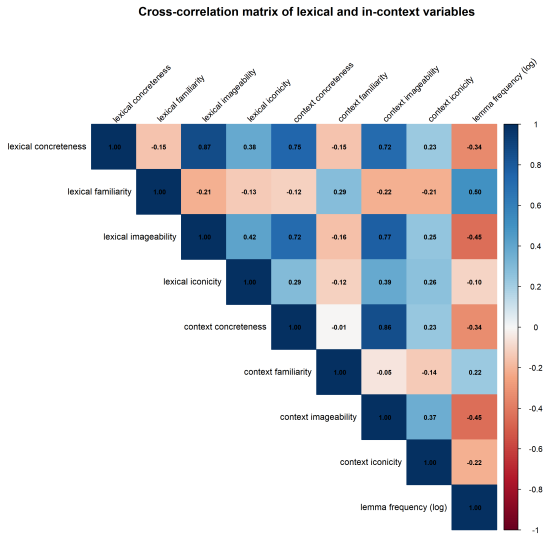


Figure 6.5: Pearson cross-correlation matrix of isolated lexical norms, GPT-predicted in-context ratings, and log lemma frequency of the sampled dataset.

We first compare the correlations between the isolated variables and their contextual counterparts. The correspondence was relatively strong for concreteness ($r = .75$) and imageability ($r = .77$), but substantially weaker for familiarity ($r = .29$) and iconicity ($r = .26$). These results suggest that different psycholinguistic variables exhibit different degrees of context dependence. Concreteness and imageability largely preserve their lexical properties across contexts, whereas familiarity and iconicity are considerably more sensitive to the specific contextual realization of a word.

To better understand these discrepancies, we examined the items with the largest absolute differences between isolated norms and contextual ratings (Table 6.1). For concreteness and imageability, the largest discrepancies occurred where the contextual meaning differed substantially from the canonical lexical sense represented by the isolated norms. Many examples involve polysemous words or named entities. For example, House and Cage receive highly concrete and imageable isolated ratings because they are typically interpreted as common nouns, but in context they function as surnames (John House, John Cage). Likewise, chair in "held a chair in physics" refers to an academic position rather than a piece of furniture, and pig in "a pig to deal with" expresses an idiomatic meaning rather than the animal itself. These examples illustrate an inherent limitation of isolated lexical norms: they assign a

single rating to all senses of a word, whereas contextual ratings distinguish between individual sense realizations.

After excluding sense disambiguation cases, the remaining largest discrepancies are primarily metaphorical uses. In these contexts, words consistently receive lower concreteness and imageability ratings than their isolated norms, reflecting the increased semantic abstraction introduced by metaphorical interpretation. This pattern supports the motivation for collecting contextual psycholinguistic norms, particularly for metaphor-rich corpora such as VUAMC.

Then, we examine the correlations between different contextual variables. Contextual concreteness and contextual imageability correlated at $r = .86$, almost identical to the correlation between their lexical counterparts. Moreover, lexical concreteness remained strongly associated with contextual imageability ($r = .72$), and lexical imageability showed an equally strong correlation with contextual concreteness ($r = .72$), indicating that the close relationship between these two semantic dimensions is preserved after contextualization. Contextual imageability was moderately correlated with contextual iconicity ($r = .37$), while contextual familiarity showed only weak correlations with the other contextual variables, suggesting that familiarity becomes relatively independent once ratings are made in context.

Lemma frequency was positively associated with lexical familiarity ($r = .50$), but its correlation with contextual familiarity was substantially weaker ($r = .22$). Frequency was negatively correlated with both lexical and contextual imageability ($r = -.45$ in both cases) and with concreteness ($r = -.34$ for both lexical and contextual ratings). Thus, although contextual ratings modify the absolute values assigned to individual words, they largely preserve the broader psycholinguistic organization of the lexical system.

6.3 Binomial logistic regression

Because metaphoricity is a binary outcome, variance explained in the ordinary linear-regression sense is not directly available. We fitted binomial logistic regression models predicting whether each token was metaphorical or literal. Model parameters were estimated by maximum likelihood.

Three models were fitted to the 10,003-token dataset. The *lexical model* contained the four isolated psycholinguistic norms. The *contextual model* replaced these variables with the four GPT-predicted in-context ratings. The *combined model* included both sets of ratings. All three models additionally controlled for lemma frequency, part of speech, and semantic class.

We first compare the models using log-likelihood, the Akaike Information Criterion (AIC), and the area under the ROC curve (AUC) (Table 6.2). In general, lower AIC indicates a better balance between model fit and complexity, while higher AUC indicates better discrimination between metaphorical and non-metaphorical tokens. The contextual model fitted the data better than the lexical model. And the combined model performed considerably better than both single-representation models.

Table 6.2: Model fit for the lexical, contextual, and combined models.

Model	N	Log-likelihood	AIC	AUC
Lexical	10,003	−3852.54	7763.08	.737
Contextual	10,003	−3754.25	7566.49	.753
Combined	10,003	−3525.15	7116.29	.794

To evaluate the generalization ability to unseen lexical items of the three models, we also tested

Table 6.1: Top five items with the largest discrepancies between isolated lexical norms and contextual ratings for each psycholinguistic variable.

(a) Concreteness

Word	Isolated	Context	Δ	Sentence
House	700	100.0	600.0	John **House**, who had organised the exhibition, came almost leaping down the stairs accompanied by a smallish woman in a pine-green tent-like coat.
Cage	700	100.2	599.8	A few years later Thomson recognised **Cage** as "the most original composer in America"...
Cage	700	100.2	599.8	He anticipated some of the enthusiasms of John **Cage**, whom he first met in 1941.
House	700	101.5	598.5	...the **House** of Lords, Law sided with the hedgers rather than the ditchers.
tires	700	101.5	598.5	"And there's no reason for Elinor to see any of this if it **tires** her."

(b) Imageability

Word	Isolated	Context	Δ	Sentence
pig	635	106.4	528.6	And sometimes non-standard discount is given, a **pig** to deal with on the old system, but if you could add one extra input box on that screen there ...
Keys	618	103.1	514.9	Salford **Keys**.
Bell	610	104.2	505.8	By ROGER **BELL**.
chair	610	104.3	505.7	He held a **chair** in physics from 1937 until his retirement, publishing more than 150 papers on atomic and nuclear physics.
House	606	100.5	505.5	John **House**, who had organised the exhibition, came almost leaping down the stairs accompanied by a smallish woman in a pine-green tent-like coat.

(c) Familiarity

Word	Isolated	Context	Δ	Sentence
ca	620	100.1	519.9	You ... comes at six thirty, you **ca** n't have it, we're bloody cooking ...
ca	620	100.7	519.3	... you just **ca** n't get anywhere.
ca	620	101.9	518.1	I bet you feel like a ... they **ca** n't truss them ...
ca	620	102.2	517.8	With other illustrators ... I **ca** n't pester them any more ...
ca	620	103.7	516.3	The Labour Party Conference is on so I **ca** n't go.

(d) Iconicity

Word	Isolated	Context	Δ	Sentence
sighs	68000	491.5	67508.5	Mr Dobbs **sighs**: "This is a typically British bureaucratic response."
banging	67500	105.7	67394.3	But I mean they were **banging** fifteen hundred in for that disco ...
snapped	64000	177.4	63822.6	"Help me secure the stairs behind us," he **snapped** at Lawton, "as much as we can."
swept	62000	237.8	61762.2	His small bright blue eyes **swept** disparagingly over the minimalized black and white flooring, stainless steel chairs and black wooden bar.
gloom	62000	303.8	61696.2	The gangway lamps seemed to give no light, only add further to the **gloom** of the dusk.

them on a lemma-disjoint train-test split (4:1). Results are shown in Table 6.3. The lexical model achieved the highest accuracy, but the lowest recall, which indicated it obtained high accuracy mainly by predicting the majority literal class. The contextual model nearly doubled metaphor recall, from .063 to .122, but at the cost of substantially lower precision and accuracy. It produced more metaphor predictions but also more false positives. The combined model provided the strongest balance in generalization test.

Table 6.3: Prediction performance on the lemma-disjoint test set. Precision, recall, and F1 are calculated for the metaphorical class.

Model	Accuracy	Precision	Recall	F1	AUC
Lexical	.866	.395	.063	.109	.660
Contextual	.779	.129	.122	.125	.653
Combined	.837	.333	.252	.287	.737

6.4 Stepwise regression

To separately examine the relative contribution of individual predictors, we performed forward stepwise logistic regression using AIC as the selection criterion. Starting from a null model containing only the intercept, predictors were added sequentially according to the reduction in AIC they produced. The procedure terminated when adding further predictors no longer improved model fit. Unlike the full logistic regression, which estimates the effect of each predictor while controlling for all others, the stepwise analysis provides an indication of the order and relative contribution with which different psycholinguistic variables improve the explanation of metaphoricity.

Table 6.4: Decrease in information loss when adding predictors to the lexical model in the forward stepwise regression procedure.

Rank	Factor	AIC	$\Delta \downarrow$
0	(none)	13867.1	—
1	Familiarity	8482.9	5384.2
2	Semantic class	7993.1	489.8
3	Iconicity	7848.8	144.3
4	Part-of-speech	7810.9	37.9
5	Concreteness	7775.0	35.9
6	Imageability	7761.1	13.9

Table 6.5: Decrease in information loss when adding predictors to the contextual model in the forward stepwise regression procedure.

Rank	Factor	AIC	$\Delta \downarrow$
0	(none)	13867.1	—
1	Contextual familiarity	8392.9	5474.2
2	Semantic class	7964.0	428.9
3	Contextual concreteness	7762.5	201.5
4	Contextual imageability	7663.9	98.6
5	Part-of-speech	7597.3	66.6
6	Contextual iconicity	7577.8	19.5
7	Lemma frequency	7566.5	11.3

Table 6.6: Decrease in information loss when adding predictors to the combined model in the forward stepwise regression procedure.

Rank	Factor	AIC	$\Delta \downarrow$
0	(none)	13867.1	—
1	Contextual familiarity	8392.9	5474.2
2	Semantic class	7964.0	428.9
3	Contextual concreteness	7762.5	201.5
4	Lexical concreteness	7397.5	365.0
5	Lexical iconicity	7275.2	122.3
6	Lexical familiarity	7178.5	96.7
7	Part-of-speech	7134.8	43.7
8	Contextual imageability	7116.5	18.3
9	Contextual iconicity	7115.1	1.4
10	Lexical imageability	7114.9	0.2

We performed three separate forward stepwise logistic regression analyses. The lexical analysis (Table 6.4) considered only the conventional lexical psycholinguistic norms, the contextual analysis (Table 6.5) considered only the GPT-predicted in-context ratings, and the combined analysis (Table 6.6) included both sets of predictors. Across all three models, familiarity was the first psycholinguistic predictor selected by the forward stepwise procedure. In the lexical model, lexical familiarity reduced the AIC by 5384.2 points, while contextual familiarity achieved an even larger reduction of 5474.2 points in both the contextual and combined models. This indicates that familiarity provides the strongest single contribution to explaining metaphoricity, regardless of whether it is measured in isolation or in context.

The contextual model showed a clear ordering among the contextual psycholinguistic variables. After contextual familiarity and semantic class, contextual concreteness produced the largest additional improvement (ΔAIC = 201.5), followed by contextual imageability (98.6), whereas contextual iconicity (19.5) and lemma frequency (11.3) contributed relatively modest gains. This suggests that contextual concreteness and imageability capture substantially more information about metaphoricity than the remaining contextual predictors.

The combined model demonstrates that lexical and contextual ratings are complementary rather than redundant. Although contextual familiarity and contextual concreteness entered first, lexical concreteness subsequently produced the largest remaining improvement in model fit (ΔAIC = 365.0), followed by lexical iconicity (122.3) and lexical familiarity (96.7). Contextual imageability and contextual iconicity continued to improve the model after the lexical predictors had entered, while lexical imageability was retained as the final predictor. The retention of variables from both representations indicates that contextual ratings do not replace traditional lexical norms but instead provide complementary information for modelling metaphoricity.

6.5 Multifactorial analysis

To quantify the independent contribution of each predictor, we fitted multifactorial binomial logistic regression models including all psycholinguistic variables simultaneously, together with lemma frequency, part of speech, and semantic class. Separate analyses were conducted for the lexical, contextual, and combined predictor sets. Model coefficients were exponentiated and reported as odds ratios (OR),

which indicate the multiplicative change in the odds of metaphorical usage associated with a one-unit increase in each predictor while holding all other predictors constant. Statistical significance was assessed using Wald tests.

Odds ratios for the four psycholinguistic ratings are reported for an increase of 100 units, corresponding to one point on the original rating scale. The effect of frequency is expressed for a tenfold increase in lemma frequency. An odds ratio above 1 indicates an increase in the odds of metaphorical usage, while an odds ratio below 1 indicates a decrease.

Table 6.7: Effects of isolated psycholinguistic variables on the likelihood of metaphorical usage in the complete filtered VUAMC data. Noun is the reference level for part of speech. *** indicates $p < .001$.

