Investigations into Semantic Underspecification in Language Models

MSc Thesis *(Afstudeerscriptie)*

written by

Franciscus Cornelis Lambertus Wildenburg
(born 4th of January, 1998 in Harderwijk)

under the supervision of **dr. Sandro Pezzelle**, 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:
28th of August, 2023

Members of the Thesis Committee:
Dr. Ekaterina Shutova (chair)
Dr. Sandro Pezzelle (supervisor)
Dr. Wilker Ferreira Aziz
Dr. Alberto Testoni

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

by Francíscus Cornelis Lambertus Wildenburg

Several (position) papers have drawn attention to the challenges semantic underspecification may bring to modern language models, yet relatively little research has been done on this topic. We contribute to this area of research by presenting DUST, a dataset of underspecified sentences annotated with their domain of underspecification. Using this dataset and three experiments using prompts, language model perplexity, and diagnostic classifiers, we study the way modern language models process sentences containing semantic underspecification. We find that the ability of language models to recognize underspecification does not correlate with some commonly used metrics for language models, and that a fine-grained approach to underspecification could greatly benefit the research community.

Acknowledgements

Looking back, it feels like writing this thesis took both many years and no time at all. The former, I believe, is inherent to writing theses, but the latter was made possible only due to the support by many people around me. Here, I would like to thank some of these people.

Firstly, I would like to thank Sandro for giving me guidance and structure during the writing of the thesis. The weekly meetings, in particular, were a key reason for me being able to make regular progress. Secondly, I thank Michael Hanna for some vital contributions to DUST and the execution of the experiments. Thirdly, the Thesis Committee, for their useful feedback and questions during and kind words after the defense.

Outside of the university, too, I have many people to thank. My fellow board members of S.V. Ergo, who definitely made the year... unique . My friends, for making sure I kept a life outside of logic. My mother, for her undying support and love. And perhaps most of all my father. If this thesis contains even a fraction of the wisdom you shared with me, I can be more than happy.

I received support from many more people, but these I leave (sometimes fittingly) underspecified. I hope this thesis will show that this does not make your contribution any less important.

Contents

Dedicated to the memory of my father
Huub Wildenburg (1955-2021)

Chapter 1

Introduction

In recent years, the use of artificial systems to handle or assist with textual tasks has increased dramatically (Kasneci et al., 2023). A key role in this process is taken up by language models, which have entered mainstream culture in roles such as writing aids (Lee, Liang, and Q. Yang, 2022) or dialogue agents (OpenAI, 2023), and it has been argued that such models are or will be able to offer benefits to society, such as aiding in education (Kasneci et al., 2023). To be able to do so, however, these language models need to be able to deal with many aspects of natural language, including some aspects that are traditionally considered more difficult for artificial systems (Khurana et al., 2023; A. Liu et al., 2023; Pezzelle, 2023).

One of these potential challenges is *semantic underspecification*. If a sentence or phrase is semantically underspecified, a language user processing it needs to use non-linguistic information to (fully) understand it (Frisson, 2009; Egg, 2010; Harris, 2020; Pezzelle, 2023). An example of such an utterance is the sentence "We will meet there." – this is a grammatical sentence, but to understand it a reader or listener would need to know, among other aspects, who "we" are and where "there" is. Although multimodal systems might have access to this information to some extent, they do still struggle with accounting for and addressing this phenomenon (Pezzelle, 2023), while text-only systems do not have any access to extra-textual information and have been known to struggle with related phenomena such as ambiguity (A. Liu et al., 2023).

Despite these shortcomings, to the best of the author's knowledge, relatively little research has been done on how language models account for semantic underspecification. What research does exists generally focuses on ambiguity in particular (such as that of A. Liu et al., 2023). This is despite the fact that the relevance of semantic underspecification, even when considered separate from ambiguity, has been pointed out in the literature (e.g. Hutchinson, Baldridge, and Prabhakaran, 2022; Pezzelle, 2023; although not explicitly named, it can also be seen in Bennett et al., 2021).

This thesis attempts to contribute to this area of research by studying the influence of semantic underspecification on language models at a more fine-grained level. To do so, we introduce DUST[1], a fine-grained dataset covering several aspects of underspecification. Experimenting on this dataset, we conclude that while some language models are able to recognize underspecification to some extent, their ability to do so does not correlate with some commonly used metrics for language models. Furthermore, we conclude that a fine-grained approach to underspecification could greatly benefit the research community.

The rest of this work is laid out as follows: in the remaining part of chapter 1, we give a more detailed introduction to semantic underspecification and its relevance

[1] DUST, as well as the code used for the experiments in the remainder of this thesis, can be found at `https://github.com/frank-wildenburg/DUST`

to language models. In chapter 2 we introduce DUST, a collection of semantically underspecified sentences and specified counterparts annotated with the domain of underspecification they belong to. Then, in chapters 3, 4 and 5 we use this dataset to study the ability of language models to recognize and self-correct for underspecification in text, the impact of underspecification on language model perplexity, and the possibility of recognizing the presence of underspecification encoded in a model's representations. Finally, in chapter 6, we summarize our findings and draw conclusions for the use of language models.

1.1 Semantic Underspecification

While specific definitions of underspecification may differ in details, it is generally agreed that the phenomenon can be roughly thought of as the situation where a word, phrase, sentence or other fragment of language leaves open possibilities by not specifying one or more semantic details (Sennet, 2023), which have to be interpreted using non-linguistic information to (fully) understand the utterance. This phenomenon is widespread in natural language, ranging from features that differ between languages such as language-specific inflectional behaviour (Franzon and Zanini, 2022) to phenomena seen in multiple languages like determiners such as *many* or *few* (Lappin, 2000), and sometimes even to animal communication or non-linguistic forms of communication (Harris, 2020).

Underspecification can be caused by the relation between language and perception. When communicating some perceived information, describing every perceived detail would prevent the receiving party from incorrectly inferring an unintended meaning, but it would also require the speaker or writer to communicate a large amount of likely redundant information (Piantadosi, Tily, and Gibson, 2012).

Yet underspecification can also be used to great effect in exclusively textual settings. An example of this can be found in the *Earthsea* cycle by writer Ursula Le Guin. Here, the skin colour of the main character, as well as that of many others, is left underspecified. No attempt is made to mislead the reader into thinking these characters are white-skinned, but because of assumptions made by the reader, one might be surprised when the book 'reveals' that this character is dark-skinned (Bellot, 2018). Such examples show that the presence or absence of underspecification might cause or prevent a natural language user to make certain inferences, which could in turn impact their future decision making, sometimes in culturally charged situations where nuance is required to avoid conflict.

1.1.1 Causes for Underspecification in Language

Given that semantic underspecification might lead to incorrect inferences being made, one might wonder why underspecification is such a commonly occurring phenomenon in natural language. However, research has shown that ambiguity and underspecification are in reality desirable features of communication systems (Piantadosi, Tily, and Gibson, 2012). Given the availability of context information, ambiguity and underspecification allow for language users to ignore redundant (and therefore inefficient) information (Piantadosi, Tily, and Gibson, 2012) as well as the re-using of words, sounds, symbols and grammatical features that are more easily produced or understood or that allow for more efficient information compression (Piantadosi, Tily, and Gibson, 2012; Franzon and Zanini, 2022; Pezzelle, 2023).

This advantage of underspecification is further strengthened by the fact that language users can generally efficiently integrate context information (Harris, 2020) and make inferences with this information (Pezzelle, 2023), especially compared to the time consumption and cognitive load of speaker articulation (Levinson, 2000). These facts cause underspecification to lead to increased processing efficiency. Hence, in order for artificial systems to master human language use, they should be able to deal with underspecification (Pezzelle, 2023) – perhaps especially so given the environmental costs of (inefficient communication by) modern language models.

1.1.2 Underspecification and Ambiguity

As can already be seen in the previous section, underspecification and ambiguity share several features. This, combined with the fact that underspecification and ambiguity are sometimes used interchangeably in common parlance (Sennet, 2023); that several comparable but distinct terms such as 'under-determined' (Lappin, 2000), 'unspecified' (Sennet, 2023), or 'lack of specification' (Zwicky and Sadock, 1975) are also used; and the fact that research often uses more than one of these terms (Zwicky and Sadock, 1975; Poesio, 1994; Hutchinson, Baldridge, and Prabhakaran, 2022; A. Liu et al., 2023) might make it difficult to distinguish between underspecification and ambiguity. Here, we briefly make explicit the relation and difference between ambiguity and underspecification. Of course, to precisely define these two terms requires more time and effort than what is given to them here; we limit ourselves to distinguishing the two.

Ambiguity, roughly defined, may be considered as the phenomenon where a word, sentence, phrase or other linguistic utterance has two or more valid readings or interpretations (Zwicky and Sadock, 1975; Sennet, 2023). For example, the sentence "I saw the man with the telescope" may just as well be read as "I saw the man through a telescope" as "I saw the man who had a telescope". Underspecification, meanwhile, requires that an utterance 'leaves open' possibilities for those reading the text; possibilities that can then be 'filled in' by non-linguistic information to fully understand the text (Zwicky and Sadock, 1975; Frisson, 2009; Egg, 2010; Harris, 2020; Pezzelle, 2023). An example of such a sentence might be "we will meet there". This sentence requires extra-lingual information to clarify who 'we' are and where 'there' is. Note, however, that once this information is available, the sentence has only one reading. In other words, even though this sentence is underspecified, it is not ambiguous.

The converse is often not true. After all, if an utterance has two or more readings, it is often, if not always, possible to determine which of these readings is the intended one by considering the context in which the utterance was made (Piantadosi, Tily, and Gibson, 2012). In this way, it is possible to see underspecification as a generalization of ambiguity. This perhaps also explains the claim of Sennet (2023) that in common parlance, "simple underspecificity will suffice for a charge of ambiguity".

1.1.3 Previous Research on Underspecification

Being primarily a linguistic phenomenon, underspecification has most often been studied from a linguistic perspective, but this is not always the case. Here, we will give a short and non-exhaustive introduction to the previous research done into underspecification.

Earlier work on semantic underspecification considered the phenomenon from the perspective of (formal) semantics. For example, Zwicky and Sadock (1975) consider ambiguity and underspecification and which tests linguists can be used to distinguish the two. Later, Lappin (2000) gave an intensional parametric semantic for vague quantifiers through the use of an underspecified parametric interpretation. Throughout the 90s and the 00s, underspecified representations were also used as a possible approach for natural language processing (NLP) systems to deal with processing the many possible readings of ambiguous sentences. To that end, Poesio (1994) gives a formal analysis of ambiguity processing that uses underspecified representations, whereas Dwivedi, Goldhawk, and Mailhot (2009) give an analysis of semantic underspecification and anaphora. This research direction can also be seen in the work of Pinkal and colleagues (e.g. Niehren, Pinkal, and Ruhrberg, 1997; Pinkal, 1999).

Throughout the latter half of the 00s and the 2010s, more research was done on the causes of underspecification. This was done from the perspective of linguistics, for example by Wasow, Perfors, and Beaver (2005), but also through that of neuroscience and language processing (see e.g. the overview by Frisson, 2009) and that of information theory (e.g. Piantadosi, Tily, and Gibson, 2012; Franzon and Zanini, 2022).

In recent years, the beginnings of an investigation of the relevance of underspecification for (large) language models have begun to form. We further investigate this research direction in section 1.3.

1.1.4 Egg's Domains of Semantic Underspecification

One particularly detailed investigation into underspecification is that of Egg (2010). Egg gives a classification of ambiguity into four classes according to two criteria: whether the reading comprise the same semantic material and whether it is possible to give a single syntactic analysis for all the readings. It is these four classes that we will use in the remainder of this work to further distinguish between types of underspecification. Here, we detail these four classes. For brevity, we may occasionally refer to the classes by the number assigned to them here.

1: Semantically and Syntactically Homogeneous Expressions

The expressions that fall under this class are both syntactically and semantically homogeneous. Here, semantically homogeneous should not be taken as meaning that all readings of the expressions have the same semantic meaning on the level of the expression as a whole (for indeed this would mean that these readings are not ambiguous or underspecified), but rather that the semantic value of the individual components of the expression are identical. As examples of these types of expressions, Egg mentions scopal ambiguities such as quantifier scope ambiguities, and ambiguities caused by other scope-bearing items such as negation and modal expressions.

2: Semantically but not Syntactically Homogeneous Expressions

Expressions that fall under this class comprise of the same semantic material, but this material is arranged in different ways in such a way that the semantic value of the whole expression is different between readings. Examples of this kind of expressions, according to Egg, are modifier attachment ambiguities.

3: Syntactically but not Semantically Homogeneous Expressions

The third type of expressions are those who share a single syntactic structure, but who do not share the same semantic material in its components. Egg categorizes these expressions into four subgroups: expressions with lexically ambiguous words, polysemy (but not homonymy), reinterpretation (metonymy and aspectual coercion) and referential ambiguity and missing information.

This last group is especially relevant for our work. Where Egg only briefly mentions the two cases, and restrict the group of missing information to those (parts of) expressions "that could not be decoded due to problems in production, transmission or reception", we find that much of the work on underspecification in language models (see below) concerns this class of missing information, especially if interpreted more broadly to also include those expressions where information is missing because it was consciously or unconsciously not included by the creator of the utterance. Examples include deictic expressions (where the missing information is that information that refers to whom/what/when the expression refers) and expressions that are underspecified due to certain information not being mentioned (e.g. a description of a person not including their skin colour).

Similarly, referential ambiguity is also mentioned for only a few sentences, but this group includes sentences such as those included in the Winograd schema challenge (Levesque, Davis, and Morgenstern, 2012), which is generally considered to be very relevant for language models

Missing information, in particular, might be seen as something considerably different from the other domains. Where all other domains can also be considered to be instances of ambiguity, information being missing cannot always be considered as such. Rather, instances of missing information such as deixis are types of underspecification that is not ambiguous.

4: Neither Syntactically nor Semantically Homogeneous Expressions

Finally, to the class consisting of expressions that are neither syntactically nor semantically homogeneous, Egg assigns only homonymic expressions and expressions that contain homonymic components.

1.2 (Large) Language Models

Formally, a language model is a model that assigns probabilities to sequences of words (Jurafsky and Martin, 2000), often (although not necessarily) assigning the highest probabilities to the words considered most likely by some measure of likeliness. Although such models can be simple in theory, the kind of model most often used associated with the term 'language model' nowadays are so-called 'large language models', or LLMs, which consist of large artificial neural networks trained using deep-learning techniques.

Although these models are often trained to simply predict the most likely word based on a context, their usage has become widespread due to their ability to perform tasks not explicitly trained on when trained using sufficient data (e.g. Radford et al., 2019; Brown et al., 2020). For this reason, the usage of language models to handle or assist with textual tasks has increased dramatically (Kasneci et al., 2023), with the release of ChatGPT (OpenAI, 2023) in particular introducing language models into the mainstream.

When trained on a word prediction task, language models may either be tasked to predict how a segment continues (so-called autoregressive models) or to predict a word missing from a segment, not necessarily at the end of the context window (so-called 'masked' language models) (Zaib, Sheng, and Emma Zhang, 2020). This thesis generally deals with autoregressive models.

The possibility of models showing such generalization when trained using sufficient training data has led to these models being trained on datasets of ever increasing size. To create such large datasets, information is often scraped from the internet without detailed curation. However, this brings with it the risk of the models displaying biases or other negative effects (Bender et al., 2021). Furthermore, the large, 'black box' nature of the models does not allow researchers to determine in advance which linguistic capacities they do or do not possess. Due to this, any research on (the negative effects of) language models being unable to deal with certain linguistic features will have to be done after training, often while the model is already in use elsewhere (Dodge et al., 2021).