Variable	Category or unit	Odds ratio
Part of speech	Noun	1.00
	Verb	1.12***
	Adjective	1.59***
	Adverb	0.70***
	Function word	1.22***
Concreteness	Per one-point increase	1.46***
Imageability	Per one-point increase	0.54***
Familiarity	Per one-point increase	0.60***
Iconicity	Per one-point increase	1.37***
Frequency	Per tenfold increase	0.82***

Table 6.8: Effects of isolated psycholinguistic variables on metaphorical usage in the 10,003-token contextual subset. Noun is the reference level for part of speech. *** indicates $p < .001$.

Variable	Category or unit	Odds ratio
Part of speech	Noun	1.00
	Verb	1.57***
	Adjective	1.81***
	Adverb	1.04
Concreteness	Per one-point increase	1.32***
Imageability	Per one-point increase	0.76***
Familiarity	Per one-point increase	0.45***
Iconicity	Per one-point increase	1.53***
Frequency	Per tenfold increase	0.94

Table 6.9: Effects of GPT-predicted in-context psycholinguistic ratings on the likelihood of metaphorical usage. Noun is the reference level for part of speech. *** indicates $p < .001$.

Variable	Category or unit	Odds ratio
Part of speech	Noun	1.00
	Verb	1.11
	Adjective	0.86
	Adverb	0.59***
Contextual concreteness	Per one-point increase	0.59***
Contextual imageability	Per one-point increase	1.52***
Contextual familiarity	Per one-point increase	0.73***
Contextual iconicity	Per one-point increase	1.43***
Frequency	Per tenfold increase	0.86***

Table 6.10: Combined effects of isolated lexical norms and GPT-predicted in-context ratings on metaphorical usage. Noun is the reference level for part of speech. *** indicates $p < .001$.

Variable	Category or unit	Odds ratio
Part of speech	Noun	1.00
	Verb	1.40***
	Adjective	1.13
	Adverb	0.82
Lexical concreteness	Per one-point increase	1.72***
Lexical imageability	Per one-point increase	0.83
Lexical familiarity	Per one-point increase	0.54***
Lexical iconicity	Per one-point increase	1.49***
Contextual concreteness	Per one-point increase	0.49***
Contextual imageability	Per one-point increase	1.31***
Contextual familiarity	Per one-point increase	0.84***
Contextual iconicity	Per one-point increase	1.18
Frequency	Per tenfold increase	0.91

Compared with the complete VUAMC corpus (Table 6.7), the sampled dataset (Table 6.8) shows a highly consistent pattern of effects. The direction of all four psycholinguistic variables remains unchanged: concreteness and iconicity increase the likelihood of metaphorical usage (OR = 1.32 and 1.53), whereas imageability and familiarity decrease it (OR = 0.76 and 0.45). The main differences are quantitative rather than qualitative. The effects of frequency and adverbs are no longer statistically significant in the sampled dataset, while the effects of verbs and adjectives become stronger (OR = 1.57 and 1.81, compared with 1.12 and 1.59 in the full VUAMC). These results suggest that the 10,003-token subset preserves the main psycholinguistic patterns observed in the full corpus, providing a reliable basis for the subsequent contextual analyses.

The largest difference between the lexical and contextual models concerns concreteness and imageability. Concreteness changes from a positive predictor in the lexical model (OR = 1.32) to a negative predictor in the contextual model (OR = 0.59), and both effects remain significant in the combined model (OR = 1.72 and 0.49). Imageability shows the opposite pattern, changing from a negative lexical effect (OR = 0.76) to a positive contextual effect (OR = 1.52), while only contextual imageability remains significant after both representations are included. This difference follows naturally from the distinction between lexical norms and contextual ratings. Lexical ratings describe a word's basic meaning, whereas contextual ratings reflect the meaning expressed in a particular sentence. According to MIPVU, metaphor is identified through a contrast between a more concrete basic meaning and a more abstract contextual meaning. It is therefore expected that words with higher lexical concreteness are more likely to be extended metaphorically, while metaphorical uses themselves receive lower contextual concreteness ratings. A similar explanation applies to imageability. Highly imageable lexical items are often concrete words that mainly occur literally, whereas contextual imageability reflects how vividly the contextual meaning can be mentally represented. These results suggest that lexical norms and contextual ratings capture different stages of metaphor representation rather than the same psycholinguistic property.

In contrast, familiarity and iconicity are considerably more stable across models. Both lexical familiarity (OR = 0.45) and contextual familiarity (OR = 0.73) remain significant in the combined model, indicating that familiarity with a word and familiarity with its contextual meaning each make an independent contribution to metaphoricity. This is consistent with the MIPVU assumption that

metaphor involves a familiar basic meaning and a less familiar contextual interpretation. Iconicity shows a different pattern. Lexical iconicity remains a significant positive predictor in both the lexical and combined models (OR = 1.53 and 1.49), whereas contextual iconicity is no longer significant after lexical iconicity is included. This suggests that iconicity is largely a stable lexical property, with comparatively little additional information introduced by context. The effect of frequency also becomes weaker than in the full VUAMC corpus and disappears in the combined model, probably because the sampled dataset contains only content words and excludes many high-frequency function words that contributed to the original corpus-level effect.

Taken together, the combined model shows that lexical norms and contextual ratings are complementary rather than interchangeable. Lexical norms capture relatively stable properties that underlie a word's metaphoric potential, whereas contextual ratings reflect how those properties are realised in a specific context. Modelling both levels together therefore provides a more complete account of metaphoricity than either representation alone.

Chapter 7

Discussion and limitations

7.1 Further discussion of LLM prediction

Potential biases in LLM ratings. We observe different rating biases across psycholinguistic properties. For isolated familiarity, GPT-4o and Qwen assign a rating of 7 to over 97% of the words, leaving little variation in their predictions. However, GPT-4o achieves a much higher correlation on the contextual familiarity dataset of Pissani and Almeida [PA26] ($\rho = .813$) than on ours ($\rho = .316$), suggesting that our human ratings' ceiling effect may also contribute to the low agreement. For iconicity, GPT-4o consistently gives ratings more than two points below human judgments. Hagihara and Miyazawa [HM26] also found relatively low agreement for iconicity, suggesting that models may judge form–meaning resemblance differently from humans. Lastly, our error analysis suggests that ratings systematically differ for specific semantic domains (e.g. psychological states).

Metaphorical usage seems to affect the prediction of imageability and concreteness. For imageability, the correlation between GPT-4o and human ratings is much lower for metaphorical than literal usages (.170 vs. .624), A similar, though smaller, difference is observed for concreteness (.417 vs. .638; Table 5.2). Reijnierse, Burgers, Krennmayr, and Steen [Rei+19] found that participants gave substantially lower concreteness ratings when metaphorical rather than non-metaphorical meanings were activated. In our data, however, this pattern is more pronounced for GPT-4o than for humans. This may be related to how word meanings are presented, as Reijnierse et al. used definitions to activate specific meanings, whereas we only highlighted target words in sentences. Zhong, Todd, Xu, and Brehm [Zho+25] found that LLMs rely more heavily on other psycholinguistic cues when making lexical judgments. Our results for imageability and concreteness suggest that metaphorical usage might also serve as a cue for model ratings.

7.2 Complementary effects of lexical and contextual ratings

The results in study 3 show that lexical norms and contextual ratings are complementary in metaphor modelling. The contextual model fitted the observed data better than the lexical model, while the combined model achieved the best overall fit, with an AUC of .794. Adding the four contextual ratings to the lexical model significantly improved model fit, and adding the lexical ratings to the contextual model produced a similarly significant improvement. In the lemma-disjoint evaluation, the combined

model also achieved the highest AUC (.737) and F1 score (.287), outperforming both the lexical and contextual models.

This result suggests that lexical and contextual ratings represent different aspects of metaphoric potential. Lexical norms capture relatively stable properties of a word and its basic meaning, whereas contextual ratings reflect how these properties are realised in a particular sentence. This distinction is consistent with previous accounts that separate relatively stable metaphoric potential from its actual contextual realisation [Cam97; Cam18; Mül08]. It also follows the logic of MIPVU, in which metaphor is identified by comparing the contextual meaning of a lexical unit with its more basic meaning [Ste+10]. Modelling the two levels together therefore provides a more complete account of metaphorical usage.

7.3 Effects of each lexical variable in metaphor modelling

7.3.1 Concreteness

The results for concreteness align well with our hypothesis. Lexical concreteness positively predicts metaphorical use (OR = 1.32), and this effect becomes stronger in the combined model (OR = 1.72). By contrast, contextual concreteness negatively predicts metaphorical use (OR = 0.59), with a stronger effect in the combined model (OR = 0.49). Thus, the more concrete the basic meaning is, the more likely the word is to be extended metaphorically, whereas the more concrete its meaning is in the current context, the less likely the token is to be metaphorical.

The effective cue may therefore be the contrast between lexical and contextual concreteness rather than either rating alone:

$$\Delta C = C_{\text{lexical}} - C_{\text{contextual}}.$$

This aligns with Conceptual Metaphor Theory: words with more concrete basic meanings tend to be used metaphorically, and when they are used metaphorically, their meanings become more abstract in context. The original meaning provides the concrete source domain, which supports the abstract target meaning expressed in the current context [LJ80].

7.3.2 Imageability

The results for imageability are contrary to our initial hypothesis. As discussed in Section 2.3.1, highly imageable source concepts may maintain vivid perceptual representations and thereby support metaphorical mapping or make the mapping more noticeable [GIB06; JJ23]. However, higher lexical imageability did not increase the likelihood of metaphorical use. Lexical imageability showed a negative effect in the lexical model (OR = 0.76) and became non-significant in the combined model. By contrast, contextual imageability positively predicted metaphorical use (OR = 1.52), and remained significant in the combined model (OR = 1.31).

An important factor here is the strong correlation between concreteness and imageability. Contextual concreteness and contextual imageability were correlated at $r = .86$, while the lexical versions showed a similarly strong correlation. In the multifactorial regression, the imageability coefficient therefore does not represent its overall relationship with metaphorical use, but its remaining effect after concreteness, familiarity, iconicity, and the other predictors are controlled.

The positive contextual effect may indicate that metaphorical contextual meanings can remain highly imageable even when they are relatively abstract. Expressions such as *grasp an idea*, *a bright future*, and *the argument collapsed* refer to abstract target meanings, but their source domains still provide clear mental scenes. This is compatible with the view that imageable source meanings support embodied or perceptual simulation during metaphor comprehension [GIB06].

However, the negative lexical effect does not support the stronger hypothesis that more imageable source words are generally more likely to be extended metaphorically. Its non-significance in the combined model suggests that much of the information provided by lexical imageability may already be captured by lexical concreteness, lexical iconicity, and contextual imageability. To obtain a clearer estimate of the independent effect of imageability, future analyses could compare separate concreteness and imageability models, or use residualized imageability after removing the variance shared with concreteness.

7.3.3 Familiarity and frequency

Familiarity turned out to be the most predictive variable in our stepwise regression. Lexical familiarity entered the lexical model first (ΔAIC = 5384.2), while contextual familiarity entered the contextual and combined models first (ΔAIC = 5474.2). In the full combined model, both lexical familiarity (OR = 0.54) and contextual familiarity (OR = 0.84) negatively predicted metaphorical use. Thus, the more familiar a lexical or contextual meaning is, the less likely the token is to be metaphorical.

The contextual familiarity effect aligns with the hypothesis that unfamiliar or less conventional metaphorical meanings retain stronger metaphorical salience (see Section 2.3.1). As a metaphorical meaning becomes more familiar through repeated use, the original source meaning becomes less salient and the metaphorical mapping becomes less noticeable [Goa97; JJ23]. This is also consistent with corpus evidence that metaphorically used verbs tend to occur with less familiar semantic contexts than literal verbs [AST24]. The negative lexical familiarity effect, however, may additionally reflect that highly familiar words have a larger proportion of common literal uses.

The results also support our hypothesis that contextual familiarity is more informative than lexical frequency. Frequency was not significant in the combined model, whereas both lexical and contextual familiarity remained significant. Frequency reflects the distribution of a word form, while familiarity also captures language users' experience with particular meanings [SCK12; TT11]. However, the stepwise order only shows that familiarity is the strongest single predictor in these models; it does not establish that familiarity has the strongest causal effect.

7.3.4 Iconicity

Our hypothesis was that an iconic form reinforces the basic or source meaning, allowing it to remain perceptually salient when the word is extended metaphorically (see Section 2.3.1). This may make the contrast between the basic and contextual meanings easier to recognise and increase the likelihood that the use is identified as metaphorical [LJ80; Ste+10].

The results mainly support this hypothesis at the lexical level. Lexical iconicity positively predicted metaphorical use in the lexical model (OR = 1.53) and remained significant in the combined model (OR = 1.49). Contextual iconicity also showed a positive effect in the contextual model (OR = 1.43), but became non-significant in the combined model (OR = 1.18).

This suggests that the relevant relation is mainly between the word form and its lexical or basic meaning, rather than between the word form and its metaphorical contextual meaning. The iconic form strengthens the perceptual representation of the source meaning, which may remain salient when the word is used in an abstract context. Contextual ratings therefore provide relatively little additional information for iconicity.

However, contextual iconicity was also one of the most difficult variables to measure in Studies 1 and 2, and its isolated–contextual correspondence was relatively weak. The non-significant contextual effect may therefore partly result from measurement error, rather than showing that iconicity is completely unaffected by context.

7.3.5 Part of speech

Part of speech also influenced metaphorical use. In the full VUAMC model, verbs (OR = 1.12) and adjectives (OR = 1.59) were more likely to be metaphorical than nouns, while adverbs were less likely to be metaphorical (OR = 0.70). In the sampled subset, the effects were stronger for verbs (OR = 1.57) and adjectives (OR = 1.81). However, after lexical and contextual psycholinguistic ratings were included in the combined model, only the verb effect remained significant (OR = 1.40).

This is broadly consistent with previous corpus findings that verbs have a higher within-category proportion of metaphorical uses than nouns [Kre11]. Nouns often provide clear and concrete source-domain concepts, which may make nominal metaphors easier to recognise, but nouns also occur frequently in literal uses [Cam03; JJ23]. Verbs, by contrast, commonly realise metaphor through relational extensions between predicates and their arguments. Their remaining effect in the combined model suggests that this predicate–argument structure is not fully captured by the psycholinguistic variables.

7.4 Limitations

Dataset. The human annotation dataset is relatively small, with approximately 600 sentences for each psycholinguistic variable. This might also cause the ceiling effect in familiarity rating, as discussed in Section 7.1. Compared with established norming datasets, the number of ratings per item is also relatively low, which may have affected rating reliability. The questionnaire was also demanding, as participants had to rate approximately 100 sentences and distinguish contextual meanings from the more general meanings of words. Collecting more ratings per item and reducing the questionnaire length could help improve consistency.

Our dataset is limited to English nouns, verbs, and adjectives from VUAMC. Some original corpus sentences were incomplete or ambiguous, making it difficult even for human annotators to determine the intended meaning.

The isolated ratings are older and not from the same set of participants, which may affect the comparison with contextualized ratings.

LLM evaluation. Our evaluation was limited to four models. The probability-based metric that we use is more effective than eliciting ratings directly [Con+25a], but does not work with reasoning models as the reasoning trace biases the probabilities too much. In addition, the relatively low human agreement for some items makes model performance difficult to interpret. A low correlation with

human ratings does not necessarily mean that the model fails to capture the contextual meaning, since human annotators may also interpret the word or the rating criteria differently. A larger and more consistent human dataset is therefore needed to evaluate LLM predictions more conclusively.

Chapter 8

Conclusion

This thesis investigates the effects of context-sensitive psycholinguistic ratings in the modelling of metaphoric potential. We focus on concreteness, imageability, familiarity, and iconicity, and combine human annotation, LLM-based prediction, and multifactorial regression on metaphor-annotated corpus data.

RQ1: How do humans rate psycholinguistic variables at the token level in sentential context?

In Study 1 (Chapter 4), we introduce a human-annotated dataset of in-context concreteness, imageability, familiarity, and iconicity ratings using naturally occurring literal and metaphorical sentences from the VU Amsterdam Metaphor Corpus. The dataset provides token-level psycholinguistic norms that capture variation across different contextual meanings of the same word. The item-level ratings show reasonable stability, although agreement is lower for familiarity and for ambiguous or idiomatic expressions.

RQ2: How well can LLMs predict isolated and in-context psycholinguistic ratings?

In Study 2 (Chapter 5), GPT-4o achieves the highest agreement with human judgments, with stronger performance for concreteness and imageability than for familiarity and iconicity. Model-human agreement is also higher for items with more consistent human ratings. These results suggest that LLMs can provide useful psycholinguistic estimates, but should not be treated as direct replacements for human judgments, particularly for context-dependent meanings. Systemic differences with human ratings observed in our error analysis suggest that validation for specific semantic domains (e.g. psychological states) is needed before using LLM-generated norms for downstream tasks.

RQ3: How do lexical and contextual psycholinguistic ratings contribute to modelling metaphoric potential?

In Study 3 (Chapter 6), the contextual model performs better than the lexical model, while the combined model achieves the best overall and lemma-disjoint performance. Lexical and contextual ratings therefore provide complementary information. Lexical ratings capture relatively stable properties of a word and its basic meaning, whereas contextual ratings reflect how these properties are realised in a particular sentence. Concreteness shows this distinction most clearly: higher lexical concreteness increases the likelihood of metaphorical usage, while higher contextual concreteness decreases it.

Beyond LLM evaluation, our dataset can support further research in both computational linguistics and psycholinguistics. It can be used to examine how psycholinguistic properties vary across word senses and literal and metaphorical usages, and to evaluate whether computational models capture these contextual differences. It may also provide stimuli and human reference ratings for experiments

on lexical processing and semantic interpretation. Future work could extend the dataset to more lexical items and contexts, to other languages, and collect more human ratings to improve its reliability.

Bibliography

[AD92] Peter Auer and Aldo Di Luzio, editors. *The Contextualization of Language*. Volume 22. Pragmatics & Beyond New Series. Amsterdam: John Benjamins, 1992. https://doi.org/10.1075/pbns.22 (cited on page 9).

[APC20] Yara Alnafesah, Mohammad Taher Pilehvar, and Nigel Collier. "Automatic Metaphor Detection Using Contextual Word Embeddings and Concreteness Scores". In: *Proceedings of the Second Workshop on Figurative Language Processing*. 2020, pages 154–158 (cited on page 12).

[AST24] Kotaro Aono, Ryohei Sasano, and Koichi Takeda. "Verifying Claims About Metaphors with Large-Scale Automatic Metaphor Identification". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Edited by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pages 711–719. https://doi.org/10.18653/v1/2024.naacl-short.62. https://aclanthology.org/2024.naacl-short.62/ (cited on pages 12–14, 49).

[Azu96] Tamiko Azuma. "Familiarity and related lexical properties of homographic meanings". In: *Behavior Research Methods, Instruments, & Computers* 28.1 (1996), pages 109–115. https://doi.org/10.3758/BF03203645 (cited on page 14).

[BC93] Dawn G. Blasko and Cynthia M. Connine. "Effects of Familiarity and Aptness on Metaphor Processing". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19.2 (1993), pages 295–308. https://doi.org/10.1037/0278-7393.19.2.295 (cited on page 14).

[Bla62] Max Black. *Models and Metaphors: Studies in Language and Philosophy*. Ithaca, NY: Cornell University Press, 1962 (cited on page 9).

[BLF15] Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. "Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples". In: *Proceedings of the Third Workshop on Metaphor in NLP*. Edited by Ekaterina Shutova, Beata Beigman Klebanov, and Patricia Lichtenstein. Denver, Colorado: Association for Computational Linguistics, June 2015, pages 11–20. https://doi.org/10.3115/v1/W15-1402. https://aclanthology.org/W15-1402/ (cited on page 12).

[BMR25] Marc Brysbaert, Gorka Martínez, and Pedro Reviriego. "Moving Beyond Word Frequency Based on Tally Counting: AI-Generated Familiarity Estimates of Words and Phrases as an Additional Index of Language Knowledge". In: *Behavior Research Methods* 57 (2025), page 28 (cited on pages 5, 16, 22).

[Bos85] Peter Bosch. "Context Dependence and Metaphor". In: *The Ubiquity of Metaphor: Metaphor in Language and Thought*. Edited by Wolf Paprotté and René Dirven. Volume 29. Current Issues in Linguistic Theory. Amsterdam: John Benjamins, 1985, pages 141–176. https://doi.org/10.1075/cilt.29.08bos (cited on page 9).

[BWK14] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior Research Methods* 46.3 (2014), pages 904–911. ISSN: 1554-3528. https://doi.org/10.3758/s13428-013-0403-5. https://doi.org/10.3758/s13428-013-0403-5 (cited on pages 4, 10–12, 15, 19, 26, 33).

[BZ21] Johan Blomberg and Jordan Zlatev. "Metalinguistic Relativity: Does One's Ontology Determine One's View on Linguistic Relativity?" In: *Language & Communication* 76 (2021), pages 35–46. ISSN: 0271-5309. https://doi.org/10.1016/j.langcom.2020.09.007 (cited on page 7).

[Cam03] Lynne Cameron. *Metaphor in Educational Discourse*. London: Continuum, 2003 (cited on pages 15, 50).

[Cam10] Lynne Cameron. "The Discourse Dynamics Framework for Metaphor". In: *Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities*. Edited by Lynne Cameron and Robert Maslen. Studies in Applied Linguistics. London: Equinox, 2010, pages 77–94. ISBN: 978-1-84553-446-2 (cited on page 7).

[Cam18] Lynne Cameron. "From Metaphor to Metaphorizing: How Cinematic Metaphor Opens Up Metaphor Studies". In: *Cinematic Metaphor in Perspective: Reflections on a Transdisciplinary Framework*. Edited by Sarah Greifenstein, Dorothea Horst, Thomas Scherer, Christina Schmitt, Hermann Kappelhoff, and Cornelia Müller. Volume 5. Cinepoetics. Berlin and Boston: De Gruyter, 2018, pages 17–35. ISBN: 978-3-11-061223-3. https://doi.org/10.1515/9783110615036 (cited on pages 7, 48).

[Cam97] Lynne Joan Cameron. "Metaphorical Use of Language in Educational Discourse: A Theoretical and Empirical Investigation". PhD thesis. London, UK: Institute of Education, University of London, 1997. https://discovery.ucl.ac.uk/id/eprint/10020284/ (cited on pages 7, 48).

[CKH10] Michael J. Cortese, Maya M. Khanna, and Sarah Hacker. "Recognition memory for 2,578 monosyllabic words". In: *Memory* 18.6 (2010). PMID: 20677075, pages 595–609. https://doi.org/10.1080/09658211.2010.493892. https://doi.org/10.1080/09658211.2010.493892 (cited on page 13).

[Col81] Max Coltheart. "The mrc psycholinguistic database". English. In: *The Quarterly Journal of Experimental Psychology Section A* 33.4 (1981), pages 497–505. ISSN: 0272-4987. https://doi.org/10.1080/14640748108400805 (cited on pages 4, 10, 26).

[Con+25a] Javier Conde, Miguel González, María Grandury, Gonzalo Martínez, Pedro Reviriego, and Mar Brysbaert. *Psycholinguistic Word Features: A New Approach for the Evaluation of LLMs Alignment with Humans*. Preprint. 2025 (cited on pages 16, 50).

[Con+25b] Javier Conde, María Grandury, Tairan Fu, Carlos Arriaga, Gonzalo Martínez, Thomas Clark, Sean Trott, Clarence Gerald Green, Pedro Reviriego, and Marc Brysbaert. "Adding LLMs to the psycholinguistic norming toolbox: A practical guide to getting the most out of human ratings". In: *arXiv preprint arXiv:2509.14405* (2025) (cited on page 5).

[Dön+25] Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Faleńska. "AI Argues Differently: Distinct Argumentative and Linguistic Patterns of LLMs in Persuasive Contexts". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2025, pages 34595–34626 (cited on page 12).

[DRM63] James L. Dolby, Howard L. Resnikoff, and Fred L. MacMurray. "A Tape Dictionary for Linguistic Experiments". In: *Proceedings of the Fall Joint Computer Conference*. Volume 24. Baltimore, MD: Spartan Books, 1963, pages 419–423. https://doi.org/10.1145/1463822.1463865 (cited on page 10).

[DWG18] Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. "Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pages 1412–1424. https://doi.org/10.18653/v1/D18-1171 (cited on page 12).

[DZ20] Simon Devylder and Jordan Zlatev. "Cutting and Breaking Metaphors of the Self and the Motivation & Sedimentation Model". In: *Figurative Meaning Construction in Thought and Language*. Edited by Annalisa Baicchi and Günter Radden. Volume 9. Figurative Thought and Language. Amsterdam: John Benjamins Publishing Company, 2020, pages 253–282. ISBN: 978-90-272-0705-0. https://doi.org/10.1075/ftl.9.11dev (cited on page 7).

[GIB06] RAYMOND W. GIBBS JR. "Metaphor Interpretation as Embodied Simulation". In: *Mind & Language* 21.3 (2006), pages 434–458. https://doi.org/https://doi.org/10.1111/j.1468-0017.2006.00285.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0017.2006.00285.x (cited on pages 13, 48, 49).

[Gio97] Rachel Giora. "Understanding Figurative and Literal Language: The Graded Salience Hypothesis". In: *Cognitive Linguistics* 8.3 (1997), pages 183–206. https://doi.org/10.1515/cogl.1997.8.3.183 (cited on page 9).