1.3 Semantic Underspecification and Language Models

In recent years, the rise of modern language models has led to a modest amount of research being done on the way ambiguity and underspecification impact modern NLP systems. This research generally focuses on disambiguating or clarifying unclear sentences. For example, Berzak et al. (2015) studied the ability of multimodal systems to disambiguate ambiguous sentences through the use of visual depictions of the sentence, while in 2022 one of the tasks in SemEval-2022 concerned identifying plausible clarifications of underspecified phrases (Roth, Anthonio, and Sauer, 2022).

More recently, in 2023, A. Liu et al. (2023) studied the way language models model ambiguity and found that even state-of-the-art language models achieve much lower scores than humans in recognizing ambiguity and generating disambiguations in natural language inference tasks. Other research supports these findings (Ortega-Martín et al., 2023; Fantechi, Gnesi, and Semini, 2023), although authors do note that such language models can already aid in ambiguity detection tasks despite their shortcomings (A. Liu et al., 2023; Ortega-Martín et al., 2023; Fantechi, Gnesi, and Semini, 2023).

However, to the best of the author's knowledge, the amount of research done on this topic remains limited. Nor can more general evaluations of language models give us much insight on the performance of such models on ambiguous or underspecified sentences, since ambiguous or underspecified instances are often excluded in the creation of curated benchmarks (A. Liu et al., 2023) or limited to specific forms of underspecification such as Winograd schemas (Levesque, Davis, and Morgenstern, 2012). Instances of underspecification that are not ambiguous, in particular, remain understudied.

This is especially problematic given that the importance of underspecification in modern language models *has* been explicitly noted in multiple (position) papers. Hutchinson, Baldridge, and Prabhakaran (2022), for example, note that many challenges in text-to-image tasks can be caused by both ambiguity in descriptions and underspecification of desired depictions. These challenges, they argue, may give rise to risks such as the amplification of societal biases, the generation of harmful or offensive content, the spread of mis- or dis-information, or the generation of unsafe images in high-risk scenarios. Pezzelle (2023), meanwhile, argues that the research community should be aware of semantic underspecification in order to develop successful multimodal NLP systems, again reporting that state-of-the-art multimodal NLP models struggle with underspecification.

The importance of language models being able to correctly process and generate underspecified sentences can also be seen in papers outside of the NLP community. Bennett et al. (2021), for example, do not use the term 'underspecification', nor do they extensively discuss technical details of NLP systems, but their discussion of (mis)representation in image descriptions of topics such as skin colour, gender and disability show that the question of when information about these topics should be shared is complicated and contested. Hence, for an artificial system to be able to successfully generate such image descriptions, it should know when to specify such features or leave them underspecified. Being able to do so will make technology more accessible and prevent biases from being amplified. Similarly, Franzon and Zanini (2022) argue that being able to account for (underspecified) grammatical features may help the scientific community pursue language inclusiveness.

Chapter 2

DUST: A Dataset of Underspecified Sentence Types

To study the impact of semantic underspecification on the workings of language models we present DUST, a **D**ataset of **U**nderspecified **S**entence by **T**ype consisting of 3226 underspecified sentences and equally many specified counterparts, collected and adapted from various datasets, each annotated with the type/domain of underspecification it belongs to in order to facilitate a detailed investigation of the various facets of underspecification.

2.1 Dataset Creation

To collect examples of underspecified expressions, we select sentences from existing datasets on underspecification and ambiguity, as well as filtering sentences that contain word that are considered indicative of ambiguity or underspecification. Where datasets are separated into training, development or test splits, we take data from all splits. Specifically, the dataset is created from the following sources:

2.1.1 Semantically and Syntactically Homogeneous Expressions

The sentences covering semantically and syntactically homogeneous expressions are collected from the Language and Vision Ambiguities (LAVA) dataset (Berzak et al., 2015), a multimodal dataset consisting of ambiguous sentences and visual data that disambiguates these sentences. Out of the sentences collected from this dataset, 40 sentences containing scopal ambiguity cover this domain of underspecification. Notably, these sentences were not collected or filtered from existing text, but rather generated using Part of Speech tag sequence templates and a fixed dictionary of words.

For each of these sentences, we create a non-ambiguous counterpart based on the visual disambiguation provided in the dataset. These counterparts create a sentence that is semantically equivalent with the intended reading of the original sentence by rephrasing the sentence. For example, the sentence “Danny approached the person with a blue telescope” might be rephrased as “Danny approached the person while holding a blue telescope” or “Danny approached the person who had a blue telescope”.

2.1.2 Semantically but not Syntactically Homogeneous Expressions

The set of sentences covering semantically but not syntactically homogeneous expressions also consists of sentences drawn from or based on the LAVA dataset. This domain is covered by 108 sentences containing VP and PP attachment ambiguity.

Non-ambiguous counterparts are created in the same way as for the previous domain.

2.1.3 Syntactically but not Semantically Homogeneous Expressions

The domain of syntactically but not semantically homogeneous expressions is the largest of the domains in the DUST dataset, with sentences originating from multiple datasets. These are:

Language and Vision Ambiguities

From the LAVA dataset, 89 more sentences containing referential ambiguity or sentences with missing information are used for this domain. Specified counterparts are once again created in the same manner as described above.

CLArifying Insertions from Revision Edits

To further extend the third class of sentences, we collect sentences from the CLArifying Insertions from Revision Edits (CLAIRE) dataset used for SemEval-2022 task 7 (Roth, Anthonio, and Sauer, 2022). This dataset consists of sentences from instructional texts from wikiHow which were edited in order to insert a word or phrase. The authors argue that these revisions may clarify sentences that were initially unclear or that left information implicit in the text. For this reason, we consider the text pre-edit to be more underspecified than the same text post-edit, and we include the non-revised sentences in our datasets with the revised sentences as a specified counterpart.

The sentences from this datasets can be divided into four categories: implicit references (216 sentences), where a reference that was initially non-verbalized is made explicit; fused heads (532 sentences), where the head noun of a noun phrase is made explicit through the edit; noun compounds (774 sentences), where the dependent noun is inserted to form a more specific compound; and metonymic references (91 sentences), where a revision adds a noun to make explicit to which aspect of another noun the text refers. Since in all these cases the sentence is made less underspecified due to information being added, we classify these sentences under 'missing information' and hence as instances of syntactically but not semantically homogeneous ambiguity. However, we do retain the original type of clarification (i.e. fused head, metonymic reference etc.) in our dataset as a subclass.

Due to the processing applied to the sentences by the original authors, not all original (pre-edit) sentences provided in the CLAIRE dataset are grammatical. To prevent this from unduly influencing our results, we remove every sentence pair for which the pre-edit sentence has a GRUEN score lower than 0.8. This metric evaluates (among others) the grammaticality of sentences, and unlike other evaluation metrics has the advantage of not requiring references or human supervision (Zhu and Bhat, 2020). The threshold of 0.8 was experimentally found to provide a sufficient recall and precision for our purposes. This procedure leaves 1613 sentences out of the original 5000. To ensure that the language model-based filtering does not impact the results of our experiments, we verify that the difference in perplexity between the filtered sentence pairs is similar to that between that of the entire dataset.

Winograd Schema Challenge

To further investigate the influence of semantic underspecification in the form of referential ambiguity, we include the original 273 sentences from the Winograd schema challenge (Levesque, Davis, and Morgenstern, 2012). In these sentences, one specific word results in a pronoun ambiguously referring to one of two antecedents earlier in the sentence. Hence, these sentences are further examples of syntactically but not semantically homogeneous ambiguity, and referential ambiguity in particular.

To create non-ambiguous counterparts to the sentences from the dataset, we change the gender or plurality of one of the two antecedents (as well as that of the pronoun, where applicable) so that the pronoun can only refer to one of the antecedents. For example, the sentence "Although they ran at about the same speed, Sue beat Sally because she had such a bad start" might become "Although they ran at about the same speed, Sue beat John because he had such a bad start".

It should be noted that the (lack of) ambiguity of these sentences is partially dependent on how the model associates certain names with their intended gender. However, this problem is also present in the original sentences, and names that were already in the dataset were used for the changes, so that the influence hereof should be equal on both the data and control data[1].

Deixis

Another type of sentences that may be considered to contain missing information are those with deictic expressions, i.e. expressions that refer to a specific time, place or person in context. To collect sentences that contain deictic expressions we create a list of commonly used deictic expressions[2], and for each word select approximately ten sentences containing that deictic expression[3] from a collection of sentences sampled from English Wikipedia[4]. We use Wikipedia because we believe that these sentences may be reasonably specified on average, leading to any impact of underspecification being due to the deictic expressions only. As counterparts to these sentences we select equally as many sentences that contain none of the words in this list. Due to the size of the sample not all words have exactly ten corresponding sentences, leading to a total of 123 underspecified and 123 relatively specified sentences.

2.1.4 Neither Syntactically nor Semantically Homogeneous Expressions

To cover the class of sentences that are neither syntactically nor semantically homogeneous, we create a selection of sentences containing homonymic expressions in the same way as we collect sentences containing deictic expressions. We do so based on a list of 100 homonyms selected by Maciejewski and Klepousniotou based on linguistic principles, dictionary entries and subjective ratings (Maciejewski and Klepousniotou, 2016). Once again, for each word we select approximately 10 sentences containing it to be in our dataset, as well as equally as many sentences containing

[1] For more information on the role of gender in the Winograd schema challenge, see, e.g., Rudinger et al. (2018) or Abdou et al. (2020)

[2] The exact list is "I", "you", "we", "she", "they", "here", "there", "that", "this", "now", "then", "tonight", "tomorrow", "yesterday", "left" and "right", based on information from studysmarter.us: `https://www.studysmarter.us/explanations/english/pragmatics/deictic-expressions/`, accessed 5th of June 2023

[3] It is theoretically possible that the sentences also contain other deictic expressions

[4] Available at `https://www.kaggle.com/datasets/mikeortman/wikipedia-sentences`, accessed 5th of June 2023

none of the homonyms for the control dataset, resulting in a total of 980 underspecified and 980 relatively specified sentences.

The sentences obtained in this manner, as well as those that are examples of deixis, are different from those mentioned previously in that the underspecified sentence and its specified counterpart do not form a minimal pair in the same way that the sentences from the previously mentioned sources do. This carries the risk of influencing results in our experiments, potentially giving rise to results that are caused by features other than underspecification alone. For this reason, we take extra care that the results found for the homonomy and deixis subclasses are comparable to those found for the subclasses which are made up of minimal pairs of sentences.

2.1.5 Various domains

Finally, during the writing of this thesis, A. Liu et al. (2023) published an investigation into ambiguity in language models. Alongside this paper they released AmbiEnt, a dataset for natural language inference with ambiguous sentences annotated (A. Liu et al., 2023). This dataset consists of premise-hypothesis pairs where either the premise or the hypothesis might be ambiguous or underspecified; for such sentences, the dataset provides disambiguating phrasings of these sentences.

Due to the way these sentences were curated and generated, this dataset does not, to our knowledge, consist of only one of Egg's classes. Due to this, we do not include these sentences in DUST. However, we do use this dataset to validate the outcomes of our experiments.

As underspecified sentences, we take any premise or hypothesis that is labelled by the authors to be ambiguous; as specified counterparts we randomly take one of the disambiguations provided by the authors. Doing so, we obtain 543 sentences that may be used for verification.

2.2 DUST Statistics

The resulting dataset consists of 3226 underspecified sentences with equally as many specified counterparts. To give an indication of the distribution of the data, we describe some basic statistics of the dataset in table 2.1. The average word frequency is based on the word frequency of the 333.333 most commonly-used single words on the English web, as derived from the Google Web Trillion Word Corpus and normalized to a value between 0 and 1[5].

These statistics already suggest that there are considerable differences between the underspecified and specified counterparts of the sentences in the dataset, but also that these differences are sometimes dependent on the domain of underspecification the sentence belongs to. Whether this is is inherently correlated with underspecification or something that results from the different domains of underspecification originating from different datasets is something we will examine later in this thesis. However, it does already show the importance of not testing across domains without considering the potential impact of the sentences having different distributions.

Looking at the subclass level for those domains where this is possible, we get the results of tables 2.2 and 2.3. Here, also, we can see that the differences between subclasses can be quite severe. For example, the subclass of expressions with metonymic

[5] available at https://www.kaggle.com/datasets/rtatman/english-word-frequency, accessed 11th of July 2023

Class	Overall	1	2	3	4
Avg. sen. len.	**17.04** (±10.65)	**7.60** (±1.72)	**7.79** (±1.41)	**14.12** (±8.26)	**24.67** (±11.73)
	15.81 (±8.52)	**8.75** (±1.69)	**8.98** (±1.49)	**14.63** (±7.28)	**19.38** (±10.11)
Avg. word len.	**4.67** (±0.74)	4.84 (±0.69)	4.66 (±0.59)	**4.53** (±0.73)	**4.99** (±0.66)
	4.82 (±0.77)	4.71 (±0.52)	4.58 (±0.52)	**4.66** (±0.73)	**5.19** (±0.75)
Average word freq.	**0.14** (±0.07)	0.17 (±0.10)	**0.16** (±0.10)	0.13 (±0.08)	**0.15** (±0.06)
	0.13 (±0.07)	0.15 (±0.11)	**0.15** (±0.11)	0.13 (±0.07)	**0.14** (±0.07)
Vocab. size	13633	36	39	7517	8196
	13767	51	58	7991	7592

TABLE 2.1: Average sentence length (in words), word length (in characters) and normalized word frequency and standard deviations (between parentheses), and vocabulary size, of the test (upper row) and control (lower row) datasets per domain of Egg. All numbers are rounded to two decimal places. For every row but vocabulary size, numbers shown in bold indicate a significant difference ($p < 0.05$) between that statistic in the test data and control data, as shown by a Wilcoxon rank-sum test.

Subclass	PP attachment ambiguity	VP attachment ambiguity
Average sentence length	**8.02** (±1.55)	**7.60** (±1.28)
	9.17 (±1.58)	**8.83** (± 1.42)
Average word length	4.64 (±0.59)	4.68 (±0.59)
	4.55 (±0.48)	4.61 (±0.56)
Average word freq.	0.18 (±0.11)	0.15 (±0.10)
	0.16 (±0.12)	0.14 (±0.10)
Vocabulary size	37	38
	52	51

TABLE 2.2: Statistics of table 2.1 split up per subclass of domain 2 of Egg

references shows no significant differences between control and test data for any of the tested statistics. This is especially remarkable when considering the 'average sentence length' statistic, given that the sentences in the control dataset consist of sentences from the test dataset with additional information inserted into the sentence.

Of course, the fact that some of these statistics show significant differences between test and control dataset is worth being cognizant of when considering the results of the experiments in the sections to come. While it might be the case that these differences are inherent in the differences between underspecified and specified sentences, it might also be the case that any effects found in the sections to come might be impacted by the difference in one or more of these features, rather than the

Subclass	Add. comp.	Deixis	Fused Head	Impl. ref.	Met. ref.	Missing info.	Ref. amb.
Avg. sen. len.	**12.35** (±7.23)	**25.11** (±14.21)	**14.13** (±7.33)	**12.87** (±7.97)	16.73 (±7.04)	**8.13** (±1.77)	15.34 (±5.85)
	13.34 (±7.23)	**18.18** (±8.30)	**15.13** (±7.33)	**14.81** (±7.96)	17.73 (±7.04)	**9.19** (±1.84)	15.49 (±5.70)
Avg. word len.	**4.59** (±0.83)	**4.85** (±0.61)	**4.56** (±0.68)	4.26 (±0.68)	4.45 (±0.61)	4.61 (±0.68)	4.38 (±0.57)
	4.74 (±0.79)	**5.14** (±0.76)	**4.73** (±0.69)	4.30 (±0.61)	4.55 (±0.60)	4.64 (±0.66)	4.43 (±0.56)
Avg. word freq.	**0.14** (±0.08)	0.14 (±0.05)	0.12 (±0.07)	**0.23** (±0.07)	0.17 (±0.07)	0.20 (±0.11)	0.13 (±0.07)
	0.12 (±0.07)	0.14 (±0.06)	0.11 (±0.06)	**0.17** (±0.08)	0.16 (±0.06)	0.18 (±0.12)	0.12 (±0.07)
Vocab. size	3464	1529	2567	1236	750	37	1060
	3836	1258	2831	1381	817	54	1072

TABLE 2.3: Statistics of table 2.1 split up per subclass of domain 3 of Egg

Domain of Egg	LAVA	CLAIRE	WSC273	Sampled from Wikipedia
1	40	0	0	0
2	108	0	0	0
3	89	1613	273	123
4	0	0	0	980

TABLE 2.4: Amount of sentences in DUST originating from each dataset, divided by the domain of Egg they belong to.

difference in specification (alone).