[GL80a] K. J. Gilhooly and R. H. Logie. "Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words". In: *Behavior Research Methods & Instrumentation* 12.4 (1980), pages 395–427. ISSN: 1554-3528. https://doi.org/10.3758/BF03201693. https://doi.org/10.3758/BF03201693 (cited on page 19).

[GL80b] Ken J. Gilhooly and Robert H. Logie. "Age-of-Acquisition, Imagery, Concreteness, Familiarity, and Ambiguity Measures for 1,944 Words". In: *Behavior Research Methods & Instrumentation* 12.4 (1980), pages 395–427. https://doi.org/10.3758/BF03201693 (cited on page 10).

[Goa97] Andrew Goatly. *The Language of Metaphors*. London and New York: Routledge, 1997. ISBN: 9780415128773. https://doi.org/10.4324/9780203210000 (cited on pages 14, 49).

[GRB14] Andrew Gargett, Josef Ruppenhofer, and John Barnden. “Dimensions of metaphorical meaning”. In: *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (COGALEX)*. 2014, pages 166–173 (cited on pages 12, 13).

[Gri75] H. P. Grice. “Logic and Conversation”. In: *Syntax and Semantics, Volume 3: Speech Acts*. Edited by Peter Cole and Jerry L. Morgan. New York: Academic Press, 1975, pages 41–58 (cited on page 9).

[GV16] Gianluca Genovesi and Michael Vertolli. “The Independent Role of Familiarity in Metaphor Processing: An Eye-Tracking Study”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. 2016 (cited on page 14).

[HM26] Hiromichi Hagihara and Kazuki Miyazawa. “How well do large language models mirror human cognition of word concepts?: A comparison of psychological ratings for early-acquired English words”. In: *Behavior Research Methods* 58.2 (2026), page 58. ISSN: 1554-3528. https://doi.org/10.3758/s13428-025-02938-2. https://doi.org/10.3758/s13428-025-02938-2 (cited on pages 16, 47).

[JJ23] Nina Julich-Warpakowski and Thomas Jensen. “Zooming in on the notion of metaphoricity: Notions, dimensions, and operationalizations”. In: *Metaphor and the Social World* 13 (July 2023), pages 16–36. https://doi.org/10.1075/msw.00027.jul (cited on pages 9, 13–15, 48–50).

[Kis+73] George R. Kiss, Christine Armstrong, Robert Milroy, and James Piper. “An Associative Thesaurus of English and Its Computer Analysis”. In: *The Computer and Literary Studies*. Edited by A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith. Edinburgh: Edinburgh University Press, 1973, pages 153–165 (cited on page 10).

[Kre11] Tina Krennmayr. *Metaphor in Newspapers*. Utrecht: LOT, 2011 (cited on pages 15, 50).

[KS17] Tina Krennmayr and Gerard Steen. “VU Amsterdam Metaphor Corpus”. In: *Handbook of Linguistic Annotation*. Edited by Nancy Ide and James Pustejovsky. Dordrecht: Springer Netherlands, 2017, pages 1053–1071. ISBN: 978-94-024-0881-2. https://doi.org/10.1007/978-94-024-0881-2_39. https://doi.org/10.1007/978-94-024-0881-2_39 (cited on page 8).

[KWC18] Douwe Kiela, Changhan Wang, and Kyunghyun Cho. “Dynamic Meta-Embeddings for Improved Sentence Representations”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pages 1466–1477. https://doi.org/10.18653/v1/D18-1176 (cited on page 12).

[LBS18a] Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. “A Report on the 2018 VUA Metaphor Detection Shared Task”. In: *Proceedings of the Workshop on Figurative Language Processing*. 2018, pages 56–66 (cited on page 15).

[LBS18b] Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. “A Report on the 2018 VUA Metaphor Detection Shared Task”. In: *Proceedings of the Workshop on Figurative Language Processing*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pages 56–66. https://doi.org/10.18653/v1/W18-0907 (cited on page 8).

[Leo+20a] Chee Wee Leong et al. “A Report on the 2020 Shared Task on Metaphor Detection”. In: *Proceedings of the Second Workshop on Figurative Language Processing*. 2020 (cited on page 15).

[Leo+20b] 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: *Proceedings of the Second Workshop on Figurative Language Processing*. Online: Association for Computational Linguistics, 2020, pages 18–29. https://doi.org/10.18653/v1/2020.figlang-1.3 (cited on page 8).

[Lit15] Jeannette Littlemore. “Metaphor in Specialist Discourse: Insights and Implications for Metaphor Studies and Beyond”. In: *Metaphor in Specialist Discourse*. Edited by J. Berenike Herrmann and Tony Berber Sardinha. John Benjamins, 2015, pages 299–314. https://doi.org/10.1075/milcc.4.12lit (cited on page 9).

[LJ80] George Lakoff and Mark Johnson. *Metaphors We Live By*. Chicago: University of Chicago Press, 1980 (cited on pages 4, 8, 12, 13, 15, 34, 48, 49).

[LW18] Gary Lupyan and Bodo Winter. “Language Is More Iconic Than You Think”. In: *Proceedings of the Royal Society B: Biological Sciences* 285.1886 (2018), page 20180696. https://doi.org/10.1098/rspb.2018.0696 (cited on page 13).

[Mar+25] Anna Marklová, Jiří Milicka, Leonid Ryvkin, L’udmila Lacková Bennet, and Libuše Kormaníková. “Iconicity in Large Language Models”. In: *Digital Scholarship in the Humanities* 40.4 (2025), pages 1203–1224 (cited on pages 5, 16, 32).

[Mit86] Roger Mitton. “A Partial Dictionary of English in Computer-Usable Form”. In: *Literary and Linguistic Computing* 1.4 (1986), pages 214–215. https://doi.org/10.1093/llc/1.4.214 (cited on page 10).

[MMG+25] Gorka Martínez, José D. Molero, Sergio González, et al. “Using Large Language Models to Estimate Features of Multi-Word Expressions: Concreteness, Valence, Arousal”. In: *Behavior Research Methods* 57 (2025), page 5 (cited on pages 5, 16, 34).

[Mon+23] Maria Montefinese, Lorenzo Gregori, Andrea Amelio Ravelli, Rossella Varvara, and Daniele Paolo Radicioni. “CONcreTEXT norms: Concreteness ratings for Italian and English words in context”. In: *PLOS ONE* 18.10 (2023), e0293031. https://doi.org/10.1371/journal.pone.0293031 (cited on pages 4, 11, 13, 16, 32, 34).

[MP17] Dominic W. Massaro and Marcus Perlman. “Quantifying Iconicity’s Contribution during Language Acquisition: Implications for Vocabulary Learning”. In: *Frontiers in Communication* Volume 2 - 2017 (2017). ISSN: 2297-900X. https://doi.org/10.3389/fcomm.2017.00004. https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2017.00004 (cited on page 13).

[Mül08] Cornelia Müller. *Metaphors Dead and Alive, Sleeping and Waking: A Dynamic View*. Chicago and London: University of Chicago Press, 2008. ISBN: 978-0-226-54825-8 (cited on pages 7, 48).

[PA26] Laura Pissani and Roberto G. de Almeida. “Metaphors in context and in isolation: Familiarity, aptness, concreteness, metaphoricity, and structure norms for 300 two-word expressions”. In: *Behavior Research Methods* 58 (2026), page 31. https://doi.org/10.3758/s13428-025-02902-0 (cited on pages 11, 16, 32, 47).

[Pai13] Allan Paivio. “Dual Coding Theory, Word Abstractness, and Emotion: A Critical Review of Kousta et al. (2011)”. In: *Journal of Experimental Psychology: General* 142.1 (2013), pages 282–287. https://doi.org/10.1037/a0027004 (cited on page 12).

[Per+18] Lynn K. Perry, Marcus Perlman, Bodo Winter, Dominic W. Massaro, and Gary Lupyan. “Iconicity in the speech of children and adults”. In: *Developmental Science* 21.3 (2018), e12572. https://doi.org/https://doi.org/10.1111/desc.12572. https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12572 (cited on page 13).

[PYM68] Allan Paivio, John C. Yuille, and Stephen A. Madigan. “Concreteness, imagery, and meaningfulness values for 925 nouns”. In: *Journal of Experimental Psychology* 76.1, Pt. 2 (1968), pages 1–25. https://doi.org/10.1037/h0025327. https://doi.org/10.1037/h0025327 (cited on pages 10, 19, 33).

[Rab+18] Ella Rabinovich, Benjamin Sznajder, Artem Spector, Ilya Shnayderman, Ranit Aharonov, David Konopnicki, and Noam Slonim. “Learning Concept Abstractness Using Weak Supervision”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pages 4854–4859. https://doi.org/10.18653/v1/D18-1522 (cited on page 12).

[Rei+19] W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. “How Polysemy Affects Concreteness Ratings: The Case of Metaphor”. In: *Cognitive Science* 43.2 (2019), e12703. https://doi.org/10.1111/cogs.12703 (cited on pages 13, 47).

[RF86] David C. Rubin and Michael Friendly. “Predicting Which Words Get Recalled: Measures of Free Recall, Availability, Goodness, Emotionality, and Pronunciability for 925 Nouns”. In: *Memory & Cognition* 14.1 (1986), pages 79–94. https://doi.org/10.3758/BF03209231 (cited on page 13).

[Ric36] I. A. Richards. *The Philosophy of Rhetoric*. New York: Oxford University Press, 1936 (cited on page 9).

[RS25] Sebastian Reimann and Tatjana Scheffler. “Using Large Language Models to Perform MIPVU-Inspired Automatic Metaphor Detection”. In: *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*. Vienna, Austria: Association for Computational Linguistics, 2025, pages 10–21. https://doi.org/10.18653/v1/2025.analogyangle-1.2 (cited on page 8).

[SA22] Elisa Sanchez-Bayona and Rodrigo Agerri. “Leveraging a New Spanish Corpus for Multilingual and Crosslingual Metaphor Detection”. In: *Proceedings of the 26th Conference on Computational Natural Language Learning*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pages 228–240. https://doi.org/10.18653/v1/2022.conll-1.16 (cited on page 8).

[SCK12] Jolanda Schock, Michael J. Cortese, and M. Maya Khanna. "Imageability Estimates for 3,000 Monosyllabic Words". In: *Behavior Research Methods* 44.2 (2012), pages 374–379. https://doi.org/10.3758/s13428-011-0162-9 (cited on pages 15, 49).

[Sco+19] Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. "The Glasgow Norms: Ratings of 5,500 Words on Nine Scales". English. In: *Behavior Research Methods* 51.3 (2019), pages 1258–1270. ISSN: 1554-3528. https://doi.org/10.3758/s13428-018-1099-3 (cited on pages 4, 10, 11).

[SD06] Hans Stadthagen-Gonzalez and Colin J. Davis. "The Bristol norms for age of acquisition, imageability, and familiarity". In: *Behavior Research Methods* 38.4 (2006), pages 598–605. ISSN: 1554-351X. https://doi.org/10.3758/BF03193891 (cited on pages 13–15, 19).

[SHB26] Xander Akiko Snelder, Yunchong Huang, and Jelke Bloem. "Prompting Instruction-tuned LLMs for Semantic Similarity Values". In: *Proceedings of the Fifteenth Language Resources and Evaluation Conference*. Edited by Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral. Palma de Mallorca, Spain: ELRA Language Resource Association, May 2026, pages 11390–11403. https://doi.org/10.63317/3kbjxx6989dg. https://aclanthology.org/2026.lrec-1.891/ (cited on page 16).

[SS89] Paula J. Schwanenflugel and Randall W. Stowe. "Context Availability and the Processing of Abstract and Concrete Words in Sentences". In: *Reading Research Quarterly* 24.1 (1989), pages 114–126. ISSN: 00340553. http://www.jstor.org/stable/748013 (visited on 08/05/2026) (cited on page 13).

[Ste+10] Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam: John Benjamins, 2010 (cited on pages 5, 8, 12, 14, 48, 49).

[SW86] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press, 1986 (cited on page 9).

[TB22] Sean Trott and Benjamin Bergen. *Contextualized Sensorimotor Norms: multi-dimensional measures of sensorimotor strength for ambiguous English words, in context*. 2022. https://arxiv.org/abs/2203.05648 (cited on page 11).

[TB78] Michael P. Toglia and William F. Battig. *Handbook of Semantic Word Norms*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1978 (cited on pages 10, 13).

[TÖS15] Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. "Exploring Sensorial Features for Metaphor Identification". In: *Proceedings of the Third Workshop on Metaphor in NLP*. Edited by Ekaterina Shutova, Beata Beigman Klebanov, and Patricia Lichtenstein. Denver, Colorado: Association for Computational Linguistics, June 2015, pages 31–39. https://doi.org/10.3115/v1/W15-1404. https://aclanthology.org/W15-1404/ (cited on page 13).

[Tro24] Sean Trott. "Can large language models help augment English psycholinguistic datasets?" In: *Behavior Research Methods* 56 (2024), pages 6082–6100. https://doi.org/10.3758/s13428-024-02337-z (cited on page 16).

[Tsv+14] 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)*. Edited by Kristina Toutanova and Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pages 248–258. https://doi.org/10.3115/v1/P14-1024. https://aclanthology.org/P14-1024/ (cited on page 13).

[TT11] Kumiko Tanaka-Ishii and Hiroshi Terada. "Word Familiarity and Frequency". In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. 2011, pages 1–10 (cited on pages 15, 49).

[WB25] Si Wu and Sebastian Bruch. "Uncovering Visual-Semantic Psycholinguistic Properties from the Distributional Structure of Text Embedding Space". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, 2025, pages 5638–5649. https://doi.org/10.18653/v1/2025.acl-long.281 (cited on page 12).

[Win+24] Bodo Winter, Gary Lupyan, Lynn K. Perry, Mark Dingemanse, and Marcus Perlman. "Iconicity ratings for 14,000+ English words". In: *Behavior Research Methods* 56.3 (2024), pages 1640–1655. ISSN: 1554-3528. https://doi.org/10.3758/s13428-023-02112-6. https://doi.org/10.3758/s13428-023-02112-6 (cited on pages 4, 10, 11, 13, 19, 26).

[WX20] Mingyu Wan and Baixi Xing. "Modality Enriched Neural Network for Metaphor Detection". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Edited by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pages 3036–3042. https://doi.org/10.18653/v1/2020.coling-main.270. https://aclanthology.org/2020.coling-main.270/ (cited on page 14).

[Zho+25] Yanlu Zhong, Simon Todd, Nicole Xu, and Laurel Brehm. "Evaluating LLMs as proxies for humans in psycholinguistic ratings: A comparison of statistical knowledge". In: *Research Methods in Applied Linguistics* 4.3 (2025), page 100274. https://doi.org/10.1016/j.rmal.2025.100274 (cited on page 47).

Appendix A

Annotation Instructions

A.1 Concreteness

Concreteness Rating Instructions

In this survey we will ask you to evaluate the degree of concreteness of a series of word senses. The aim of our study is to understand whether and to what extent the degree of concreteness of words depends on the context in which words are presented.
Your task is to assess the concreteness of the concept expressed by a given word within a sentence, on a scale from 1 (completely abstract) to 7 (completely concrete).
Some words refer to things or actions in reality, which you can experience directly through one of the five senses. We call these words concrete words. Other words refer to meanings that cannot be experienced directly but which we know because the meanings can be defined by other words. These are abstract words. Still other words fall in-between the two extremes, because we can experience them to some extent and in addition we rely on language to understand them. We want you to indicate how concrete the meaning of each word is for you by using a 7-point rating scale going from abstract to concrete.
A concrete word comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your five senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it (e.g. To explain 'sweet' you could have someone eat sugar; To explain 'jump' you could simply jump up and down or show people a movie clip about someone jumping up and down; To explain 'couch', you could point to a couch or show a picture of a couch).
An abstract word comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words (e.g. There is no simple way to demonstrate 'justice'; but we can explain the meaning of the word by using other words that capture parts of its meaning).
Importantly, the concreteness of a word may change depending on the context in which it is used. For example, in the sentence "In summer, wheat FIELDS are colored in yellow", the noun FIELD refers to an extremely concrete concept because it can be seen, touched, and pointed to. In this case, in a scale ranging from 1 to 7 its concreteness may be evaluated as 7, because it refers to an extremely concrete concept. In contrast, in the sentence "Magnetic FIELD attracts iron", the noun FIELD refers to something more abstract compared to "wheat FIELDS" and may receive a lower rating.
There are no correct or wrong answers. Please, when you read the word in the sentence, do not spend a lot of time thinking about your rating, because your first considerations are of greatest interest. It is important that you try to use the entire range of the scale (from 1 to 7).

A.2 Iconicity

Iconicity Rating Instructions

In this survey, you will evaluate the degree of iconicity of words as they are used in sentences. The aim of this study is to understand whether and to what extent the iconicity of a word depends on the context in which it appears.

Your task is to assess the iconicity of the concept expressed by a highlighted word within a sentence, on a scale from 1 (not iconic at all) to 7 (highly iconic).

Some English words sound like what they mean. These words are iconic. You might be able to guess the meaning of such a word even if you did not know English.

Words with high iconicity not only mimic sounds (e.g., "screech"), but also sensory qualities (e.g., manner of movement as in "twirl"), touches (e.g., "mushy"), sizes (e.g., "teeny"), textures (e.g., "crisp") or visual events (e.g., "flash").

Some words that people have rated moderate in iconicity are "porcupine," "glowing," and "steep," because they sound somewhat like what they mean.

Some words rated low in iconicity are "menu," "amateur," and "are," because they do not sound at all like what they mean. In this task, you are going to rate words for how iconic they are.

Sometimes, the same word may have different degrees of iconicity in different contexts. For example, in the sentence "The only sound was the soft TICK of the clock", the sound of the word TICK imitates the real-world sound of a clock, and may receive a higher rating. In contrast, in the sentence "His exercise books were full of well-deserved red TICKS", TICK refers to a written check mark indicating a correct answer, which is no longer tied to or resembles its sound, and it may therefore receive a lower iconicity rating.

You will rate each word on a scale from 1 to 7. A rating of 1 indicates that the word is not at all iconic and does not at all sound like what it means. 7 indicates that the word is high in iconicity and sounds very much like what it means.

There are no correct or wrong answers. Please do not spend too much time thinking about your ratings. Your first intuitions are of greatest interest for this study. It is important that you try to use the entire range of the scale (from 1 to 7).

A.3 Imageability

Imageability Rating Instructions

In this survey, you will evaluate the degree of imageability of words as they are used in sentences. The aim of this study is to understand whether and to what extent the imageability of a word depends on the context in which it appears.

Your task is to assess how easily the concept expressed by a highlighted word within a sentence can evoke a mental image, on a scale from 1 (very difficult to imagine) to 7 (very easy to imagine).

Words differ in their capacity to arouse mental images of things and events. Some words arouse a sensory experience, such as a mental picture or sound, very quickly and easily, whereas others may do so only with difficulty (i.e., after a long delay) or not at all. The purpose of this experiment is to rate a list of words as to the ease or difficulty with which they arouse mental images. Any word which, in your estimation, arouses a mental image (i.e., a mental picture, or sound, or other sensory experience) very quickly and easily should be given a high imagery rating; any word that arouses a mental image with difficulty or not at all should be given a low imagery rating.

Think of the words "apple" or "fact." "Apple" would probably arouse an image relatively easily and would be rated as high imagery; "fact" would probably do so with difficulty and would be rated as low imagery. Since words tend to make you think of other words as associates, e.g., knife–fork, it is important that you note only the ease of getting a mental image of an object or an event to the word.

Importantly, the imageability of a word may change depending on the context in which it is used. Please evaluate the highlighted word only in the context of the sentence you are reading. For example, in the sentence "Two men murmured with their heads together at the end of the BAR", the word BAR refers to a place where alcoholic and other drinks are sold and drunk, and is easy to imagine. In this context, the word may receive a relatively high imageability rating. In contrast, in the sentence "You need to pass the BAR if you really want to practice law", the word BAR refers to the law licensing exam and may receive a lower imageability rating.

Your ratings will be made on a seven-point scale, where one is the low imagery end of the scale and seven is the high imagery end of the scale. Make your rating by typing a number from 1 to 7 that best indicates your judgment of the ease or difficulty with which the word arouses imagery. The words that arouse mental images most readily for you should be given a rating of 7; words that arouse images with the greatest difficulty or not at all should be rated 1; words that are intermediate in ease or difficulty of imagery, of course, should be rated appropriately between the two extremes.

It is important that you try to use the entire range of the scale (from 1 to 7). At the same time, don't be concerned about how often you use a particular number as long as it is your true judgment. Work fairly quickly but do not be careless in your ratings.

A.4 Familiarity

Familiarity Rating Instructions

In this survey, you will evaluate the degree of familiarity of words as they are used in sentences. The aim of this study is to understand whether and to what extent the familiarity of a word depends on the context in which it appears.

Your task is to assess how familiar the concept expressed by a highlighted word within a sentence feels to you, on a scale from 1 to 7. In this scale, 1 represents "NEVER," that is, you have never seen or heard or used the word in your life; the number 2 represents "RARELY," that is you have seen or heard or used the word at least once before, but only rarely; and so on until 7, which represents "VERY OFTEN," that is, you have seen or heard or used the word nearly every day of your life.

Do not be bothered if you are unable to give a definition of some of the words. Simply rate each one as to the number of times you have come in contact with it regardless of its meaning.

There may be some words which you have used or heard more often than you have seen them. Or there may be other words which you have seen more often than you have used or heard them. In such cases, always give the word in the highest rating of the three. For example, you probably use or hear the word "cheers" often, but you may never have seen it in print. In this case, you would rate "cheers" as "OFTEN" and write down the number 6.

It is important that sometimes, the same word may have different degrees of familiarity depending on the context in which it is used. For example, in the two sentences "After dinner she ate a MINT" and "The government expanded the MINT", the word "MINT" may feel somewhat less familiar to some participants in the second sentence because this meaning may be used less frequently in everyday life. In this case, the word in the second sentence may receive a lower familiarity rating.

It is important that you try to use the entire range of the scale (from 1 to 7). At the same time, don't be concerned about how often you use a particular number as long as it is your true judgment. Work fairly quickly but do not be careless in your ratings.

Appendix B

Model rating prompts

B.1 Concreteness

Concreteness isolated rating prompt 1

Complete the following task as a native speaker of English.
Please indicate how concrete the TARGET word is on a scale from 1 (not at all concrete) to 7 (highly concrete).
Target word: {word}
Only answer with an integer from 1 to 7.

Concreteness isolated rating prompt 2

Could you please rate the concreteness of the following word on a scale from 1 to 7, where 1 means very abstract and 7 means very concrete? Examples of words that would get a rating of 1 are essentialness, although and hope. Examples of words that would get a rating of 7 are bat, frangipane, and blackbird. The word is: {word}. Only answer a number from 1 to 7. Please limit your answer to numbers.

Concreteness in-context rating prompt 1

Could you please rate the concreteness of the word in the sentence on a scale from 1 to 7, where:
1 = very abstract 7 = very concrete
Examples of words in context that would get a rating of 1 are:
field: Magnetic field attracts iron.
run: The student runs the risk of failing the exam.
ambivalent: I felt very ambivalent about leaving home.
Examples of words in context that would get a rating of 7 are:
field: In summer, wheat fields are coloured in yellow.
run: John runs the marathon.
wooden: Stir the mixture with a wooden spoon.
Rate the word: {word} in the sentence: {sentence}.
Only answer with a number from 1 to 7. It is important that you try to use the entire range of the scale (from 1 to 7). Please limit your answer to numbers.

Concreteness in-context rating prompt 2

Could you please rate the concreteness of the word in the sentence on a scale from 1 to 7, where:
1 = very abstract 7 = very concrete
A concrete word refers to things or actions in reality. For example, in the sentence "In summer, wheat FIELDS are colored in yellow", the noun FIELD may be evaluated as 7. In contrast, in the sentence "Magnetic FIELD attracts iron", the noun FIELD may receive a lower rating. Now rate the word: {word} in the sentence: {sentence}.
Only answer with a number from 1 to 7. It is important that you try to use the entire range of the scale (from 1 to 7). Please limit your answer to numbers.

B.2 Iconicity

Iconicity isolated rating prompt

Could you please rate the iconicity of the following word on a scale from 1 to 7, where 1 means the sound of the word does not resemble its meaning at all and 7 means the sound strongly resembles its meaning?
Examples of words that would get a rating of 1 are menu, amateur and are.
Examples of words that would get a rating of 7 are splash, twirl and mushy.
The word is: {word}.
Only answer a number from 1 to 7.

Iconicity in-context rating prompt

Could you please rate the iconicity of the word in the sentence on a scale from 1 to 7, where:
1 = not iconic at all 7 = extremely iconic
Words with high iconicity have forms that resemble their meanings. Please evaluate the word based on the context of the sentence you are reading. For example, in the sentence "The only sound was the soft TICK of the clock", the word TICK may be evaluated very high. In contrast, in the sentence "His exercise books were full of well-deserved red TICKS", the word TICK may receive a lower rating. Now rate the word: {word} in the sentence: {sentence}.
Only answer with a number from 1 to 7. It is important that you try to use the entire range of the scale (from 1 to 7). Please limit your answer to numbers.

B.3 Imageability

Imageability isolated rating prompt

Could you please rate the imageability of the following word on a scale from 1 to 7, where 1 means very difficult to form a mental image and 7 means very easy to form a mental image?
Examples of words that would get a rating of 1 are fact, circumstance and able.
Examples of words that would get a rating of 7 are apple, pizza and mountain.
The word is: {word}.
Only answer a number from 1 to 7. Please limit your answer to numbers.