Finally, the amount of sentences originating from each dataset and belonging to each domain can be seen in table 2.4.

Chapter 3

Can Language Models Identify and Self-Correct Underspecification?

One of the first steps towards being able to correctly handle underspecification is being able to recognize the presence or absence of such underspecification. Hence, if language models are able to correctly distinguish underspecified sentences from specified sentences, we may have reasons to believe that they are less susceptible to the risk described in the previous chapters. A second step would be for the models to self-correct for underspecified language. This would mean that the pitfalls of underspecified language might be able to be avoided by prompting a language model to avoid such language.

To test if models are able to do so, we use sentences from the DUST dataset in prompts for language models. We find that language models are, to some degree, able to recognize underspecified sentences, although this ability differs per domain of underspecification. We then give a qualitative analysis of the explanations the models give of the underspecification of the sentences as well as possible disambiguations generated by the language models.

3.1 Introduction

If we wish to know how language models respond to underspecification, we may start by giving it an underspecified text and seeing how the model responds. Although simple, such 'prompt-based learning' has recently become an oft-used method of study. Here, we give a short introduction to this paradigm.

3.1.1 Language Models and Recognizing Through Prompts

The ability of language models to recognize (linguistic) features of text can be studied through prompts given to a language models. Through such a prompt – a set of instructions provided to a language model that specifies what the model's output should contain – a language model may be 'asked' to identify whether or not said linguistic feature occurs in a text. Compared to fine-tuning existing models for a specific research direction, research through such prompts requires less time and resources (P. Liu et al., 2023; Wang et al., 2023), as well as requiring only textual input and hence less technical know-how. Such 'prompt-based learning' has been called a new paradigm in natural language processing (P. Liu et al., 2023) and has been the focus of a large amount of research in the past years.

It is perhaps unsurprising, then, that some research on linguistic ambiguity through prompts has already been done. Fantechi, Gnesi, and Semini (2023) and Ortega-Martín et al. (2023) have tested the ability of ChatGPT to detect the presence of absence of ambiguity in a text – the former with a focus on different types of ambiguity, and the latter in a comparison between rule-based NLP systems. However, these studies only consider a small amount of text. A more comprehensive study was done by A. Liu et al. (2023); however, this study focuses on Natural Language Inference (NLI) and does not distinguish explicitly between different kinds of ambiguity or underspecification.

3.1.2 Language Models and Self-Correcting

One specific form of research using prompts is research in the ability of large language models to self-correct. In such research, the prompt(s) are altered – possibly, though not necessarily, with the use of automated tools in between prompts – in a way that attempts to automatically prevent the language model from showing some undesired behaviour. Previous research has investigated the ability of language models to auto-correct for, among others, generating faulty code (Chen et al., 2023; Gou et al., 2023), biases and stereotypes (Ganguli et al., 2023; Gou et al., 2023) and hallucination (Gou et al., 2023).

Here, too, previous research on ambiguity exists. A. Liu et al. (2023) generates disambiguations for ambiguous sentences in the AmbiEnt datasets and evaluates these disambiguations both automatically and by means of human evaluation, finding that the correctness of these disambiguations is considerably lower than that of the disambiguations in the dataset itself. Ortega-Martín et al. (2023) take the opposite approach, asking ChatGPT to generate ambiguous sentences, and find that the model does not consider these sentences to be ambiguous when later asked whether or not they are..

3.2 Can Language Models Recognize Underspecification?

We first investigate whether language models are able to recognize semantic underspecification when prompted to do so. To do so, we prompt XLNet (large; Z. Yang et al., 2019), GPT-2 (xl; Radford et al., 2019), Flan-T5 (xxl; Chung et al., 2022) and OPT (13b; Zhang et al., 2022)[1] with both a semantically underspecified sentence from the DUST dataset and a specified counterpart, asking the model which of these sentences is more semantically underspecified. Specifically, the prompt used is:

> “Here are two sentences. A: “_”. B: “_”. Which one of these is more semantically underspecified? Please respond by outputting only A or B. Answer:”

where the underscores are replaced by one of the two sentences. The order in which the sentences are placed is randomized so as to prevent biases in the model from unduly influencing the final accuracy.

Each sentence pair in the dataset was included in a prompt once. We noted both the accuracy of the model as well as whether they answered A or B. The results of this can be seen in tables 3.1 and 3.4. Details per subclass can be seen in tables 3.2 and 3.3.

[1] We limited our investigation to autoregressive language models; the models used and size thereof were limited by availability and compute required for the models. Future research could be aimed at extending our research to models with higher requirements as well as different types of models, such as masked language models through the use of pseudo-perplexity.

Model	Overall	1	2	3	4
XLNet large	0.1265	0.1250	0.1296	0.1196	0.1408
GPT-2 xl	0.3084	0.175	0.3056	0.3146	0.3010
OPT 13b	0.5043	0.4500	0.5648	0.4857	**0.5398**
Flan-T5 xxl	**0.7117**	**0.7500**	**0.6389**	**0.8356**	0.4531

TABLE 3.1: Accuracy of models in recognition task per domain of Egg. All numbers are rounded to four decimal places. Numbers shown in bold indicate an accuracy significantly above chance ($p < 0.05$)

Model	PP attachment ambiguity	VP attachment ambiguity
XLNet	0.0625	0.1833
GPT-2	0.3333	0.2833
OPT	0.5625	0.5667
Flan-T5	0.5208	**0.7333**

TABLE 3.2: Accuracy of models in recognition task per subclass in domain 2 of Egg. All numbers are rounded to four decimal places. Numbers shown in bold indicate an accuracy significantly above chance ($p < 0.05$)

Model	Add. comp.	Deixis	Fused Head	Impl. ref.	Met. ref.	Missing info.	Ref. amb.
XLNet large	0.1059	0.1220	0.1297	0.1065	0.1319	0.0566	0.1521
GPT-2 xl	0.3152	0.3577	0.3064	0.3056	0.2527	0.2453	0.3463
OPT 13b	0.4858	0.4715	0.4643	0.5417	0.5164	0.5849	0.4628
Flan-T5 xxl	**0.9677**	0.5203	**0.8590**	**0.9583**	**0.8132**	**0.6604**	0.5405

TABLE 3.3: Accuracy of models in recognition task per subclass in domain 3 of Egg. All numbers are rounded to four decimal places. Numbers shown in bold indicate an accuracy significantly above chance ($p < 0.05$)

Model	A	B	Neither
XLNet large	804	49	2373
GPT-2 xl	1109	914	1203
OPT 13b	3116	110	0
Flan-T5 xxl	1010	2116	0

TABLE 3.4: Amount of times a model chose an output during recognition task. 'Neither' comprises all answers other than 'A' or 'B'.

Based on the results, we can see that although all models decide which sentence is underspecified partially through biases between A or B, it is possible for the models to perform above chance level. This is especially visible for Flan-T5, which performs above chance level for the entire dataset as well as for every domain except for the domain consisting of neither syntactically nor semantically homogeneous expressions. In contrast, OPT performs above chance level for only that domain[2]. Once again, it is important to note here that unlike domains 1 and 2 and the large majority of domain 3, domain 4 does not consist of true minimal pairs, which might be the cause of this result.

That Flan-T5 is often able to recognize semantically underspecified sentences might be considered as surprising, given that that same model was unable to achieve a performance much higher than chance when recognizing possible disambiguations in the task of A. Liu et al. (2023). To verify the correctness of these findings, we also run the above task on the ambiguous sentences and disambiguations of the AmbiEnt dataset, which results in an accuracy of 0.7459. This might indicate that while language models *are* able to recognize underspecified or ambiguous language, they are *not* able to correctly reason with such ambiguous language – a hypothesis strengthened by the findings of A. Liu et al. (2023) and of Ortega-Martín et al. (2023) that language models are not internally consistent across questions regarding ambiguity. We may take this as meaning that recognizing underspecified or ambiguous text is an easier task for language models than correctly reasoning with such underspecification/ambiguity.

To investigate the influence of prompt variation, we also prompt Flan-T5 with several variations of the prompt. Several variations were found to to increase the model's bias to one of the two answers or increase the amount of errors made by the model (e.g. "outputting only B or A", "outputting only "A" or "B"", "outputting only 1 or 2").

Interestingly, changing the prompt from 'more semantically underspecified' to 'more ambiguous' shows no significant difference in accuracy for domain 1 and 2 of Egg ($p > 0.05$) but a statistically significantly lower accuracy of 0.7793 for the third domain ($p < 0.001$). Although this may be (partially) due to the difference in amount of sentences available for each domain, it could also indicate that the model is correctly able to discern that the sentences that belong to this domain, which often originate from the CLAIRE dataset, are often underspecified but not necessarily ambiguous. Changing the prompt to 'more unclear' shows a significantly lower accuracy ($p < 0.001$) for every domain except the second domain ($p > 0.05$), suggesting that the model might not always consider underspecified sentences as unclear, even when it is able to recognize them as underspecified.

3.3 Can Language Models Explain Underspecification?

To test whether language models are also able to explain why they classify a sentence as underspecified, we give a qualitative analysis of Flan-T5 xxl prompted with the following prompt for a sample of 10 sentences for each subclass in our dataset:

> "Here is a sentence: "_". This sentence is underspecified. Please explain why this sentence is underspecified. Answer:"

[2] Although OPT has an accuracy considerably above 0.5 for domain 2 and some subclasses of domain 3, this accuracy is not significantly above chance level, likely due to the relatively small size of these (sub)domains

The full results can be found in our GitHub repository. We find that in general, Flan-T5 does not answer with an explanation, but rather with a sentence similar to the one in the prompt. Often, this sentence is simply the prompt, potentially with some information added or subtracted. When information is added, this does sometimes make the sentence provided by the model less underspecified than the sentence used in the prompt.

Rarely, the model does respond with a (correct) explanation for why a sentence is underspecified. For example, when the prompt includes the sentence “Add whatever extra twists you like and your doll is done!", the model output is “The extra twists are not specified to be a specific type of twist.". However, and especially in the case of sentences originating from the CLAIRE dataset, the part of the sentence the model mentions as being underspecified or that the model adds more information to is not the part of the sentence that differs between the test and control datasets. This seems to indicate that the sentences originating from the CLAIRE dataset are perceived by the model as being underspecified in more locations than the one they were selected for.

For sentences originating from the LAVA dataset, the model often hallucinates new information. This information does generally make the sentence more specified. For example, when prompted with the sentence “Someone moved the bags”, the model outputs “Someone moved the bags from the car to the garage.". Occasionally, the generated sentence specifies the underspecified prompt. For example, “Danny left the person holding a blue telescope" results in the model outputting “The person holding the telescope is Danny.". However, more often the new sentence is less underspecified by adding more information, rather than by solving the ambiguity present in the sentences.

Finally, for the sentences from the homonymy or deixis subclasses, the model output is often (a rephrasing of part of) the sentence used in the prompt, potentially with added (hallucinated) information added, without giving an explanation for why the sentence is underspecified.

3.4 Can Language Models Generate Specified Sentences?

In the previous section, we have seen that language models are sometimes able to rephrase underspecified sentence in less underspecified ways. To see whether such self-correction can be explicitly provoked, we give a qualitative analysis of Flan-T5 xxl prompted with the following prompt for a sample of 10 sentences for each subclass in our dataset:

> “Here is a sentence: “_". This sentence is underspecified. Please formulate a version of this sentence that is less underspecified. Answer:"

The full results can also be found in our GitHub repository. We find that in general, Flan-T5 appears to be better at this task than at that of the previous section, giving more specified phrasings of the provided sentences quite often (although it still often copies the provided sentence entirely, too). This is perhaps to be expected, given that in the previous experiment the model often provided a comparable sentence rather than an explanation, which benefits it in its task here.

When the model returns a more specified sentence, it generally does so by adding some information to the provided sentence. This means that the kind of underspecification corrected by the model is that of the ‘missing information’ kind. Compared to the attention paid to other forms of underspecification in research this may be considered surprising; in preliminary research for this thesis, the author found that this is also the kind of underspecification often corrected by more modern language models such as ChatGPT.

Critically, this type of fix for underspecification by adding more information is also often applied for the provided sentences that contain some form of ambiguity. For example, the sentence “Danny looked at Andrei moving a yellow chair" is rephrased as “Danny looked at Andrei moving a yellow chair with a sledgehammer." without clarifying whether the chair was moved by Andrei or Danny. This is not always the case – for example, the sentence “Danny left Andrei; he held a telescope" is correctly disambiguated as “Danny left Andrei and held a telescope." – but does happen regularly.

Finally, we note that the results between the two experiments above are not always internally consistent. For example, while Flan-T5 was correctly able to recognize that, and why, the sentence “Add whatever extra twists you like and your doll is done!" was underspecified, it does not change the sentence when asked to phrase it in a way that makes it less underspecified. This seems to match the claims of Ortega-Martín et al. (2023) that language models are not internally consistent when handling ambiguous text.

3.5 Discussion

In the above chapter, we have tested whether language models can recognize, explain, and correct for underspecified sentences. The results suggest that some, although not all, language models are able to do so to some extent. Considering the general phenomenon of language models achieving higher performance as scale increases, it perhaps comes as no surprise that the largest models tested achieve the highest performance. Yet it is important to note that model size does not seem to be the only factor at play. For example, the version of OPT with 13 billion parameters only achieved above chance performance on sentences that were neither syntactically nor semantically ambiguous – and even then only slightly above chance – whereas Flan-T5 xxl, with 2 billion fewer parameters, performed significantly and considerably above chance level on all domains of Egg *except* for the one OPT performed well on.

This difference also suggests that a model performing well at recognizing one type of underspecification or ambiguity does not necessarily imply that that model will be able to do well at recognizing all types of underspecification in sentences. This is an important distinction, considering that the distinction between underspecification and ambiguity is not always made clear in existing research.

The above findings support some previous research while contrasting others. For example, the finding that language models are not always internally consistent when regarding ambiguity or underspecification matches the findings of Ortega-Martín et al. (2023), whereas the results of Flan-T5 at recognizing underspecification and ambiguity might be surprising when considering that A. Liu et al. (2023) found that these same models are not able to correctly reason with ambiguity.

Finding out that some language models are able to correctly recognize and correct for underspecification is good news, considering the importance of underspecification in language described in the first chapter. However, it does not explain why and how language models are able to do so. This is something we will investigate in chapters 4 and 5.

Chapter 4

Underspecification and Language Model Perplexity

The previous chapter has shown that language models are sometimes able to recognize underspecified language to some extent, yet this does not yet tell us whether language models process specified and underspecified language in different ways. Hence, we might wish to know if a change in (under)specification is correlated with a change in model behaviour 'under the hood'.

One metric by which language models can be evaluated is *perplexity*, which can be intuitively seen as how 'surprised' a language model is by some supplied text or collection of texts. Since perplexity is a core measure of language model behaviour, a difference in perplexity between specified and underspecified text could be a clear sign that language models truly do treat underspecified and specified language differently. One reason to believe this may be the case is the fact that human reading behaviour, which is thought to correlate with language model perplexity, is (at least in some cases) indeed influenced by the presence of absence of semantic underspecification in a text (Swets et al., 2008; Rayner and Duffy, 1986).

To test this hypothesis, we investigate the perplexities of several commonly-used language models on the sentences in the DUST dataset, testing whether there is a significant difference between the perplexity of underspecified sentences compared to their specified counterparts, whether the way this difference manifests differs between different domains of underspecification, and whether language model perplexity may be used to predict the (under)specification of a sentence.

4.1 Introduction

Perplexity is an old, oft-used metric of language models, having been used since long before large language models claimed the place they currently have in modern AI research. Here, we give a short introduction to this metric.

4.1.1 Perplexity and Language Models

Since the task of a language model is, at its core, to assign probabilities to sequences of words, one can intrinsically evaluate language models by comparing the sequences of words with the highest probabilities with the sequences of words that are encountered in practice. This is the main idea behind *perplexity*, an intrinsic metric for evaluating language models.

Intuitively, the perplexity of a text represents how 'surprised' a language model is by that text, i.e. the probability assigned to that text by the language model. Formally, this is represented by the inverse probability of the text normalized by the number of words. Hence, for a sentence $S = w_1 w_2 \dots w_N$:

$$\text{perplexity}(S) = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}} \tag{4.1}$$

Alternatively and equivalently, we may see perplexity as the exponential of the cross-entropy of a language model:

$$\text{perplexity}(S) = 2^{-\frac{1}{N}\log P(w_1 w_2 \dots w_N)} \tag{4.2}$$

The higher the perplexity assigned to a text by a language model, the lower the probability that that language model assigns to that text.

Although intrinsic evaluation metrics in general and perplexity in particular have their drawbacks as evaluation metrics for language models (Kuribayashi et al., 2021; Meister and Cotterell, 2021), perplexity has nevertheless often been used as such an evaluation metric (Jelinek et al., 1997; Jurafsky and Martin, 2000), in part due to the fact that language models with lower perplexity were reported to have more human-like behaviour (Kuribayashi et al., 2021).

Given this popularity, it should come as no surprise that perplexity has been used as an informative metric in many studies into (large) language models. For examples, studies have been done on what kind of linguistic structures affect perplexity in neural language models (Miaschi et al., 2021), how perplexity can be used to efficiently select training data for language models (Toral, 2013), how perplexity can be used to choose better performing prompts for language models (Gonen et al., 2022), how language model perplexity relates to the ability of said model to learn in-context (Shin et al., 2022), and so on. Given the drawbacks of perplexity alternative approaches to language model evaluation have been proposed (Meister and Cotterell, 2021), yet as of the time of writing perplexity is still a commonly used metric for language model performance.

4.1.2 Perplexity and Human Language Use

This relation between perplexity and human behaviour has been oft-studied by researchers working on the relation between artificial language models and human language processing. For example, it has been found that language models with lower perplexity often improve at predicting human reading times (Goodkind and Bicknell, 2018) – although not always (Wilcox et al., 2020; Kuribayashi et al., 2021).

Since ambiguity and underspecification are also known to impact human reading times – sometimes resulting in ambiguous sentences being read faster than disambiguated strings (Swets et al., 2008), but at other times increasing the amount of time required to process a sentence (Rayner and Duffy, 1986) – one might wonder if semantic underspecification in sentences then also impacts the perplexity of language models on these sentences. For some types of (temporarily) ambiguous sentences, such as so-called garden-path sentences, such influences have already been found, although they do not perfectly correlate with human processing difficulty (Van Schijndel and Linzen, 2021; Arehalli, Dillon, and Linzen, 2022). However, to the best of our knowledge, no previous work has investigated differences in language model perplexity across different domains of ambiguity and underspecification.

This is especially relevant since different domains of underspecification might have different effects on language model perplexity. For example, where a (temporarily) ambiguous sentence might be expected to increase the perplexity of a language model, the removal of words – another form of underspecification – might sometimes be expected to decrease perplexity. For example, we might expect a language model, like a human, to be less surprised by the sentence 'I ate some chocolate' than by the sentence 'I ate some floating blue chocolate'. In particular, we might expect the effects of the third domain of Egg to behave differently than the other domains, for reasons already discussed in chapter 1.

4.2 Does Underspecification impact Perplexity?

To investigate whether underspecification impacts the perplexity of language models, we calculate the perplexity of the sentences in our datasets as produced by XLNet (base and large; Z. Yang et al., 2019), GPT-2 (base and xl; Radford et al., 2019), Flan-T5 (base and xxl; Chung et al., 2022) and OPT (125m and 13b; Zhang et al., 2022). We then test if there is a significant difference in perplexity between the test and control datasets for each domain of Egg. The results can be seen in figure 4.1.

We find that for most models and domains of Egg, there is a significant difference between the perplexity of the underspecified sentences and that of the (relatively) specified sentences in that domain. For instances of syntactically and semantically homogeneous sentences and instances of semantically but not syntactically homogeneous sentences, this perplexity is generally higher for the underspecified sentences, whereas for the other two domains it is generally lower. However, this occasionally differs per model (such as XLNet for domain 2) or even the size of the model (such as XLNet base for domain 1). It is not necessarily the case that larger versions of the same model show a more significant difference in perplexity between the test and control set (such as seen for Flan-T5 xxl in domain 1), although this is sometimes the case (such as XLNet large for domain 1).

Of particular interest here is the fact that domain 3, consisting of syntactically but not semantically homogeneous sentences, does indeed behave differently from the other domains. Although most models still show a significant difference between perplexity for test and control sentences in this domain, this significant difference occurs here less often than in other domains (with only 5 models showing a significant difference), and even when the models show a significant difference, that difference is often less reliable than that of other domains. Critically, as seen in figure 4.2, this difference in behaviour often remains even when considering only one subclass or dataset, indicating that this difference is not (only) caused by class 3 consisting of multiple datasets, unlike the other classes. This appears to match our assertion in section 1.1.4 that the third domain of Egg appears to be different from the other domains.

Also notable is that figure 4.2 shows that there are considerable differences between the subclasses in the third domain of Egg. Particularly interesting here is the fact that the subclass of referential ambiguity shows no significant differences in perplexity for any model, even though reading time is known to change when referential ambiguity is introduced (Cunnings, Fotiadou, and Tsimpli, 2017). Also notable is that there appears to be a difference between the different subclasses from the CLAIRE dataset, potentially indicating that there may be differences even within the domains and subgroups defined by Egg.

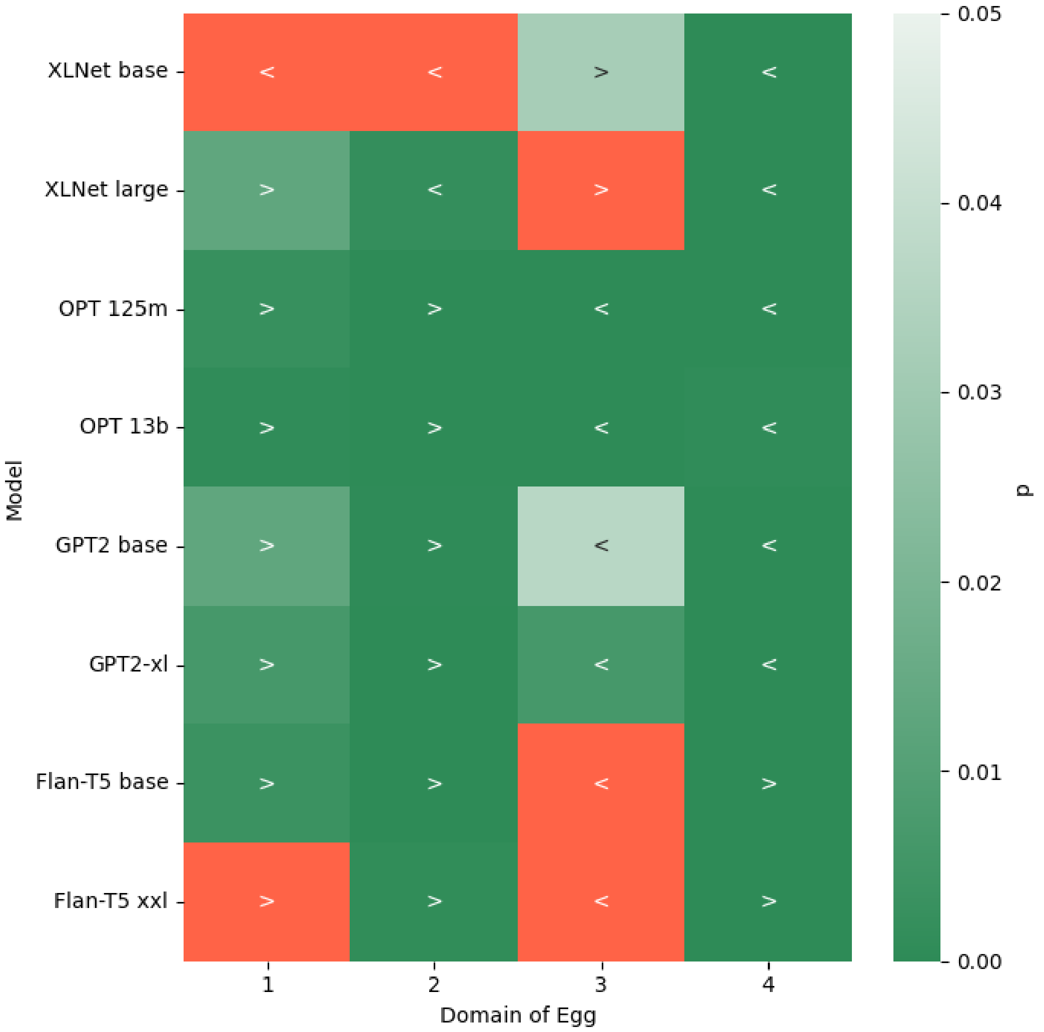

FIGURE 4.1: Significance of differences in perplexity between control and test set per domain of Egg, as calculated using a Wilcoxon rank-sum test. Red fields indicate a significance of $p > 0.05$. ">" indicates the perplexity of the (underspecified) test dataset is higher than that of the control dataset, whereas "<" indicates the opposite.

We verify our findings in two ways. First, to investigate whether the absence of minimal pairs in the deixis and homonomy subclasses may influence our findings, we compare the deixis subclass with all subclasses that do consist of minimal pairs. As can be seen by comparing figures 4.2 and 4.3, we find that although there is occasionally a difference between the sentences consisting of minimal pairs and those without, those differences are comparable to those differences between subclasses that both consist of minimal pairs. This suggests that our findings are valid even for those sentence pairs not consisting of minimal pairs, and hence that also the results for the fourth domain of Egg can be safely considered.

Secondly, we run our experiment for the sentence pairs from the AmbiEnt dataset, the results of which can be seen in figure 4.4. We find that there is a significant difference in perplexity between control and test sentences for every model used, although the direction of this difference differs per model used. Given that figure 4.1 shows that the direction generally differs per domain, and that we believe the AmbiEnt dataset covers multiple domains of Egg, this may be considered to be expected.

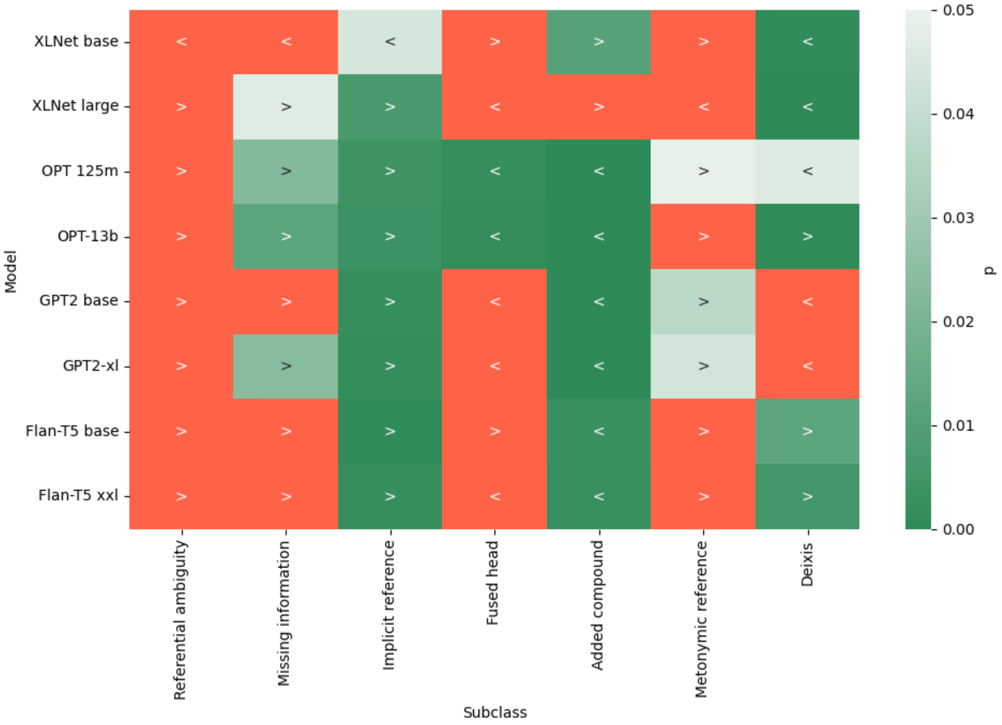

FIGURE 4.2: Significance of differences in perplexity between control and test set per subclass of domain 3 of Egg, as calculated using a Wilcoxon rank-sum test. Red fields indicate a significance of $p > 0.05$. ">" indicates the perplexity of the (underspecified) test dataset is higher than that of the control dataset, whereas "<" indicates the opposite.

4.3 Is Perplexity Predictive of Underspecification?

Given that there appears to often be a significant difference in perplexity between underspecified and specified sentences, one might then wonder if the perplexity of sentences may be used as a method of predicting whether a model will classify a sentence as underspecified or not. Should this be the case, then language model perplexity might be used as a way of predicting whether a language model will perceive a text or dataset as underspecified, which might give an indication as to whether or not underspecification in a text could be a risk towards language model behaviour.

To test whether this can be done, we fit logistic regression models to the sentences with the sentence length, average concreteness, average age of acquisition, average word frequency and perplexity of the Flan-T5 xxl model as independent variables, and the Flan-T5 xxl classification of the sentences in the dataset as the binary dependent variable (where '1' is 'classified by the model as underspecified' and '0' is not). The average concreteness of words is based on concreteness ratings for words by Brysbaert, Warriner, and Kuperman (2014) and is rated on a scale of 1 to 5, where 1 is most abstract and five is most concrete. The average age of acquisition, meanwhile, is based on Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) and represents the average age of acquisition in a word in years. The coefficient of these features can then give us an indication of whether, and to what degree, perplexity is predictive of model output. We use Flan-T5 xxl because of its success in the previous chapter.