Imageability in-context rating prompt

Could you please rate the imageability of the word in the sentence on a scale from 1 to 7, where:
1 = not imageable at all 7 = extremely imageable
A imageable word easily arouses images of things and events. Please evaluate the word based on the context of the sentence you are reading. For example, in the sentence "Two men murmured with their heads together at the end of the BAR", the noun BAR may be evaluated very high. In contrast, in the sentence "You need to pass the BAR if you really want to practice law", the noun BAR may receive a lower rating. Now rate the word: {word} in the sentence: {sentence}.
Only answer with a number from 1 to 7. It is important that you try to use the entire range of the scale (from 1 to 7). Please limit your answer to numbers.

B.4 Familiarity

Familiarity isolated rating prompt

Could you please rate the familiarity of the following word on a scale from 1 to 7, where 1 means very unfamiliar and 7 means very familiar?
Examples of words that would get a rating of 1 are astrolabe, bedel and convocation.
Examples of words that would get a rating of 7 are house, water and dog.
The word is: {word}.
Only answer a number from 1 to 7. Please limit your answer to numbers.

Familiarity in-context rating prompt 2

Could you please rate the familiarity of the word in the sentence on a scale from 1 to 7, where:
1 = not familiar at all 7 = extremely familiar
Please evaluate the word based on the context of the sentence you are reading. For example, in the sentence "After dinner she ate a MINT", the word MINT may be evaluated very high. In contrast, in the sentence "The government expanded the MINT", the word MINT may receive a lower rating. Now rate the word: {word} in the sentence: {sentence}.
Only answer with a number from 1 to 7. It is important that you try to use the entire range of the scale (from 1 to 7). Please limit your answer to numbers.

Appendix C

Human rating highest disagreement

Table C.1: Items with the largest standard deviations in human in-context concreteness ratings.

Sentence	Target	POS	Type	Mean	SD	*N*
Do n't **SOUND** a lot of weight but	sound	Verb	Met/WIDLII	4.00	2.65	7
Having put the **CAT** among the pigeons, made mischief in his special way, he was lying low now and waiting.	cat	Noun	Met	4.60	2.55	10
THROW us a drink!	throw	Verb	Met	4.11	2.52	9
Yeah you know what Cathy **SPEAK** at the radio	speak	Verb	–	4.71	2.43	7
To **HOLD** all of the 1/50 000 scale topographic map coverage alone for the European Community would probably require about 3 terabytes of storage.	hold	Verb	–	4.57	2.37	7
Winterburn almost put the ball into his own net from Milligan's cross, O'Leary recovering an awkward **SITUATION**.	situation	Noun	Met	4.14	2.34	7
A painter **STOPPED** to kiss her: a journalist waved.	stopped	Verb	–	4.22	2.33	9
Bonar **LAW** then sent for ...	law	Noun	–	4.78	2.33	9
They returned to the bridge, fearful of every **DARK** place— every corner, every door.	dark	Adj.	–	4.22	2.33	9
They climb into the car and **SET** off.	set	Verb	–	3.60	2.32	10

Table C.2: Items with the largest standard deviations in human in-context imageability ratings.

Sentence	Target	POS	Type	Mean	SD	*N*
Well should have **WON**	won	Verb	–	4.44	2.70	9
If he had done the shock would have been less but, on the other **HAND**, his holiday would have been spoilt.	hand	Noun	Met	3.89	2.62	9
Just make **SURE** you squeeze their knickers off, make them	sure	Adj.	Met	3.64	2.62	11
But it's **TRUE**, it does make a noise.	true	Adj.	Met	3.82	2.60	11
Who the **HELL** can?	hell	Noun	–	4.56	2.55	9
No, not more **ACCURATE**: 'thirst' is not accurate at all as a description for this apparently indescribable condition.	accurate	Adj.	Met	4.33	2.55	9
The above equation, in spite of its utmost simplicity, represents an extremely **GOOD** guess.	good	Adj.	Met	4.56	2.51	9
That could work—though it is easy to imagine conflicts of **INTEREST** with the housing and planning functions of the department.	interest	Noun	Met/WIDLII	4.00	2.49	11
He even persuaded 12 European powers to **TURN** cottages into full embassies.	turn	Verb	Met	4.20	2.49	10
This **AREA** of application would appear to be an obvious market for such a product.	area	Noun	Met	3.89	2.47	9

Table C.3: Items with the largest standard deviations in human in-context iconicity ratings.

Sentence	Target	POS	Type	Mean	SD	*N*
LOOK macho.	look	Verb	–	3.50	2.55	10
'He **CAUGHT** a flight to Frankfurt.'	caught	Verb	Met	3.60	2.55	10
Perhaps he will be **ABLE** to say, you know?	able	Adj.	–	3.60	2.46	10
They used the **RING** because of their line into Switzerland.	ring	Noun	–	4.27	2.45	11
Keep your **MOUTH** closed when you've got something in it or we'll take something away.	mouth	Noun	–	3.10	2.42	10
Public relations **JOB** or something similar.	job	Noun	–	3.30	2.41	10
Do it **LOT**'s and lot's.	lot	Noun	Met/WIDLII	3.10	2.38	10
There is a danger in reading back from this figure another stereotype, similar to the Pavlovian 'drug addiction/inner city' assumed **LINK**.	link	Noun	Met	3.50	2.37	10
This is only one vertebra of an animal much too bulky to appear in a book of this **SIZE**!	size	Noun	–	3.60	2.37	10
It's a killing **JOB**.	job	Noun	–	3.40	2.37	10

Table C.4: Items with the largest standard deviations in human in-context familiarity ratings.

Sentence	Target	POS	Type	Mean	SD	*N*
They helped him kit up, Nell connecting the leads from the **BACK** pack into the suit attachments.	back	Noun	–	4.43	2.44	7
But well I mean it's rather **STUPID** for how long?	stupid	Adj.	Met/WIDLII	4.43	2.37	7
Because the ba the children were **SMALL**.	small	Adj.	Met	4.78	2.28	9
'In the air, that risk factor would **KILL** you.'	kill	Verb	Met/PP	5.00	2.24	7
Start thinking, got nothing **LOSE**, sta start	lose	Verb	–	5.00	2.18	9
Not necessarily, no no a **HOUSE** would do.	house	Noun	–	4.56	2.13	9
MIND of its own job, yeah?	mind	Noun	Met/PP	4.44	2.13	9
In private conversation he told Asquith, "I am afraid that I shall have to **SHOW** myself very vicious Mr Asquith this session."	show	Verb	Met/WIDLII	4.33	2.12	9
'I've **LEFT** the suit's radio on full,' he said.	left	Verb	Met	4.78	2.11	9
This desire that cannot find its name, though it would dare **SPEAK** it if it could, is pleasurable.	speak	Verb	Met/PP	4.44	2.07	9

Appendix D

Model rating full results

D.1 Concreteness

D.1.1 Isolated results

Table D.1: Correlation matrix for isolated concreteness ratings. Upper triangle: Spearman's ρ; lower triangle: Pearson's r.

	Human	Llama	Qwen	GPT-4o-mini	GPT-4o
Human	1.000	0.697	0.680	0.726	0.860
Llama	0.664	1.000	0.696	0.726	0.762
Qwen	0.670	0.646	1.000	0.740	0.784
GPT-4o-mini	0.696	0.670	0.635	1.000	0.833
GPT-4o	0.854	0.696	0.679	0.772	1.000

Table D.2: Correlations between human and LLM concreteness ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 40$)	Noun ($N = 40$)	Verb ($N = 38$)
GPT-4o	0.803 / 0.829	0.858 / 0.883	0.823 / 0.803
GPT-4o-mini	0.442 / 0.446	0.759 / 0.798	0.783 / 0.745
Llama	0.419 / 0.413	0.798 / 0.804	0.555 / 0.516
Qwen	0.559 / 0.529	0.646 / 0.702	0.613 / 0.649

D.1.2 In-context results

Table D.3: Pairwise correlations between human and LLM concreteness ratings using probability-weighted outputs. Values above the diagonal are Spearman's ρ, whereas values below the diagonal are Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	0.397	0.428	0.438	0.543
Qwen	0.396	1.000	0.416	0.574	0.577
Llama	0.451	0.463	1.000	0.647	0.625
GPT-4o-mini	0.464	0.526	0.593	1.000	0.770
GPT-4o	0.561	0.531	0.582	0.711	1.000

Table D.4: Correlations between human and LLM concreteness in-context ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 196$)	Noun ($N = 200$)	Verb ($N = 197$)
GPT-4o	0.484 / 0.515	0.663 / 0.663	0.426 / 0.392
GPT-4o-mini	0.370 / 0.359	0.584 / 0.587	0.282 / 0.277
Llama	0.440 / 0.411	0.569 / 0.568	0.246 / 0.236
Qwen	0.366 / 0.393	0.488 / 0.503	0.218 / 0.229

Table D.5: Spearman correlations by human agreement level. Items are divided into tertiles based on the standard deviation of human ratings. High ($n = 198$), Medium ($n = 197$), Low ($n = 198$).

Model	High	Medium	Low
GPT-4o	0.577	0.577	0.470
GPT-4o-mini	0.494	0.473	0.339
Qwen	0.376	0.460	0.340
Llama	0.474	0.464	0.326

Table D.6: GPT-4o's performance on in-context concreteness ratings by metaphor type. Items without a metaphor annotation and those tagged as literal are grouped as Literal. Types with items less than 20 are excluded.

Metaphor type	n	Spearman ρ	p	MAE	RMSE	Bias
Literal	317	0.638	$< .001$	1.915	2.205	-1.466
Metaphorical	225	0.417	$< .001$	2.211	2.455	-1.958
Metaphorical/PP	22	0.286	.197	2.468	2.769	-2.431
Metaphorical/WIDLII	27	0.529	.005	2.197	2.360	-2.142

D.2 Iconicity

D.2.1 Isolated results

Table D.8: Correlation matrix for iconicity ratings. Upper triangle: Spearman's ρ; lower triangle: Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	0.074	0.337	0.472	0.517
Qwen	0.059	1.000	0.128	0.206	0.150
Llama	0.320	0.190	1.000	0.527	0.539
GPT-4o-mini	0.502	0.175	0.532	1.000	0.668
GPT-4o	0.566	0.161	0.547	0.664	1.000

Table D.7: Items with the 10 largest absolute errors in the LLM's in-context concreteness ratings. The target word is shown in bold.

Sentence	Target	POS	Type	Human	SD	LLM	Abs. error
But its future **direction** remains unknown and indeterminate.	direction	Noun	Met	5.71	1.50	2	3.71
"Do not **worry**."	worry	Verb	Met	5.57	1.62	2	3.57
He is perhaps **surprised** by her composure.	surprised	Adj.	Met/PP	6.43	0.79	3	3.43
"**Working** for the Stasi."	working	Verb	Lit	6.29	0.76	3	3.29
Through it, Palestinians are struggling on a daily basis to **shake** off Israel's authority and to construct their own.	shake	Verb	Met	6.14	0.90	3	3.14
No **thanks**.	thanks	Noun	Lit	5.14	1.95	2	3.14
On the other **hand**, sometimes the drama of the auction and the competition between bidders can have the opposite effect.	hand	Noun	Met	5.14	1.77	2	3.14
I searched him and his luggage, and he was **clean**.	clean	Adj.	Met	6.00	1.15	3	3.00
Some parents were very **worried** that more integrated settings would lead to lack of attention for children with special needs.	worried	Adj.	Met	6.00	1.41	3	3.00
Furthermore, the **scales** at which population estimates are often required mean that even EDs are too coarse for risk assessments.	scales	Noun	Met	5.71	1.80	3	2.71

Table D.9: Correlations between human and LLM iconicity ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 40$)	Noun ($N = 40$)	Verb ($N = 40$)
GPT-4o	0.565 / 0.504	0.570 / 0.522	0.660 / 0.593
GPT-4o-mini	0.512 / 0.506	0.497 / 0.395	0.556 / 0.544
Llama	0.499 / 0.528	0.357 / 0.332	0.422 / 0.382
Qwen	0.194 / 0.229	0.049 / 0.090	-0.065 / -0.113

D.2.2 In-context results

Table D.10: Pairwise correlations between human and LLM iconicity ratings using probability-weighted outputs. Values above the diagonal are Spearman's ρ, whereas values below the diagonal are Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	0.179	0.207	0.203	0.349
Qwen	0.194	1.000	0.317	0.430	0.398
Llama	0.200	0.379	1.000	0.358	0.312
GPT-4o-mini	0.210	0.413	0.334	1.000	0.566
GPT-4o	0.353	0.335	0.356	0.552	1.000

Table D.11: Correlations between human and LLM in-context iconicity ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 200$)	Noun ($N = 198$)	Verb ($N = 196$)
GPT-4o	0.352 / 0.345	0.257 / 0.304	0.336 / 0.286
GPT-4o-mini	0.230 / 0.187	0.188 / 0.169	0.170 / 0.211
Llama	0.283 / 0.288	0.275 / 0.279	0.237 / 0.240
Qwen	0.197 / 0.201	0.210 / 0.218	0.178 / 0.115

Table D.12: GPT-4o's performance by metaphor type for iconicity.

Metaphor type	N	ρ	p	MAE	RMSE	Bias
Literal	312	.362	¡.001	2.151	2.305	-2.149
Metaphorical	229	.345	¡.001	2.322	2.478	-2.306
Metaphorical/PP	26	.430	.028	2.172	2.270	-2.172
Metaphorical/WIDLII	26	-.062	.764	2.183	2.362	-2.183

Table D.13: Spearman correlations by human agreement level for iconicity. Items are divided into tertiles based on the standard deviation of human ratings. High ($n = 198$), Medium ($n = 198$), Low ($n = 198$).

Model	High	Medium	Low
GPT-4o	0.395	0.490	0.219
GPT-4o-mini	0.231	0.320	0.080
Qwen	0.229	0.202	0.079
Llama	0.337	0.152	0.148

D.3 Imageability

D.3.1 Isolated results

Table D.14: Correlation matrix for imageability ratings. Upper triangle: Spearman's ρ; lower triangle: Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	0.539	0.625	0.718	0.786
Qwen	0.595	1.000	0.632	0.662	0.634
Llama	0.691	0.707	1.000	0.775	0.706
GPT-4o-mini	0.747	0.630	0.722	1.000	0.773
GPT-4o	0.815	0.542	0.611	0.743	1.000

Table D.15: Correlations between human and LLM imageability ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 40$)	Noun ($N = 40$)	Verb ($N = 40$)
GPT-4o	0.725 / 0.729	0.850 / 0.844	0.831 / 0.760
GPT-4o-mini	0.552 / 0.548	0.771 / 0.749	0.805 / 0.691
Llama	0.415 / 0.374	0.791 / 0.761	0.690 / 0.588
Qwen	0.328 / 0.304	0.634 / 0.703	0.623 / 0.513

D.3.2 In-context results

Table D.16: Pairwise correlations between human and LLM in-context imageability ratings using the model probability-weighted output. Values above the diagonal are Spearman's ρ, whereas values below the diagonal are Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	0.369	0.391	0.466	0.494
Qwen	0.374	1.000	0.572	0.602	0.561
Llama	0.419	0.577	1.000	0.608	0.555
GPT-4o-mini	0.494	0.572	0.573	1.000	0.783
GPT-4o	0.542	0.536	0.539	0.745	1.000

Table D.17: Correlations between human and LLM in-context imageability ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 193$)	Noun ($N = 200$)	Verb ($N = 200$)
GPT-4o	0.491 / 0.471	0.617 / 0.570	0.474 / 0.423
GPT-4o-mini	0.399 / 0.378	0.586 / 0.559	0.460 / 0.439
Llama	0.397 / 0.340	0.407 / 0.406	0.423 / 0.397
Qwen	0.381 / 0.390	0.324 / 0.326	0.392 / 0.363

Table D.18: GPT-4o's performance by metaphor type for imageability.

Metaphor type	N	ρ	p	MAE	RMSE	Bias
Literal	313	.624	¡.001	1.093	1.342	-.660
Metaphorical	218	.170	.012	1.368	1.623	-1.151
Metaphorical/PP	28	.297	.125	1.147	1.407	-.887
Metaphorical/WIDLII	30	.426	.019	1.224	1.373	-.890

Table D.19: Spearman correlations by human agreement level for imageability. Items are divided into tertiles based on the standard deviation of human ratings. High ($n = 198$), Medium ($n = 197$), Low ($n = 198$).

Model	High	Medium	Low
GPT-4o	0.593	0.407	0.208
GPT-4o-mini	0.556	0.367	0.272
Qwen	0.428	0.315	0.203
Llama	0.457	0.413	0.133

D.4 Familiarity

D.4.1 Isolated results

Table D.20: Correlation matrix for familiarity ratings. Upper triangle: Spearman's ρ; lower triangle: Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	-0.065	0.579	0.287	0.230
Qwen	-0.118	1.000	0.161	0.062	0.205
Llama	0.534	0.141	1.000	0.291	0.216
GPT-4o-mini	0.383	0.137	0.291	1.000	0.567
GPT-4o	0.344	0.264	0.177	0.541	1.000

Table D.21: Correlations between human and LLM familiarity ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 40$)	Noun ($N = 40$)	Verb ($N = 40$)
GPT-4o	0.471 / 0.271	nan / nan	0.448 / 0.122
GPT-4o-mini	0.511 / 0.444	0.081 / 0.173	0.533 / 0.358
Llama	0.599 / 0.584	0.392 / 0.368	0.611 / 0.676
Qwen	nan / nan	-0.251 / -0.271	-0.002 / 0.062

D.4.2 In-context results

Table D.22: Pairwise correlations between human and LLM familiarity ratings using the model probability-weighted output. Values above the diagonal are Spearman's ρ, whereas values below the diagonal are Pearson's r.

	Human	Qwen	Llama	GPT-4o-mini	GPT-4o
Human	1.000	0.258	0.263	0.316	0.316
Qwen	0.243	1.000	0.491	0.557	0.445
Llama	0.251	0.472	1.000	0.482	0.406
GPT-4o-mini	0.289	0.571	0.504	1.000	0.719
GPT-4o	0.333	0.521	0.425	0.770	1.000

Table D.23: Correlations between human and LLM in-context familiarity ratings by part of speech. Each cell reports Pearson's r / Spearman's ρ.

Model	Adjective ($N = 200$)	Noun ($N = 199$)	Verb ($N = 198$)
GPT-4o	0.315 / 0.273	0.310 / 0.302	0.415 / 0.423
GPT-4o-mini	0.233 / 0.278	0.271 / 0.304	0.389 / 0.402
Llama	0.270 / 0.290	0.227 / 0.236	0.275 / 0.280
Qwen	0.235 / 0.252	0.174 / 0.184	0.305 / 0.352

Table D.24: GPT-4o's performance by metaphor type for familiarity.

Metaphor type	N	ρ	p	MAE	RMSE	Bias
Literal	328	.251	¡.001	.626	.859	-.030
Metaphorical	208	.284	¡.001	.746	.996	-.274
Metaphorical/PP	20	.456	.043	.627	.741	.107
Metaphorical/WIDLII	38	.483	.002	.742	1.018	-.249

Table D.25: Spearman correlations by human agreement level for familiarity. Items are divided into tertiles based on the standard deviation of human ratings. High ($n = 199$), Medium ($n = 199$), Low ($n = 199$).

Model	High	Medium	Low
GPT-4o	0.218	0.239	0.395
GPT-4o-mini	0.266	0.249	0.306
Qwen	0.248	0.160	0.249
Llama	0.277	0.119	0.242

Appendix E

Item level comparison lists

E.1 Concreteness

Table E.1: Largest and smallest absolute errors between human and GPT-4o isolated concreteness ratings.

Largest 10					Smallest 10				
Lemma	POS	Human iso.	GPT iso.	Abs. error	Lemma	POS	Human iso.	GPT iso.	Abs. error
world	Noun	4.360	2.333	2.027	*bed*	Noun	5.000	5.000	0.000
dark	Adj.	4.290	2.333	1.957	*comfortable*	Adj.	2.330	2.333	0.003
talk	Verb	4.070	2.333	1.737	*elegant*	Adj.	2.310	2.333	0.023
market	Noun	4.700	3.000	1.700	*good*	Adj.	1.640	1.667	0.027
call	Verb	4.000	2.333	1.667	*dull*	Adj.	2.370	2.333	0.037
slim	Adj.	3.930	2.333	1.597	*keep*	Verb	2.370	2.333	0.037
soft	Adj.	3.880	2.333	1.547	*clean*	Adj.	3.070	3.000	0.070
team	Noun	3.790	2.333	1.457	*assume*	Verb	1.750	1.667	0.083
serve	Verb	3.780	2.333	1.447	*save*	Verb	2.420	2.333	0.087
scale	Noun	4.390	3.000	1.390	*throw*	Verb	3.570	3.667	0.097

Table E.2: Largest and smallest absolute shifts between GPT-4o isolated ratings and mean GPT-4o context ratings for concreteness.

Largest 10						Smallest 10					
Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*	Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*
hand	Noun	7.000	4.120	2.880	5.000	*complicated*	Adj.	1.000	1.008	0.008	5.000
bit	Noun	4.000	1.276	2.724	5.000	*sure*	Adj.	1.000	1.015	0.015	4.000
sound	Verb	4.000	1.462	2.538	5.000	*save*	Verb	3.000	2.958	0.042	5.000
light	Noun	5.000	2.631	2.369	5.000	*double*	Adj.	3.000	2.950	0.050	5.000
picture	Noun	6.000	3.656	2.344	5.000	*sick*	Adj.	3.000	3.052	0.052	5.000
link	Noun	4.000	1.690	2.310	5.000	*sit*	Verb	5.000	4.944	0.056	5.000
scale	Noun	4.000	1.693	2.307	5.000	*speak*	Verb	3.000	3.082	0.082	5.000
cat	Noun	7.000	4.962	2.038	5.000	*shake*	Verb	5.000	4.906	0.094	5.000
view	Noun	4.000	1.969	2.031	5.000	*accept*	Verb	2.000	1.899	0.101	5.000
mean	Verb	3.000	1.075	1.925	5.000	*couple*	Noun	4.000	4.118	0.118	5.000

Table E.3: Largest and smallest standard deviations in GPT-4o context concreteness ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
hand	Noun	4.120	3.007	5.000	*worry*	Adj.	1.006	0.005	5.000
cat	Noun	4.962	2.712	5.000	*worry*	Verb	1.005	0.009	5.000
fire	Noun	4.533	2.410	5.000	*complicated*	Adj.	1.008	0.011	5.000
fall	Verb	2.327	2.403	5.000	*confident*	Adj.	1.011	0.013	5.000
slim	Adj.	3.837	2.363	5.000	*sure*	Adj.	1.015	0.016	4.000
office	Noun	4.167	2.350	5.000	*assume*	Verb	1.025	0.035	5.000
twin	Noun	5.196	2.333	5.000	*expect*	Verb	1.073	0.076	5.000
picture	Noun	3.656	2.281	5.000	*mean*	Verb	1.075	0.092	5.000
break	Verb	2.389	2.219	5.000	*clear*	Adj.	1.118	0.114	5.000
throw	Verb	4.638	2.165	5.000	*talk*	Verb	2.838	0.125	5.000

Table E.4: Largest and smallest standard deviations in human context concreteness ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
hand	Noun	4.992	1.400	5.000	*thick*	Adj.	5.064	0.216	5.000
complicated	Adj.	3.962	1.381	5.000	*sick*	Adj.	5.004	0.267	5.000
work	Verb	3.987	1.377	5.000	*clear*	Adj.	3.220	0.356	5.000
break	Verb	3.770	1.293	5.000	*light*	Noun	4.175	0.360	5.000
expect	Verb	4.063	1.288	5.000	*speak*	Verb	5.002	0.362	5.000
fire	Noun	4.690	1.242	5.000	*miss*	Verb	3.650	0.369	5.000
sweet	Adj.	3.986	1.213	5.000	*team*	Noun	5.749	0.376	5.000
successful	Adj.	4.411	1.212	4.000	*detail*	Noun	4.889	0.395	5.000
eat	Verb	5.145	1.191	5.000	*sound*	Verb	3.890	0.404	5.000
system	Noun	4.308	1.178	5.000	*plant*	Noun	5.154	0.407	5.000

E.2 Imageability

Table E.5: Largest and smallest absolute errors between human and GPT-4o isolated imageability ratings.

Largest 10					Smallest 10				
Lemma	POS	Human iso.	GPT iso.	Abs. error	Lemma	POS	Human iso.	GPT iso.	Abs. error
current	Adj.	497.000	300.000	197.000	*door*	Noun	599.000	600.000	1.000
speak	Verb	488.000	300.000	188.000	*ring*	Verb	601.000	600.000	1.000
tough	Adj.	486.000	300.000	186.000	*hand*	Noun	598.000	600.000	2.000
link	Noun	481.000	300.000	181.000	*matter*	Noun	298.000	300.000	2.000
true	Adj.	365.000	200.000	165.000	*try*	Verb	303.000	300.000	3.000
work	Noun	458.000	300.000	158.000	*accurate*	Adj.	305.000	300.000	5.000
bet	Verb	453.000	300.000	153.000	*heavy*	Adj.	495.000	500.000	5.000
make	Verb	351.000	200.000	151.000	*carry*	Verb	393.000	400.000	7.000
poor	Adj.	447.000	300.000	147.000	*extensive*	Adj.	307.000	300.000	7.000
business	Noun	441.000	300.000	141.000	*return*	Verb	293.000	300.000	7.000

Table E.6: Largest and smallest absolute shifts between GPT-4o isolated ratings and mean GPT-4o context ratings for imageability. N refers to the number of sentences containing the lemma as the target word.