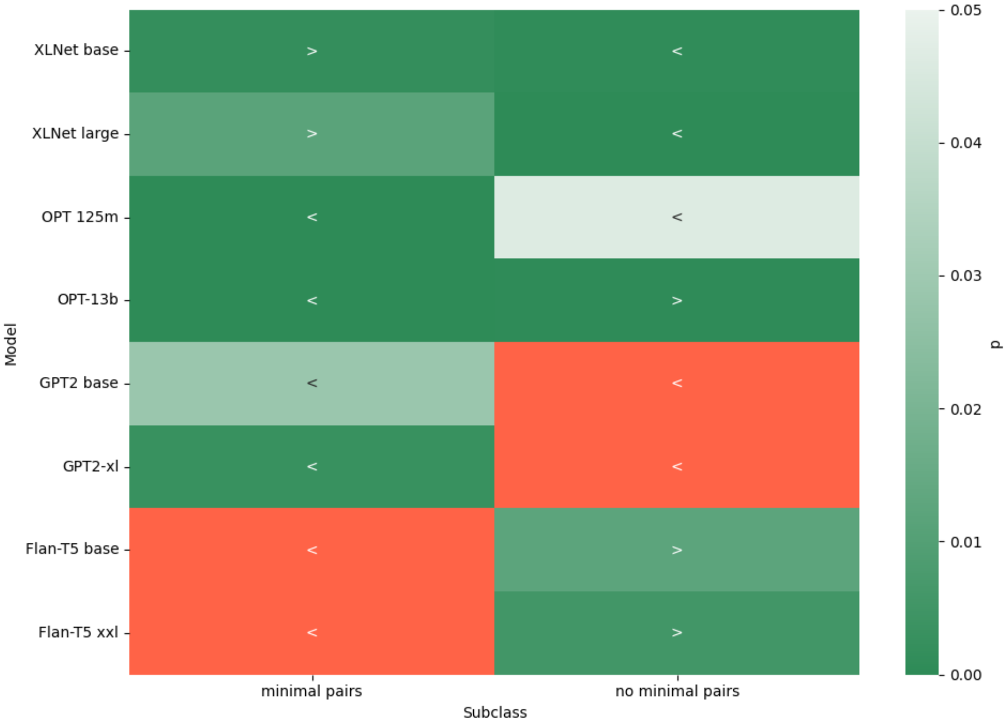

FIGURE 4.3: Significance of differences in perplexity between control and test set for subclasses with and without minimal pairs in the third domain of Egg, as calculated using a Wilcoxon rank-sum test. Red fields indicate a significance of $p > 0.05$. ">" indicates the perplexity of the (underspecified) test dataset is higher than that of the control dataset, whereas "<" indicates the opposite.

The results of this experiment can be found in appendix A. We find that the coefficient of the perplexity feature is generally very small, and although it is statistically significant over the entire dataset, at the domain level it is only statistically significant for domains 3 and 4. These domains differ, however, in that for the fourth domain a higher perplexity makes it more likely for a sentence to be classified as underspecified, whereas the opposite is true for the other three domains. This may be interesting when considering that chapter 3 showed that Flan-T5 performed above chance for every domain except the fourth, possibly suggesting that the model has a lower accuracy for this domain because the difference in perplexity is different than usual.

Further experiments (not included in appendix) show that similar behaviour also occurs when averaging coefficients over all models or all large variants of the models used in the previous experiment, rather than considering specifically Flan-T5 xxl. When the experiment is repeated using only the model perplexity of the sentence as an independent variable, this coefficient is still small and often not significant, indicating that this is not a result of interaction effects between features but rather inherent in the perplexity feature.

Interestingly, we find that when similar models are fitted with the true underspecifiedness of the sentences, rather than the model prediction, as the dependent variable, the coefficient of the perplexity feature is not statistically significant at the domain level for domain 3, but it is for domains 2 and 4. This is interesting when

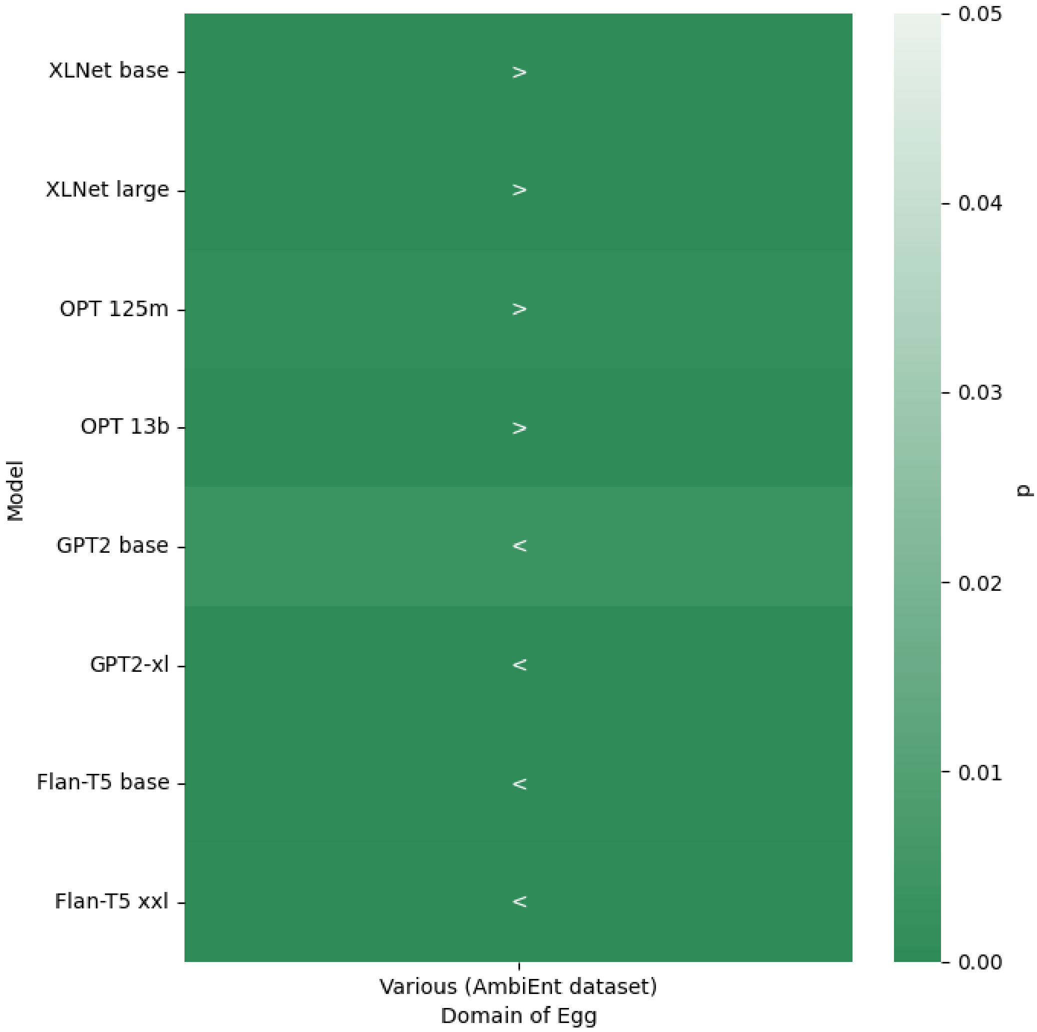

FIGURE 4.4: Significance of differences in perplexity between control and test set for sentence pairs from the AmbiEnt dataset, as calculated using a Wilcoxon rank-sum test. Red fields indicate a significance of $p > 0.05$. ">" indicates the perplexity of the (underspecified) test dataset is higher than that of the control dataset, whereas "<" indicates the opposite.

considering that chapter 3 showed that Flan-T5 performed highest on the recognition task for domain 3, and might indicate that perplexity is to some degree predictive as to whether a language model will recognize a sentence as underspecified. Further results of this experiment did not differ much from the previous one; for completeness, they are also included in appendix A.

Finally, we might consider the fact that the coefficient of the perplexity feature is significant for, and only for, the classes that consist of semantically heterogeneous expressions. This may indicate that the semantic difference between the underspecified and the specified variant of the sentence changes a language model's perplexity of said sentence to such a degree that this difference in perplexity may then be used to identify the underspecified sentence. However, the small size of this coefficient weakens this argument.

4.4 Discussion

The above experiments appear to show that while there is a significant difference in perplexity between specified and underspecified sentences, this difference cannot be used to predict the presence or absence of underspecification of a sentence based on its perplexity. Although this results seems intuitively surprising – one would expect that a difference in distribution between two classes might also be used to predict to which class one instance belongs – the fact that significant variables are not automatically good predictors has been previously studied in various fields. (Lo et al., 2015; Hofman, Sharma, and Watts, 2017).

Nevertheless, knowing that the perplexity of sentences containing underspecification is significantly different to that of sentences without such underspecification can be relevant for future research. For example, since we know that ambiguous (and therefore also at least some underspecified) text is systematically excluded from language model benchmarks (A. Liu et al., 2023), yet underspecification is a common feature of language actually used by humans (Piantadosi, Tily, and Gibson, 2012; Franzon and Zanini, 2022; Roth, Anthonio, and Sauer, 2022), we may conclude that the distribution of text as used by humans is different than that of the text used to test language models, which might result in potential negative influences on language models.

Also relevant is our finding that this difference in perplexity between underspecified and specified sentences is different for underspecified but not ambiguous sentences than it is for ambiguous sentences. The relevance of underspecification even when considered separately from ambiguity has been made clear by several position papers (Hutchinson, Baldridge, and Prabhakaran, 2022; Pezzelle, 2023), yet in the literature underspecification and ambiguity are sometimes used seemingly interchangeably. Our findings show that by researching underspecification through ambiguity, not all aspects of underspecification are covered. This is relevant even outside of the scope of natural language processing: Egg (2010) makes clear that also in the research on underspecification formalisms, (specific forms of) ambiguity are addressed more often than, for example, the elusive subgroup of missing information.

The fact that the difference in perplexity is significant less often for Egg's third domain might also be considered surprising when taking into account the fact that in chapter 3, it was for this domain that Flan-T5 achieved the highest accuracy in recognizing underspecification. This might be taken as an sign that perplexity does not serve as a factor that influences language models to classify (or not classify) a sentence as underspecified. This would also be in line with the small coefficient perplexity has in prediction underspecification, although it would contrast the fact that this coefficient is significant for domain 3, unlike domains 1 and 2. The latter might once again be used as an argument that perplexity is used to accurately predict underspecification, but this argument might be weakened by the fact that it is for domain 4 – for which this coefficient is also significant – that the models achieve the lowest accuracy on the recognition task.

In the next chapter, we will continue our investigation into how language models treat underspecification, now focusing on how and where – if at all – this phenomenon is processed.

Chapter 5

Probing Language Models for Underspecification

The previous chapters have shown that language models are, at least in some situations, able to handle underspecification to some extent, and that this underspecification sometimes correlates with language model perplexity. However, this does not yet tell us where and how this phenomenon is processed.

In this chapter, we shall attempt to answer these questions through the use of *diagnostic classifiers*. Diagnostic classifiers (Hupkes, Veldhoen, and Zuidema, 2018), sometimes referred to as *probes*, are simple, interpretable classifiers (such as linear classifier models), trained with the hidden states of language models as input and some feature or phenomenon as the value that should be predicted. The use of simple classifiers allows us to interpret the hidden states of language models, which are normally considered a black box. If such classifiers are able to correctly classify the presence of absence of a (linguistic) phenomenon (such as underspecification) based on these representations, this can serve as a suggestion that this phenomenon is encoded in this representation.

In our experiments[1], we probe language models using the sentences in the DUST dataset in order to discover whether underspecification is encoded in their hidden states. We also train separate probes for each domain of Egg and for ambiguous vs. underspecified but not ambiguous sentences, to investigate whether these features are encoded differently in the hidden states of language models.

5.1 Introduction

Diagnostic classifiers have recently been used to analyse properties of neural networks, and in particular linguistic properties of language models. Here, we give a short introduction to these classifiers.

5.1.1 Diagnostic Classifiers

With the rise of large (language) models, there has been an increase of research into interpreting these models and analysing which (linguistic) capacities they have. Unfortunately, this is made more difficult by the fact that these models are generally considered to be a black box that is not understandable by humans. For this reason, tests based on human neuroscientific experiments have been developed to help interpret these models (Ettinger, 2020).

[1] The code for the experiments using probing is adapted from code by Jaap Jumelet and Jelle Zuidema for the course Interpretability & Explainability in AI, given at the University of Amsterdam in 2022

One technique that may be used for this task can be found in so-called *diagnostic classifiers* (Hupkes, Veldhoen, and Zuidema, 2018). This approach takes the representations of a neural network such as a language model, which may encode the feature one wishes to examine but are generally too complicated for human interpretation, and trains diagnostic classifiers on these representations to test whether they indeed encode the relevant feature. In order for the diagnostic classifiers to be interpretable by humans, simple classifiers that are generally considered to be human-interpretable (such as linear classifiers) are often used (although not always; see Belinkov, 2022).

If the diagnostic classifier is able to accurately predict the presence or absence of the relevant feature based on the network representations, this may indicate that the feature is encoded within those representations. By training different diagnostic classifiers on different representations (e.g. different hidden layers of a model), these classifiers may give insight on which layers of a (language) model process which features or which parts of a structure, and give insight in potential recurrent or recursive structures in neural networks (Hupkes, Veldhoen, and Zuidema, 2018).

'Probing' language models, as the paradigm has come to be known, has its shortcomings. Especially noteworthy is the fact that the correlation found by the diagnostic classifiers does not necessarily indicate a causal relationship. Indeed, it may be the case that representations encode (something correlated with) a feature, but that the original model does not use the information that was discovered by the probe (Belinkov, 2022). Despite this shortcoming, the probing paradigm has emerged as a prominent analysis strategy, especially in the study of NLP models (Belinkov, 2022).

5.1.2 Probing for Linguistic Features

With this focus on probing for studying NLP models, it should come as no surprise that probing has often been used to investigate whether or not language models encode certain linguistic properties. Diagnostic classifiers have been used to study properties such as identification of main auxiliaries, subject nouns and n$^{\text{th}}$ tokens (representing hierarchical/syntactic and linear information) (Lin, Tan, and Frank, 2019) and number agreement between subject and verb (Giulianelli et al., 2018), as well as how such properties behave under processes such as transfer learning (Durrani, Sajjad, and Dalvi, 2021).

However, to the best of the author's knowledge, probing has not yet been used as an approach for studying ambiguity or underspecification in language models. In this chapter, we aim to remedy this shortcoming, using the DUST dataset as a collection of both specified and underspecified sentences which may be used to train diagnostic classifiers to recognize underspecification based on the hidden states of language models.

5.2 Probing Experiments

To test whether underspecification or ambiguity is encoded in the hidden states of language models, we first train diagnostic classifiers to classify whether a hidden state is the result of a model processing a sentence from the (underspecified) test dataset or the (specified) control dataset.

Due to computational limitations, we restrict ourselves to the 125 million parameter version of OPT (Zhang et al., 2022) and the base version of XLNet (Z. Yang et al., 2019). Although it would have been preferred to (also) study the models found to be successful at recognizing underspecification in the previous chapters, we believe even the results of probing these models give us relevant information as to whether underspecification is encoded in the hidden states of language models such as these.

After extracting the hidden states of the models when fed with the sentences from the DUST dataset, we fit logistic regression models on these hidden states with the (under)specification of the sentences that result in those hidden states as the dependent variable (with this variable being 1 if the sentence was underspecified, and 0 otherwise). We use the logistic regression model from the sklearn python library, with the 'liblinear' solver and a L2 penalty term. This is done for every hidden layer of the models, including one layer (layer 0) for the embeddings of the sentences. We fit the models using 80% of the DUST dataset, and test them using the remaining 20%. After doing so, we calculate their accuracy, precision, recall and F_1 score, calculating metrics for each label and using their unweighted mean for the final value of the latter three metrics.

The results for this first experiment can be found in figure 5.1. We find that for both tested models, the probes achieve an accuracy of about 60%. Although this is not considerably higher than chance level, it is significantly higher than chance, indicating that some feature somewhat correlated with the presence or absence of underspecification is encoded in the hidden states of both of the models. Of course, as described above, it is not sure whether this feature actually is the underspecification itself.