Largest 10						Smallest 10					
Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*	Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*
rise	Verb	7.000	2.319	4.681	5.000	*keep*	Verb	3.000	2.997	0.003	5.000
ring	Verb	6.000	2.224	3.776	5.000	*come*	Verb	3.000	2.991	0.009	5.000
face	Verb	7.000	4.045	2.955	5.000	*sure*	Adj.	2.000	2.012	0.012	5.000
cut	Verb	5.000	2.137	2.863	5.000	*war*	Noun	6.000	6.041	0.041	5.000
build	Noun	6.000	3.245	2.755	5.000	*distinct*	Adj.	3.000	2.952	0.048	4.000
world	Noun	6.000	3.413	2.587	5.000	*wait*	Verb	3.000	3.052	0.052	5.000
view	Noun	5.000	2.498	2.502	5.000	*tight*	Adj.	4.000	3.945	0.055	5.000
cold	Adj.	6.000	3.648	2.352	5.000	*black*	Adj.	5.000	4.944	0.056	5.000
hand	Noun	6.000	3.776	2.224	5.000	*finish*	Verb	3.000	2.942	0.058	5.000
point	Noun	4.000	1.802	2.198	5.000	*show*	Verb	3.000	3.068	0.068	5.000

Table E.7: Largest and smallest standard deviations in GPT-4o context imageability ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
hand	Noun	3.776	2.835	5.000	*sure*	Adj.	2.012	0.071	5.000
room	Noun	5.040	2.201	5.000	*responsible*	Adj.	2.241	0.096	5.000
face	Verb	4.045	2.000	5.000	*friendly*	Adj.	3.036	0.099	5.000
build	Noun	3.245	1.981	5.000	*child*	Noun	6.122	0.102	5.000
body	Noun	4.040	1.882	5.000	*consider*	Verb	2.093	0.110	5.000
stand	Verb	3.899	1.852	5.000	*proud*	Adj.	2.920	0.136	4.000
short	Adj.	3.351	1.811	5.000	*extensive*	Adj.	2.331	0.139	4.000
mother	Noun	4.950	1.749	5.000	*dull*	Adj.	2.715	0.145	4.000
side	Noun	4.441	1.690	5.000	*war*	Noun	6.041	0.149	5.000
eat	Verb	4.977	1.688	5.000	*health*	Noun	3.109	0.151	5.000

Table E.8: Largest and smallest standard deviations in human context imageability ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
come	Verb	4.287	1.517	5.000	*war*	Noun	5.309	0.205	5.000
house	Noun	5.602	1.341	5.000	*send*	Verb	3.600	0.268	5.000
job	Noun	4.144	1.338	5.000	*return*	Verb	3.747	0.278	5.000
make	Verb	3.964	1.277	5.000	*heart*	Noun	4.667	0.294	5.000
company	Noun	4.302	1.274	5.000	*turn*	Verb	4.524	0.310	5.000
face	Verb	4.702	1.251	5.000	*big*	Adj.	3.704	0.313	5.000
mother	Noun	5.011	1.181	5.000	*eat*	Verb	5.250	0.319	5.000
leave	Verb	3.830	1.161	5.000	*community*	Noun	3.824	0.322	5.000
matter	Noun	2.947	1.133	5.000	*thick*	Adj.	5.120	0.336	5.000
place	Noun	3.962	1.103	5.000	*life*	Noun	4.142	0.348	5.000

E.3 Iconicity

Table E.9: Largest and smallest absolute errors between human and GPT-4o isolated iconicity ratings.

Largest 10					Smallest 10				
Lemma	POS	Human iso.	GPT iso.	Abs. error	Lemma	POS	Human iso.	GPT iso.	Abs. error
drive	Verb	4.700	1.000	3.700	*little*	Adj.	3.000	3.000	0.000
original	Adj.	5.400	2.000	3.400	*funny*	Adj.	2.900	3.000	0.100
high	Adj.	5.300	2.000	3.300	*face*	Verb	2.300	2.000	0.300
top	Adj.	5.167	2.000	3.167	*come*	Verb	1.600	2.000	0.400
ring	Noun	6.100	3.000	3.100	*friendly*	Adj.	3.400	3.000	0.400
add	Verb	4.100	1.000	3.100	*mouth*	Noun	3.400	3.000	0.400
helpful	Adj.	5.000	2.000	3.000	*heavy*	Adj.	3.700	3.000	0.700
fit	Verb	5.000	2.000	3.000	*help*	Verb	3.700	3.000	0.700
go	Verb	4.900	2.000	2.900	*return*	Noun	2.700	2.000	0.700
central	Adj.	3.900	1.000	2.900	*slim*	Adj.	5.727	5.000	0.727

Table E.10: Largest and smallest absolute shifts between GPT-4o isolated ratings and mean GPT-4o context ratings for iconicity. N refers to the number of sentences containing the lemma as the target word.

Largest 10						Smallest 10					
Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*	Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*
slim	Adj.	5.000	2.928	2.072	5.000	*able*	Adj.	1.000	1.000	0.000	5.000
help	Verb	3.000	1.031	1.969	5.000	*sell*	Verb	1.000	1.000	0.000	5.000
stuff	Noun	3.000	1.061	1.939	5.000	*office*	Noun	1.000	1.001	0.001	5.000
friendly	Adj.	3.000	1.076	1.924	5.000	*area*	Noun	1.000	1.001	0.001	4.000
funny	Adj.	3.000	1.147	1.853	5.000	*know*	Verb	1.000	1.001	0.001	5.000
link	Noun	3.000	1.202	1.798	5.000	*fresh*	Adj.	3.000	3.001	0.001	5.000
ring	Noun	3.000	1.249	1.751	5.000	*follow*	Adj.	1.000	1.001	0.001	5.000
wicked	Adj.	3.000	1.300	1.700	5.000	*distribution*	Noun	1.000	1.002	0.002	5.000
speak	Verb	3.000	1.313	1.687	5.000	*expect*	Verb	1.000	1.003	0.003	5.000
low	Adj.	3.000	1.415	1.585	5.000	*reasonable*	Adj.	1.000	1.003	0.003	5.000

Table E.11: Largest and smallest standard deviations in GPT-4o context iconicity ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
stop	Verb	2.685	1.642	5.000	*able*	Adj.	1.000	0.000	5.000
fresh	Adj.	3.001	1.531	5.000	*sell*	Verb	1.000	0.000	5.000
heavy	Adj.	2.504	1.462	5.000	*know*	Verb	1.001	0.000	5.000
mouth	Noun	2.142	1.345	5.000	*area*	Noun	1.001	0.001	4.000
fire	Noun	2.465	1.041	5.000	*helpful*	Adj.	1.002	0.001	5.000
huge	Adj.	2.211	1.025	5.000	*office*	Noun	1.001	0.001	5.000
live	Adj.	2.061	1.010	5.000	*follow*	Adj.	1.001	0.001	5.000
drive	Verb	1.945	0.976	5.000	*distribution*	Noun	1.002	0.002	5.000
wait	Verb	1.536	0.940	4.000	*house*	Noun	1.002	0.002	5.000
vast	Adj.	2.205	0.865	5.000	*give*	Verb	1.004	0.002	5.000

Table E.12: Largest and smallest standard deviations in human context iconicity ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
popular	Adj.	2.915	1.334	5.000	*break*	Verb	4.144	0.137	5.000
little	Adj.	3.993	1.248	5.000	*sergeant*	Noun	3.200	0.200	5.000
helpful	Adj.	3.180	1.232	5.000	*stuff*	Noun	4.160	0.207	5.000
fresh	Adj.	4.455	1.230	5.000	*approach*	Noun	3.227	0.209	5.000
model	Noun	3.057	1.227	5.000	*word*	Noun	2.749	0.215	5.000
follow	Verb	2.877	1.145	5.000	*return*	Noun	3.120	0.228	5.000
friendly	Adj.	3.597	1.116	5.000	*size*	Noun	3.409	0.235	5.000
familiar	Adj.	3.550	1.108	5.000	*change*	Noun	3.460	0.254	5.000
slim	Adj.	4.319	1.087	5.000	*take*	Verb	3.267	0.272	5.000
link	Noun	3.400	1.078	5.000	*work*	Noun	3.113	0.272	5.000

E.4 Familiarity

Table E.13: Largest and smallest absolute errors between human and GPT-4o isolated familiarity ratings.

Largest 10					Smallest 10				
Lemma	POS	Human iso.	GPT iso.	Abs. error	Lemma	POS	Human iso.	GPT iso.	Abs. error
bit	Noun	460.000	700.000	240.000	*fall*	Verb	546.000	500.000	46.000
mean	Verb	499.000	700.000	201.000	*bed*	Noun	636.000	700.000	64.000
community	Noun	499.000	700.000	201.000	*tired*	Adj.	630.000	700.000	70.000
level	Adj.	504.000	700.000	196.000	*room*	Noun	627.000	700.000	73.000
allow	Verb	504.000	700.000	196.000	*popular*	Adj.	626.000	700.000	74.000
manage	Verb	508.000	700.000	192.000	*man*	Noun	623.000	700.000	77.000
hide	Verb	515.000	700.000	185.000	*good*	Adj.	621.000	700.000	79.000
clever	Adj.	515.000	700.000	185.000	*go*	Verb	618.000	700.000	82.000
management	Noun	518.000	700.000	182.000	*funny*	Adj.	617.000	700.000	83.000
show	Verb	519.000	700.000	181.000	*old*	Adj.	616.000	700.000	84.000

Table E.14: Largest and smallest absolute shifts between GPT-4o isolated ratings and mean GPT-4o context ratings for familiarity. N refers to the number of sentences containing the lemma as the target word.

Largest 10						Smallest 10					
Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*	Lemma	POS	GPT iso.	Mean context	Abs. shift	*N*
thick	Adj.	7.000	4.658	2.342	5.000	*child*	Noun	7.000	6.903	0.097	5.000
build	Noun	7.000	4.658	2.342	5.000	*go*	Verb	7.000	6.899	0.101	5.000
soft	Adj.	7.000	4.693	2.307	5.000	*team*	Noun	7.000	6.738	0.262	5.000
run	Verb	7.000	4.861	2.139	5.000	*elegant*	Adj.	6.000	5.729	0.271	5.000
air	Noun	7.000	4.930	2.070	5.000	*time*	Noun	7.000	6.722	0.278	5.000
black	Adj.	7.000	5.130	1.870	5.000	*drive*	Verb	6.000	5.671	0.329	5.000
top	Adj.	7.000	5.145	1.855	5.000	*small*	Adj.	7.000	6.667	0.333	5.000
sweet	Adj.	7.000	5.150	1.850	5.000	*come*	Verb	7.000	6.638	0.362	5.000
close	Adj.	7.000	5.161	1.839	5.000	*choose*	Verb	7.000	6.622	0.378	5.000
view	Noun	7.000	5.163	1.837	5.000	*know*	Verb	7.000	6.613	0.387	5.000

Table E.15: Largest and smallest standard deviations in GPT-4o context familiarity ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
number	Noun	5.267	1.784	5.000	*go*	Verb	6.899	0.113	5.000
role	Noun	5.198	1.763	5.000	*team*	Noun	6.738	0.133	5.000
company	Noun	5.699	1.632	5.000	*stupid*	Adj.	6.132	0.154	5.000
dead	Adj.	5.277	1.617	5.000	*child*	Noun	6.903	0.155	5.000
run	Verb	4.861	1.590	5.000	*know*	Verb	6.613	0.185	5.000
drive	Verb	5.671	1.554	5.000	*business*	Noun	6.344	0.188	5.000
full	Adj.	5.381	1.522	5.000	*story*	Noun	5.954	0.193	5.000
house	Noun	6.246	1.494	5.000	*sure*	Adj.	6.519	0.199	5.000
body	Noun	5.203	1.458	4.000	*face*	Noun	6.116	0.203	5.000
bed	Noun	5.869	1.446	5.000	*elegant*	Adj.	5.729	0.211	5.000

Table E.16: Largest and smallest standard deviations in human context familiarity ratings. N refers to the number of sentences containing the lemma as the target word.

Largest 10					Smallest 10				
Lemma	POS	Mean context	SD	*N*	Lemma	POS	Mean context	SD	*N*
serious	Adj.	6.133	1.210	5.000	*care*	Verb	5.989	0.202	5.000
wall	Noun	5.978	1.123	5.000	*raise*	Verb	6.067	0.253	5.000
black	Adj.	5.500	1.118	5.000	*place*	Noun	6.056	0.266	5.000
show	Verb	5.500	1.041	3.000	*stop*	Verb	6.211	0.273	5.000
face	Noun	5.600	1.012	5.000	*go*	Verb	6.689	0.285	5.000
drop	Verb	5.867	0.973	5.000	*mean*	Verb	6.122	0.298	5.000
size	Noun	6.033	0.968	5.000	*funny*	Adj.	6.211	0.303	5.000
sound	Verb	5.389	0.958	5.000	*natural*	Adj.	6.344	0.313	5.000
top	Adj.	5.744	0.921	5.000	*bed*	Noun	5.933	0.320	5.000
elegant	Adj.	5.444	0.919	5.000	*know*	Verb	6.567	0.325	5.000