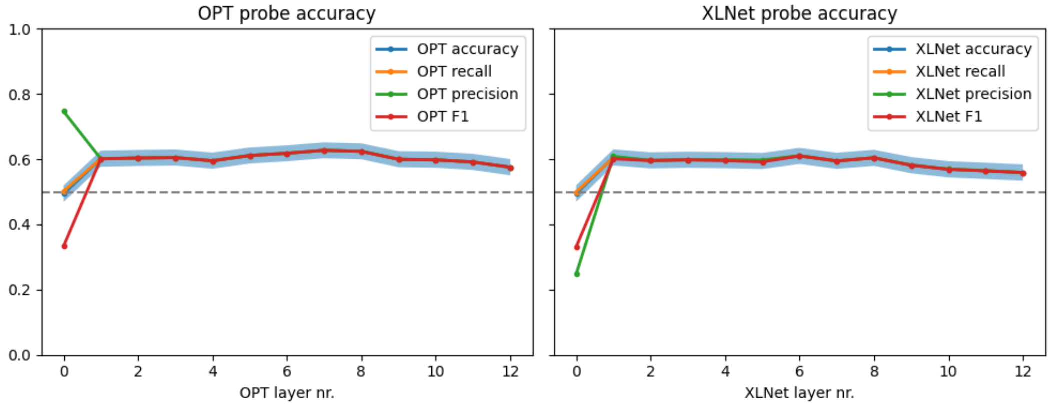

FIGURE 5.1: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of OPT 125m and XLNet base with presence or absence of underspecification as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

Next, we fit a probe in a similar manner for each domain of Egg. The results of this can be seen in figures 5.2 and 5.3. We find that for both models, the accuracy of the probes is generally highest for the hidden states representing semantically but not syntactically homogeneous expressions, then for those of both semantically and syntactically homogeneous expressions, then for those that are homogeneous in neither dimension, then finally for the third domain. This seems to contrast our findings from chapter 3, where the accuracy for the first and second domains (now the highest) was generally lower than that of the third domain, which is now the lowest. Of course, this was mainly the case for Flan-T5, which is not tested here — future research could investigate whether these results differ when tested using the hidden states of a Flan-T5 model.

Another caveat can be found in the fact that for OPT, the accuracy for the second domain was highest for the second domain, although in chapter 3 this difference was not large enough to be significant. It might be the case that, had the second domain consisted of more sentences, this difference would have been significant, in which case we might have seen this as an agreement between those results and the findings of the probing experiment. Even in this case, however, there is still a difference between the accuracy of the probes and that of the model itself, suggesting that either the feature encoded is not underspecification itself or the encoded feature is not used by the model.

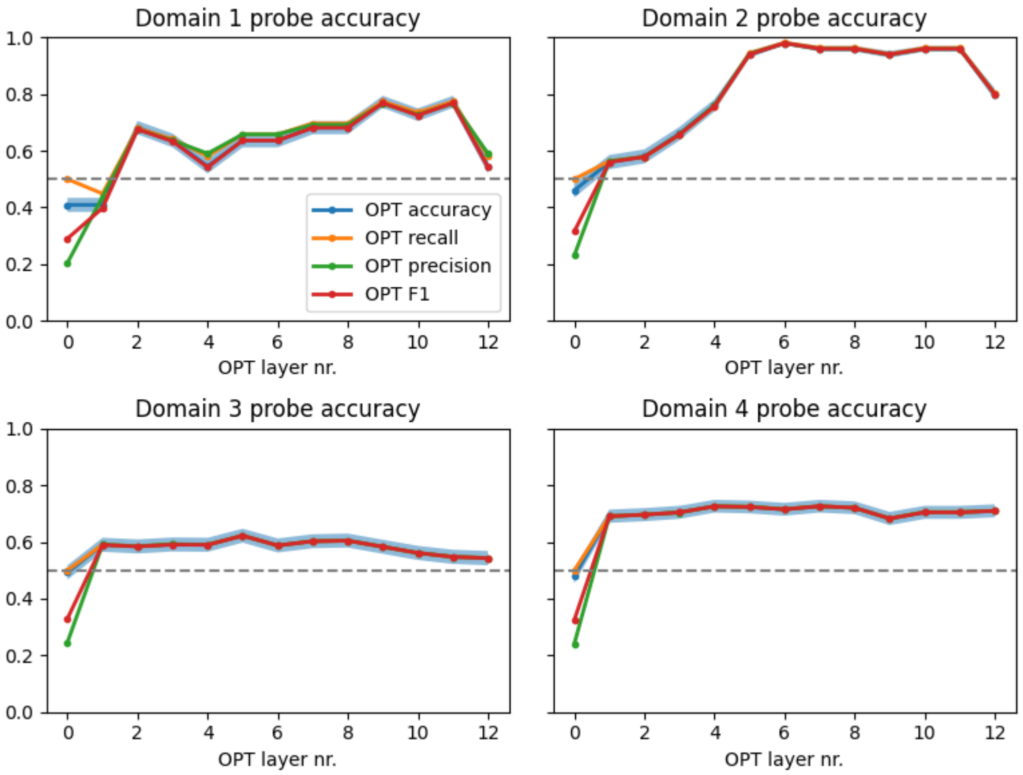

FIGURE 5.2: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of OPT 125m per domain of Egg with presence or absence of underspecification as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

Another interesting observation is the fact that while the probe accuracy for the OPT model is generally higher, or at least not lower, in the later layers, the opposite is true for the probes deployed on the XLNet model. This might indicate that different models process (the feature correlated with) the presence or absence of underspecification differently, although research on more models would be require to reinforce this argument.

Finally, we fit probes for the hidden states of both the ambiguous and the underspecified but not ambiguous expressions, to test whether there is a difference in how ambiguous and underspecified but not ambiguous sentences are encoded in language models. As ambiguous expressions we take all sentences belonging to the subclasses of referential ambiguity, PP attachment ambiguity, VP attachment ambiguity, scopal ambiguity and homonymy, while for the underspecified but not ambiguous expressions we take the sentences belonging to the subclasses of missing

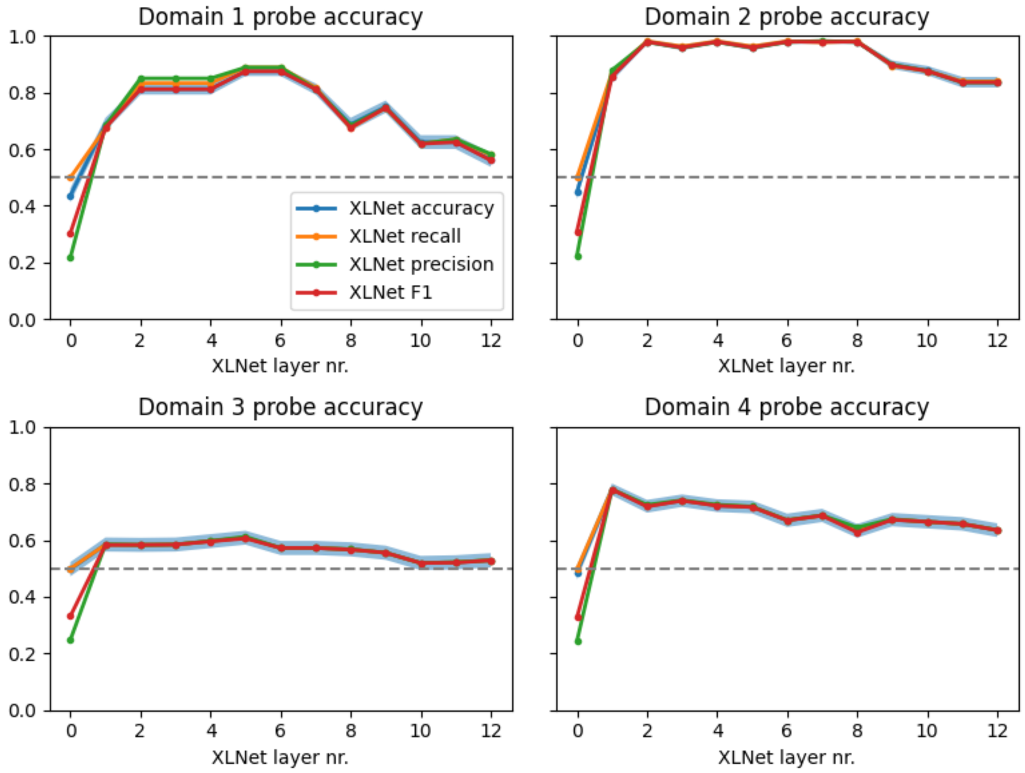

FIGURE 5.3: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of XLNet base per domain of Egg with presence or absence of underspecification as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

information, implicit reference, fused head, added compound, metonymic reference and deixis.

The results of this experiment can be found in figures 5.4 and 5.5. We can see that the probe accuracy for non-ambiguous expressions shows almost the same pattern as that for the probes of domain 3 for the same model, displaying that even though domain 3 is divided into both ambiguous (e.g. referential ambiguity) and non-ambiguous (e.g. missing information) subclasses, the removal of those ambiguous subclasses does not substantially change the way these groups are encoded in the models.

This seems to suggest that ambiguous and underspecified but not ambiguous expressions are not encoded differently in the language models tested here. This is supported by the fact that, unlike the considerable differences in probe accuracy between the domains of Egg, the probe accuracy does not show large differences between ambiguous and non-ambiguous but still underspecified sentences.

We also investigated a variation on the above two experiments by fitting probes with the domain of Egg and ambiguity/underspecification without ambiguity as the value that should be predicted directly. However, both of these experiments resulted in an exceptionally high accuracy for the probes, which we hypothesize is due to the different distributions of the datasets these groups originate from. Hence, we hypothesize, this accuracy cannot give us useful information about underspecification/ambiguity with any certainty, and therefore we do not report on these results here. However, the results are included for completeness in appendix B.

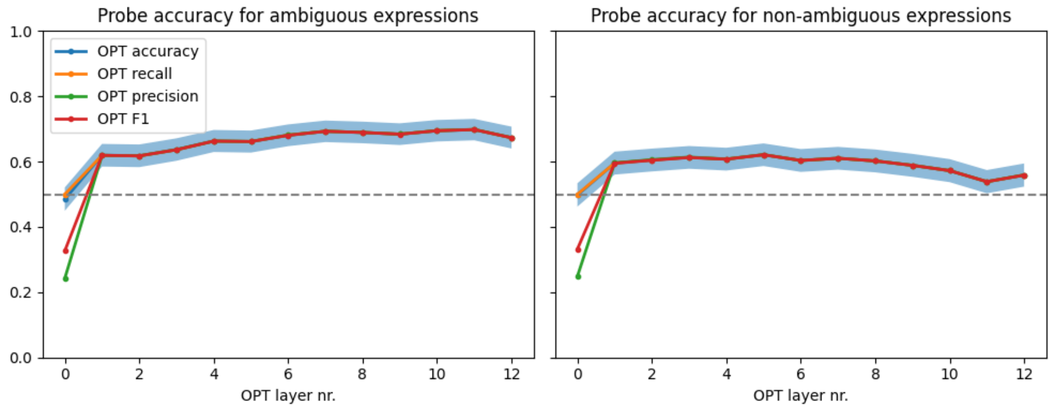

FIGURE 5.4: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of OPT 125m for ambiguous and underspecified but not ambiguous sentences with presence or absence of underspecification as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

5.3 Discussion

The above experiments seem to show that underspecification, or some feature correlated with underspecification, is encoded in the hidden states of the tested language models when these are fed with sentences from the DUST dataset.

These findings are remarkable when considering that the models tested here, OPT and XLNet, were not able to correctly recognize most domains of underspecification – not even when larger variants of the same model were used. This seems to indicate that, should the encoded feature truly be underspecification, this feature isn't used by the models in practice. Future research could potentially investigate whether other tasks, such as variations in prompts or fine-tuning of the models, might be able to use this encoded feature.

Another observation suggested by the above experiments is that different language models might encode underspecification in different ways, with OPT seemingly encoding underspecification (or the feature correlated with it) in later layers, while XLNet seems to encode that same feature within its earlier layers. This, should it be the case, might explain why in chapter 3 the Flan-T5 model achieved higher accuracy for most domains than the OPT model even though the OPT model has more parameters. Besides the possible consequences for research – suggesting that research on underspecification or ambiguity performed on one model might not generalize to different models of the same size – this might also be of use for practical applications of language models, perhaps suggesting that dangers of language models mishandling underspecified language can be alleviated (or exacerbated) by using a different model.

Finally, like in the previous chapters, we once again see a difference between the different domains of Egg. The findings of this chapter appear to match those of chapter 4, in that they once again show that something that is present for the first, second and fourth domains – a significant difference in perplexity for chapter 4, and probe accuracy here – is not present for the third domain. This further gives support to our hypothesis in chapter 1 that this domain is somehow different than the other domains.

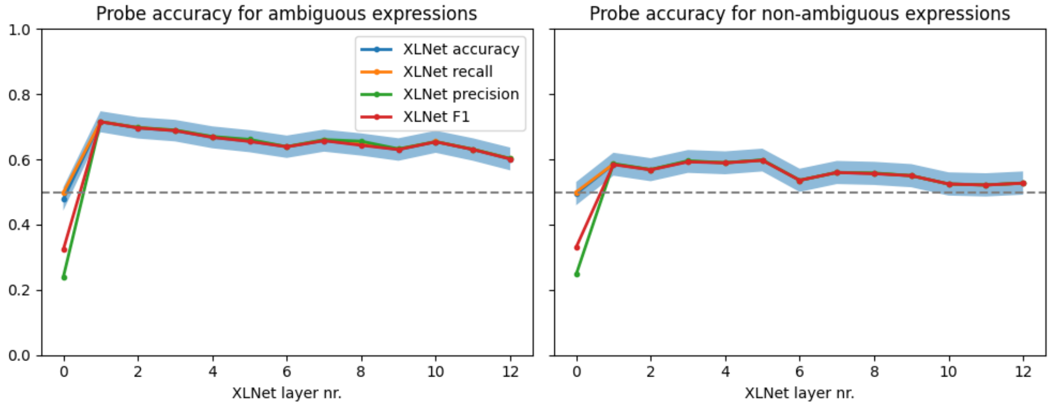

FIGURE 5.5: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of XLNet base for ambiguous and underspecified but not ambiguous sentences with presence or absence of underspecification as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

This difference also arises when considering the difference between ambiguous and non-ambiguous but still underspecified expressions, seemingly suggesting that some forms of ambiguous language are better encoded in the hidden states of the tested model than underspecified language that is not ambiguous

At the same time, however, these findings contrast earlier chapters, in that the fact that the probes for the third domain having a considerably lower accuracy – suggesting that for this domain, underspecification is less encoded within the model's hidden states – seems incompatible with the fact that it is for this domain that the highest accuracy in the recognition task was achieved by the models tested. This may be caused by the fact that this high accuracy was achieved by Flan-T5, which is not tested here – future research might investigate whether the probing accuracy for the third domain increased when tested using Flan-T5.

Chapter 6

General Discussion and Conclusion

6.1 General Discussion

In this thesis, we have investigated semantic underspecification and how different variations of this phenomenon are processed in language models. After introducing the topics in the first chapter, we proposed a new dataset – DUST – that offers examples of semantically underspecified sentences annotated with the domain of underspecification they belong to. We then used this dataset to investigate how language models handle underspecification. In the third chapter, we used an extrinsic method of evaluation to see whether language models are able to recognize underspecification in prompts, and found that some language models are able to do so to some degree. In the fourth chapter, we investigated the correlation between semantic underspecification and language model perplexity, finding that there is often – but not always – a significant difference in perplexity between underspecified sentences and their specified counterparts. Finally, in the fifth chapter, we used diagnostic classifiers to study how underspecification is encoded in language models. Based on our findings, we make the following claims:

Language models can recognize underspecification, yet this ability does not correlate with commonly used metrics for language models

In our thesis, we have shown that underspecification is reflected in commonly used metrics such as perplexity and is encoded in the hidden states of the models, and that some language models are able to recognize or self-correct underspecification to some extent, yet there appears to be no correlation between this encoded knowledge and the degree to which a language model is able to recognize and self-correct underspecification. This can be seen by contrasting the results of chapter 3 with those of chapters 4 and 5. In chapter 4, we can see that the only domain of underspecification where there is often not a significant difference in perplexity is the third domain, and in chapter 5 we can see that this is also the domain where knowledge of underspecification is least encoded in the hidden states of the tested models. Yet it is this domain where the highest accuracy is achieved on the recognition task in chapter 3.

Of course, there is the possibility of our experiments not identifying some influencing factor that is present. This may, for example, be the case because the Flan-T5 model, which achieved this high score, was not tested in our probing experiment. Yet the results of chapter 4.3 do seem to suggest that underspecification may also be correlated with other factors, such as the concreteness of words of word frequency.

Should this be the case, then it is possible that language models learn to 'recognize' underspecification by correlated factors such as these.

Underspecification is heterogeneous in many ways

As we have claimed in our introduction, the terms 'ambiguity' and 'underspecification' are sometimes used seemingly interchangeably, but these phenomena are treated differently by language models. Sentences that contain underspecification but not ambiguity are recognized better as being underspecified by language models, often result in a language model perplexity closer to their specified counterpart, and appear to be encoded less clearly in the hidden states of a language models.

Furthermore, even between sentences that are both examples of ambiguity there may be a difference in how language models treat them. For example, chapter 3 shows that while Flan-T5 xxl is able to recognize forms of syntactic ambiguity such as scopal and phrasal attachment ambiguity at a level significantly above chance, it is unable to do so for lexical ambiguity in the form of homonymy.

These results are relevant for researchers studying ambiguity or underspecification, showing that results for one kind of underspecification do not necessarily generalize to others. This is especially relevant for corpora used to investigate these phenomena, which at the moment often either do not distinguish between different kinds of underspecification or contain only one kind of underspecification.

Finally, our results show that even when considering one kind of underspecification, it may be treated differently by different language models. This can be seen extrinsically in chapter 3, where OPT 13b is able to recognize lexical ambiguity above chance level but Flan-T5 xxl is not, but also intrinsically in chapter 5, where we see that OPT models appear to encode (a feature correlated with) underspecification in later hidden layers, whereas XLNet does this in earlier layers. Hence, it remains relevant for researchers to investigate multiple (kinds of) models when making general statements about language models and underspecification.

Language models are not internally consistent w.r.t. underspecification

Finally, we claim that our findings mirror those of Ortega-Martín et al. (2023) in that they show that language models are not internally consistent w.r.t. underspecification. This can be found in our findings in chapter 3, which show that a language model can recognize a sentence as underspecified yet not chance this sentence when prompted to remove underspecification, but also in those of chapter 5, which suggests that language models encode (some feature correlated with) underspecification in their hidden states for (certain domains of) underspecification, yet do not use this knowledge when prompted to handle underspecification.

6.2 Future Work

Of course, we would much rather see this thesis as part of the first research on underspecification and language models than as the final word on this topic, and we believe there are many ways to build upon this work.

First, and perhaps foremost, is the question of what makes a model decide that an expression is underspecified. Our results seem to suggest a few factors that do *not* play a role in this, yet discovering which feature(s) *do* play a role in this process could be a great step in preventing negative consequences arriving from underspecification and language models.

At the dataset level, there are several ways through which the DUST dataset could be improved. First of all, at present, we were unable to create minimal pairs for the deixis and homonymy subclasses. Creating such minimal pairs for these subclasses would considerably strengthen the results from experiments based on this dataset, especially with regards to claims about the fourth domain of Egg. Similarly, although minimal pairs do exist for the first and second domains, there are relatively few sentences for these domains, and what sentences there are all originate from the same fixed dictionary of words. Even replacing these sentences with sentences based upon a different vocabulary might strengthen the robustness of the dataset against biases based on this restricted dictionary.

Finally, we believe that at the level of linguistics and semantics, much could be achieved by further defining and studying the subclass of 'missing information'. This group was briefly mentioned by Egg (2010), applied to only a small group of sentences, yet we believe that position papers within the field of NLP (such as Hutchinson, Baldridge, and Prabhakaran, 2022 and Pezzelle, 2023) have shown that in reality, this group covers a large amount of the sentences which could be the cause of negative consequences of language models.

6.3 Conclusion

In this thesis, we have introduced the DUST dataset as a dataset containing underspecified sentences annotated with the domain of underspecification the belong to, and used this dataset to perform some preliminary experiments into the way underspecification is handled by language models. Our results show that while some language models are able to recognize semantic underspecification to some extent, they appear to not do so based on commonly used metrics or encoded knowledge, and that a more fine-grained approach to underspecification could greatly benefit the research community.

Appendix A

Logistic regression coefficients for perplexity

TABLE A.1: Regression coefficients of Flan-T5 xxl on all domains of Egg

Variable	Coefficient	Standard Error	z	$P > \|z\|$	[0.0.25	0.095]
Sentence length	0.0173	0.003	6.073	0.0	0.012	0.023
Avg. concreteness	0.0511	0.06	0.846	0.398	-0.067	0.169
Avg. age of acq.	-0.0759	0.02	-3.796	0.0	-0.115	-0.037
Avg. word freq.	0.712	0.334	2.134	0.033	0.058	1.366
Perplexity	0.0006	0.0	3.307	0.001	0.0	0.001

TABLE A.2: Regression coefficients of Flan-T5 xxl on domain 1 of Egg

Variable	Coefficient	Standard Error	z	$P > \|z\|$	[0.0.25	0.095]
Sentence length	-1.5617	0.417	-3.749	0.0	-2.378	-0.745
Avg. concreteness	3.324	1.005	3.306	0.001	1.353	5.295
Avg. age of acq.	0.0548	0.262	0.209	0.834	-0.459	0.568
Avg. word freq.	21.2047	6.678	3.175	0.001	8.117	34.293
Perplexity	-0.0011	0.001	-0.908	0.364	-0.004	0.001

TABLE A.3: Regression coefficients of Flan-T5 xxl on domain 2 of Egg

Variable	Coefficient	Standard Error	z	$P > \|z\|$	[0.0.25	0.095]
Sentence length	-1.7541	0.263	-6.668	0.0	-2.27	-1.238
Avg. concreteness	3.8679	0.656	5.898	0.0	2.582	5.153
Avg. age of acq.	0.0846	0.185	0.456	0.648	-0.279	0.448
Avg. word freq.	21.205	3.926	5.401	0.0	13.51	28.9
Perplexity	-0.0016	0.001	-2.086	0.037	-0.003	-0.0001

Table A.4: Regression coefficients of Flan-T5 xxl on domain 3 of Egg

Variable	Coefficient	Standard Error	z	$P > \lvert z \rvert$	[0.0.25	0.095]
Sentence length	-0.0057	0.004	-1.35	0.177	-0.014	0.003
Avg. concreteness	-0.3478	0.076	-4.573	0.0	-0.497	-0.199
Avg. age of acq.	0.1147	0.025	4.497	0.0	0.065	0.165
Avg. word freq.	1.3824	0.402	3.44	0.001	0.595	2.17
Perplexity	-0.0001	0.0	-0.682	0.495	-0.001	0.0

Table A.5: Regression coefficients of Flan-T5 xxl on domain 4 of Egg

Variable	Coefficient	Standard Error	z	$P > \lvert z \rvert$	[0.0.25	0.095]
Sentence length	0.0525	0.005	10.032	0.0	0.042	0.063
Avg. concreteness	2.4941	0.198	12.607	0.0	2.106	2.882
Avg. age of acq.	-0.957	0.062	-15.331	0.0	-1.079	-0.835
Avg. word freq.	-1.6151	0.799	-2.021	0.043	-3.181	-0.049
Perplexity	0.0065	0.002	4.018	0.0	0.003	0.01

Table A.6: Regression coefficients of Flan-T5 xxl on all domains of Egg when using model prediction as the dependent variable

Variable	Coefficient	Standard Error	z	$P > \lvert z \rvert$	[0.0.25	0.095]
Sentence length	-0.022	0.003	-7.658	0.0	-0.028	-0.016
Avg. concreteness	-0.2411	0.061	-3.978	0.0	-0.36	-0.122
Avg. age of acq.	0.1124	0.02	5.602	0.0	0.073	0.152
Avg. word freq.	1.3696	0.335	4.086	0.0	0.713	2.027
Perplexity	-0.0004	0.0	-2.167	0.03	-0.001	-0.0

Table A.7: Regression coefficients of Flan-T5 xxl on domain 1 of Egg when using model prediction as the dependent variable

Variable	Coefficient	Standard Error	z	$P > \lvert z \rvert$	[0.0.25	0.095]
Sentence length	-0.5946	0.246	-2.413	0.016	-1.078	-0.112
Avg. concreteness	1.2118	0.741	1.636	0.102	-0.24	2.664
Avg. age of acq.	0.0389	0.216	0.18	0.857	-0.384	0.462
Avg. word freq.	8.1099	4.119	1.969	0.049	0.037	16.183
Perplexity	-0.0005	0.001	-0.477	0.633	-0.002	0.001

Table A.8: Regression coefficients of Flan-T5 xxl on domain 2 of Egg when using model prediction as the dependent variable

Variable	Coefficient	Standard Error	z	$P > \lvert z \rvert$	[0.0.25	0.095]
Sentence length	-0.3051	0.124	-2.452	0.014	-0.549	-0.061
Avg. concreteness	0.4121	0.384	1.074	0.283	-0.34	1.164
Avg. age of acq.	0.1241	0.141	0.883	0.377	-0.151	0.4
Avg. word freq.	4.3036	2.124	2.026	0.043	0.141	8.466
Perplexity	-0.0004	0.001	-0.738	0.461	-0.002	0.001

TABLE A.9: Regression coefficients of Flan-T5 xxl on domain 3 of Egg when using model prediction as the dependent variable

Variable	Coefficient	Standard Error	z	$P > \|z\|$	[0.0.25	0.095]
Sentence length	-0.0162	0.004	-3.769	0.0	-0.025	-0.008
Avg. concreteness	-0.3404	0.076	-4.456	0.0	-0.49	-0.191
Avg. age of acq.	0.1427	0.026	5.571	0.0	0.092	0.193
Avg. word freq.	1.0983	0.403	2.728	0.006	0.309	1.887
Perplexity	-0.0005	0.0	-2.4	0.016	-0.001	-0.0001

TABLE A.10: Regression coefficients of Flan-T5 xxl on domain 4 of Egg when using model prediction as the dependent variable

Variable	Coefficient	Standard Error	z	$P > \|z\|$	[0.0.25	0.095]
Sentence length	-0.0294	0.004	-6.559	0.0	-0.038	-0.021
Avg. concreteness	0.0882	0.163	0.54	0.589	-0.232	0.408
Avg. age of acq.	0.0007	0.049	0.013	0.989	-0.096	0.097
Avg. word freq.	2.125	0.733	2.899	0.004	0.689	3.561
Perplexity	0.0028	0.001	2.009	0.045	0.0001	0.006

Appendix B

Further Probing Results

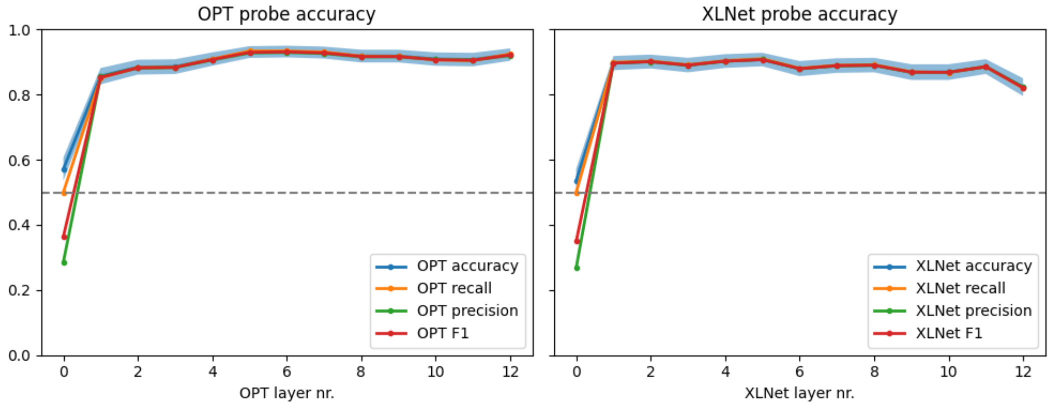

FIGURE B.1: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of OPT 125m and XLNet base for for underspecified sentences with presence or absence of ambiguity as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

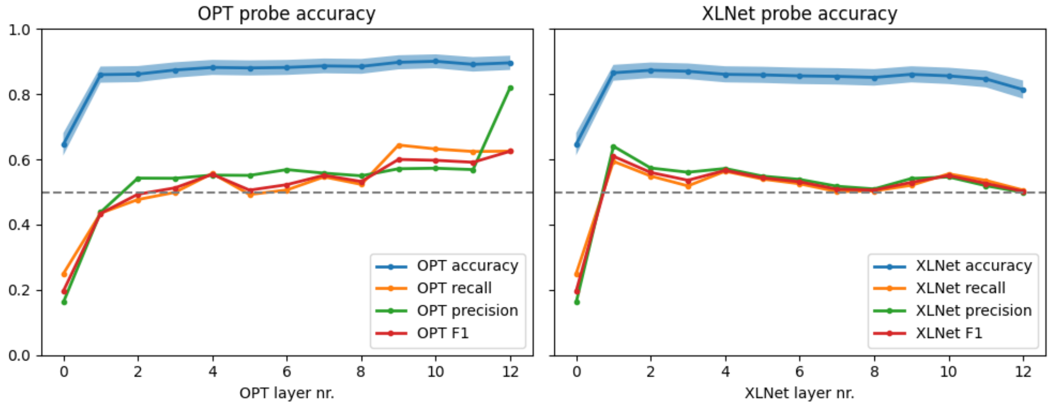

FIGURE B.2: Accuracy, precision, recall and F_1 score per layer of model for probes on hidden layers of OPT 125m and XLNet base for for underspecified sentences with the domain of Egg they belong to as dependent variable. Layer number 0 represents the sentence embeddings. The blue area indicates the 95% confidence interval around the accuracy score.

Bibliography

Abdou, Mostafa et al. (July 2020). "The Sensitivity of Language Models and Humans to Winograd Schema Perturbations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7590–7604. DOI: 10.18653/v1/2020.acl-main.679. URL: https://aclanthology.org/2020.acl-main.679 (visited on 07/31/2023).

Arehalli, Suhas, Brian Dillon, and Tal Linzen (Dec. 2022). "Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities". In: *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 301–313. URL: https://aclanthology.org/2022.conll-1.20 (visited on 06/20/2023).

Belinkov, Yonatan (Apr. 2022). "Probing Classifiers: Promises, Shortcomings, and Advances". In: *Computational Linguistics* 48.1, pp. 207–219. ISSN: 0891-2017. DOI: 10.1162/coli_a_00422. URL: https://doi.org/10.1162/coli_a_00422 (visited on 07/21/2023).

Bellot, Gabrielle (Oct. 2018). *How Le Guin's A Wizard of Earthsea Subverts Racism (But Not Sexism)*. en-US. URL: https://www.tor.com/2018/10/30/how-le-guins-a-wizard-of-earthsea-subverts-racism-but-not-sexism/ (visited on 06/11/2023).

Bender, Emily M. et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://dl.acm.org/doi/10.1145/3442188.3445922 (visited on 07/23/2023).

Bennett, Cynthia L. et al. (May 2021). ""It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, pp. 1–19. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445498. URL: https://doi.org/10.1145/3411764.3445498 (visited on 03/07/2023).

Berzak, Yevgeni et al. (Sept. 2015). "Do You See What I Mean? Visual Resolution of Linguistic Ambiguities". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1477–1487. DOI: 10.18653/v1/D15-1172. URL: https://aclanthology.org/D15-1172 (visited on 03/09/2023).

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a
Abstract.html (visited on 07/22/2023).

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman (Sept. 2014). "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior Research Methods* 46.3, pp. 904–911. ISSN: 1554-3528. DOI: 10.3758/s13428-

013-0403-5. URL: https://doi.org/10.3758/s13428-013-0403-5 (visited on 06/13/2023).

Chen, Xinyun et al. (Apr. 2023). *Teaching Large Language Models to Self-Debug*. arXiv:2304.05128 [cs]. URL: http://arxiv.org/abs/2304.05128 (visited on 06/25/2023).

Chung, Hyung Won et al. (Dec. 2022). *Scaling Instruction-Finetuned Language Models*. en. arXiv:2210.11416 [cs]. URL: http://arxiv.org/abs/2210.11416 (visited on 06/20/2023).

Cunnings, Ian, Georgia Fotiadou, and Ianthi Tsimpli (Dec. 2017). "Anaphora resolution and reanalysis during L2 sentence processing: Evidence from the Visual World Paradigm". In: *Studies in Second Language Acquisition* 39.4. Publisher: Cambridge University Press, pp. 621–652. ISSN: 0272-2631, 1470-1545. DOI: 10.1017/S0272263116000292. URL: https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/abs/anaphora-resolution-and-reanalysis-during-l2-sentence-processing/AF25F70369832F546F8B03B8DF19B562 (visited on 07/04/2023).

Dodge, Jesse et al. (2021). "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus". en. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1286–1305. DOI: 10.18653/v1/2021.emnlp-main.98. URL: https://aclanthology.org/2021.emnlp-main.98 (visited on 07/23/2023).

Durrani, Nadir, Hassan Sajjad, and Fahim Dalvi (Aug. 2021). "How transfer learning impacts linguistic knowledge in deep NLP models?" In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 4947–4957. DOI: 10.18653/v1/2021.findings-acl.438. URL: https://aclanthology.org/2021.findings-acl.438 (visited on 07/21/2023).

Dwivedi, Veena D., Matthew Goldhawk, and Frédéric Mailhot (2009). "Semantic underspecification and anaphora". In: *Proceedings of the 2009 annual conference of the Canadian Linguistic Association*.

Egg, Markus (2010). "Semantic Underspecification". In: *Language and Linguistics Compass* 4.3, pp. 166–181. ISSN: 1749-818X. DOI: 10.1111/j.1749-818X.2010.00188.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2010.00188.x (visited on 02/08/2023).

Ettinger, Allyson (Jan. 2020). "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00298. URL: https://doi.org/10.1162/tacl_a_00298 (visited on 07/21/2023).

Fantechi, Alessandro, Stefania Gnesi, and Laura Semini (2023). "Rule-based NLP vs ChatGPT in Ambiguity Detection, a Preliminary Study". In: *Joint Proceedings of REFSQ-2023 Workshops*. Barcelona.

Franzon, Francesca and Chiara Zanini (May 2022). "The Entropy of Morphological Systems in Natural Languages Is Modulated by Functional and Semantic Properties". In: *Journal of Quantitative Linguistics* 30.1, pp. 42–66. ISSN: 0929-6174. DOI: 10.1080/09296174.2022.2063501. URL: https://doi.org/10.1080/09296174.2022.2063501 (visited on 02/09/2023).

Frisson, Steven (2009). "Semantic Underspecification in Language Processing". In: *Language and Linguistics Compass* 3.1, pp. 111–127. ISSN: 1749-818X. DOI: 10.1111/j.1749-818X.2008.00104.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2008.00104.x (visited on 02/13/2023).

Ganguli, Deep et al. (Feb. 2023). *The Capacity for Moral Self-Correction in Large Language Models*. arXiv:2302.07459 [cs]. URL: http://arxiv.org/abs/2302.07459 (visited on 06/25/2023).

Giulianelli, Mario et al. (Nov. 2018). "Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 240–248. DOI: 10.18653/v1/W18-5426. URL: https://aclanthology.org/W18-5426 (visited on 07/21/2023).

Gonen, Hila et al. (Dec. 2022). *Demystifying Prompts in Language Models via Perplexity Estimation*. arXiv:2212.04037 [cs]. DOI: 10.48550/arXiv.2212.04037. URL: http://arxiv.org/abs/2212.04037 (visited on 07/23/2023).

Goodkind, Adam and Klinton Bicknell (Jan. 2018). "Predictive power of word surprisal for reading times is a linear function of language model quality". In: *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*. Salt Lake City, Utah: Association for Computational Linguistics, pp. 10–18. DOI: 10.18653/v1/W18-0102. URL: https://aclanthology.org/W18-0102 (visited on 06/20/2023).

Gou, Zhibin et al. (May 2023). *CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing*. arXiv:2305.11738 [cs]. URL: http://arxiv.org/abs/2305.11738 (visited on 06/25/2023).

Harris, Daniel W. (2020). "What Makes Human Communication Special?" In: *Unpublished book manuscript*. Draft of October 27, 2020. CUNY Graduate Center.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts (Feb. 2017). "Prediction and explanation in social systems". In: *Science* 355.6324. Publisher: American Association for the Advancement of Science, pp. 486–488. DOI: 10.1126/science.aal3856. URL: https://www.science.org/doi/10.1126/science.aal3856 (visited on 07/04/2023).

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema (Apr. 2018). "Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure". In: *Journal of Artificial Intelligence Research* 61, pp. 907–926. ISSN: 1076-9757. DOI: 10.1613/jair.1.11196. URL: https://www.jair.org/index.php/jair/article/view/11196 (visited on 07/21/2023).

Hutchinson, Ben, Jason Baldridge, and Vinodkumar Prabhakaran (Nov. 2022). "Underspecification in Scene Description-to-Depiction Tasks". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, pp. 1172–1184. URL: https://aclanthology.org/2022.aacl-main.86 (visited on 02/20/2023).

Jelinek, F. et al. (1997). "Perplexity—a measure of the difficulty of speech recognition tasks". In: *The Journal of the Acoustical Society of America* 62.S1, S63. ISSN: 0001-4966. DOI: 10.1121/1.2016299. URL: https://doi.org/10.1121/1.2016299 (visited on 06/20/2023).

Jurafsky, Daniel and James H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Kasneci, Enkelejda et al. (2023). "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences* 103, p. 102274. ISSN: 1041-6080. DOI: https://doi.org/10.1016/j.lindif.

2023.102274. URL: https://www.sciencedirect.com/science/article/pii/S1041608023000195.

Khurana, Diksha et al. (Jan. 2023). “Natural language processing: state of the art, current trends and challenges”. In: *Multimedia Tools and Applications* 82.3, pp. 3713–3744. ISSN: 1573-7721. DOI: 10.1007/s11042-022-13428-4. URL: https://doi.org/10.1007/s11042-022-13428-4 (visited on 06/07/2023).

Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert (Dec. 2012). “Age-of-acquisition ratings for 30,000 English words”. In: *Behavior Research Methods* 44.4, pp. 978–990. ISSN: 1554-3528. DOI: 10.3758/s13428-012-0210-4. URL: https://doi.org/10.3758/s13428-012-0210-4 (visited on 06/13/2023).

Kuribayashi, Tatsuki et al. (Aug. 2021). “Lower Perplexity is Not Always Human-Like”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5203–5217. DOI: 10.18653/v1/2021.acl-long.405. URL: https://aclanthology.org/2021.acl-long.405 (visited on 06/20/2023).

Lappin, Shalom (2000). “An Intensional Parametric Semantics for Vague Quantifiers”. In: *Linguistics and Philosophy* 23.6, pp. 599–620. ISSN: 0165-0157. URL: https://www.jstor.org/stable/25001796 (visited on 02/08/2023).

Lee, Mina, Percy Liang, and Qian Yang (Apr. 2022). “CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New York, NY, USA: Association for Computing Machinery, pp. 1–19. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3502030. URL: https://dl.acm.org/doi/10.1145/3491102.3502030 (visited on 06/07/2023).

Levesque, Hector, Ernest Davis, and Leora Morgenstern (2012). “The Winograd Schema Challenge”. In: *Proceeding of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. Institute of Electrical and Electronics Engineers Inc., pp. 552–561. ISBN: 978-1-57735-560-1.

Levinson, Stephen C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. OCLC: 956673720. The MIT Press. ISBN: 978-0-262-27825-6.

Lin, Yongjie, Yi Chern Tan, and Robert Frank (Aug. 2019). “Open Sesame: Getting inside BERT’s Linguistic Knowledge”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 241–253. DOI: 10.18653/v1/W19-4825. URL: https://aclanthology.org/W19-4825 (visited on 07/21/2023).

Liu, Alisa et al. (Apr. 2023). *We’re Afraid Language Models Aren’t Modeling Ambiguity*. arXiv:2304.14399 [cs]. URL: http://arxiv.org/abs/2304.14399 (visited on 06/07/2023).

Liu, Pengfei et al. (Jan. 2023). “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”. In: *ACM Computing Surveys* 55.9, 195:1–195:35. ISSN: 0360-0300. DOI: 10.1145/3560815. URL: https://dl.acm.org/doi/10.1145/3560815 (visited on 06/25/2023).

Lo, Adeline et al. (Nov. 2015). “Why significant variables aren’t automatically good predictors”. In: *Proceedings of the National Academy of Sciences* 112.45. Publisher: Proceedings of the National Academy of Sciences, pp. 13892–13897. DOI: 10.1073/pnas.1518285112. URL: https://www.pnas.org/doi/10.1073/pnas.1518285112 (visited on 07/04/2023).

Maciejewski, Greg and Ekaterini Klepousniotou (Dec. 2016). “Relative Meaning Frequencies for 100 Homonyms: British eDom Norms”. In: *Journal of Open Psychology Data* 4, e6. ISSN: 2050-9863. DOI: 10.5334/jopd.28. URL: http://openpsychologydata.metajnl.com/articles/10.5334/jopd.28/ (visited on 06/07/2023).

Meister, Clara and Ryan Cotterell (Aug. 2021). “Language Model Evaluation Beyond Perplexity”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5328–5339. DOI: 10.18653/v1/2021.acl-long.414. URL: https://aclanthology.org/2021.acl-long.414 (visited on 07/23/2023).

Miaschi, Alessio et al. (June 2021). “What Makes My Model Perplexed? A Linguistic Investigation on Neural Language Models Perplexity”. In: *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Online: Association for Computational Linguistics, pp. 40–47. DOI: 10.18653/v1/2021.deelio-1.5. URL: https://aclanthology.org/2021.deelio-1.5 (visited on 07/23/2023).

Niehren, Joachim, Manfred Pinkal, and Peter Ruhrberg (July 1997). “A Uniform Approach to Underspecification and Parallelism”. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, pp. 410–417. DOI: 10.3115/976909.979670. URL: https://aclanthology.org/P97-1053 (visited on 07/31/2023).

OpenAI (2023). *Introducing ChatGPT*. en-US. URL: https://openai.com/blog/chatgpt (visited on 06/07/2023).

Ortega-Martín, Miguel et al. (Feb. 2023). *Linguistic ambiguity analysis in ChatGPT*. arXiv:2302.06426 [cs]. URL: http://arxiv.org/abs/2302.06426 (visited on 06/12/2023).

Pezzelle, Sandro (July 2023). “Dealing with Semantic Underspecification in Multimodal NLP”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 12098–12112. URL: https://aclanthology.org/2023.acl-long.675 (visited on 07/31/2023).

Piantadosi, Steven T., Harry Tily, and Edward Gibson (Mar. 2012). “The communicative function of ambiguity in language”. In: *Cognition* 122.3, pp. 280–291. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2011.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0010027711002496 (visited on 02/08/2023).

Pinkal, Manfred (1999). “On Semantic Underspecification”. In: *Computing Meaning: Volume 1*. Ed. by Harry Bunt and Reinhard Muskens. Studies in Linguistics and Philosophy. Dordrecht: Springer Netherlands, pp. 33–55. ISBN: 978-94-011-4231-1. DOI: 10.1007/978-94-011-4231-1_2. URL: https://doi.org/10.1007/978-94-011-4231-1_2 (visited on 06/11/2023).

Poesio, Massimo (Dec. 1994). “Ambiguity, Underspecification and Discourse Interpretation”. In: *Proceedings of the First International Workshop on Computational Semantics*.

Radford, Alec et al. (2019). “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8, p. 9.

Rayner, Keith and Susan A. Duffy (May 1986). “Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity”. In: *Memory & Cognition* 14.3, pp. 191–201. ISSN: 1532-5946. DOI: 10.

3758/BF03197692. URL: https://doi.org/10.3758/BF03197692 (visited on 06/20/2023).

Roth, Michael, Talita Anthonio, and Anna Sauer (July 2022). "SemEval-2022 Task 7: Identifying Plausible Clarifications of Implicit and Underspecified Phrases in Instructional Texts". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, pp. 1039–1049. DOI: 10.18653/v1/2022.semeval-1.146. URL: https://aclanthology.org/2022.semeval-1.146 (visited on 02/14/2023).

Rudinger, Rachel et al. (June 2018). "Gender Bias in Coreference Resolution". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 8–14. DOI: 10.18653/v1/N18-2002. URL: https://aclanthology.org/N18-2002 (visited on 06/07/2023).

Sennet, Adam (2023). "Ambiguity". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2023/entries/ambiguity/ (visited on 02/22/2023).

Shin, Seongjin et al. (July 2022). "On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 5168–5186. DOI: 10.18653/v1/2022.naacl-main.380. URL: https://aclanthology.org/2022.naacl-main.380 (visited on 07/23/2023).

Swets, Benjamin et al. (Jan. 2008). "Underspecification of syntactic ambiguities: Evidence from self-paced reading". en. In: *Memory & Cognition* 36.1, pp. 201–216. ISSN: 1532-5946. DOI: 10.3758/MC.36.1.201. URL: https://doi.org/10.3758/MC.36.1.201 (visited on 06/20/2023).

Toral, Antonio (Aug. 2013). "Hybrid Selection of Language Model Training Data Using Linguistic Information and Perplexity". In: *Proceedings of the Second Workshop on Hybrid Approaches to Translation*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 8–12. URL: https://aclanthology.org/W13-2803 (visited on 07/23/2023).

Van Schijndel, Marten and Tal Linzen (2021). "Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty". en. In: *Cognitive Science* 45.6, e12988. ISSN: 1551-6709. DOI: 10.1111/cogs.12988. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12988 (visited on 04/20/2023).

Wang, Jiaqi et al. (Apr. 2023). *Prompt Engineering for Healthcare: Methodologies and Applications*. arXiv:2304.14670 [cs]. URL: http://arxiv.org/abs/2304.14670 (visited on 06/25/2023).

Wasow, Thomas, Amy Perfors, and David Beaver (2005). "The Puzzle of Ambiguity". In: *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*.

Wilcox, Ethan Gotlieb et al. (June 2020). "On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior". In: *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pp. 1707–1713. (Visited on 06/20/2023).

Yang, Zhilin et al. (2019). "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (visited on 06/20/2023).

Zaib, Munazza, Quan Z. Sheng, and Wei Emma Zhang (Feb. 2020). “A Short Survey of Pre-trained Language Models for Conversational AI-A New Age in NLP”. In: *Proceedings of the Australasian Computer Science Week Multiconference*. ACSW ’20. New York, NY, USA: Association for Computing Machinery, pp. 1–4. ISBN: 978-1-4503-7697-6. DOI: 10.1145/3373017.3373028. URL: https://doi.org/10.1145/3373017.3373028 (visited on 07/23/2023).

Zhang, Susan et al. (May 2022). *OPT: Open Pre-trained Transformer Language Models*. en. URL: https://arxiv.org/abs/2205.01068v4 (visited on 06/20/2023).

Zhu, Wanzheng and Suma Bhat (Nov. 2020). “GRUEN for Evaluating Linguistic Quality of Generated Text”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 94–108. DOI: 10.18653/v1/2020.findings-emnlp.9. URL: https://aclanthology.org/2020.findings-emnlp.9 (visited on 07/31/2023).

Zwicky, Arnold M. and Jerrold M. Sadock (Jan. 1975). “Ambiguity tests and how to fail them”. In: *Syntax and Semantics* 4, pp. 1–36.