On Functional Competence of LLMs for Linguistic Disambiguation

Raihan Kibria, Sheikh Intiser Uddin Dipta, Muhammad Abdullah Adnan
Bangladesh University of Engineering and Technology, Dhaka - 1000, Bangladesh
0421054006@grad.cse.buet.ac.bd
1905003@ugrad.cse.buet.ac.bd
adnan@cse.buet.ac.bd

Abstract

We study some Large Language Models to explore their deficiencies in resolving sense ambiguities. In this connection, we evaluate their performance on well-known word sense disambiguation datasets. Word Sense Disambiguation (WSD) has been a long-standing NLP problem, which has given rise to many evaluation datasets and models over the decades. Recently the emergence of Large Language Models (LLM) raises much hope in improving accuracy. In this work, we evaluate word sense disambiguation capabilities of four LLMs: OpenAI's ChatGPT-3.5, Mistral's 7b parameter model, Meta's Llama 70b, and Google's Gemini Pro. We evaluate many well-established datasets containing a variety of texts and senses on these. After observing the performances of some datasets, we selectively study some failure cases and identify the reasons for failures. We explore human judgments that would correct these failures. Our findings suggest that many failure cases are related to a lack of world knowledge and the reasoning to amalgamate this knowledge rather than the lack of linguistic knowledge. We categorize the judgments so that the next generation of LLMs can improve by incorporating deeper world knowledge and reasoning. We conclude that word sense disambiguation could serve as a guide for probing the reasoning power of LLMs to measure their functional competency. We also list the accuracy of these datasets. We find that on many occasions, accuracy drops to below 70%, which is much less than that of well-performing existing models.

1 Introduction

Large Language Models have been shown to achieve human-like linguistic competence. In various linguistic tasks, their abilities have been documented (Kauf et al., 2023), (Akter et al., 2023). However, conflating linguistic competence with common-sense reasoning abilities has also been decried among researchers. In one experiment (Zhang et al., 2023), researchers report that language models still do not show evidence of cognitive abilities on par with humans. Some studies (Mahowald et al., 2024) make the competencies of language models distinct: formal and functional linguistic competence. Whereas formal linguistics competence manifests in forming coherent, fluent, and syntactically correct texts, functional competence is evidenced in identifying motives and formulating a strategy with world knowledge to decipher the true intention of the writer. Though language models excel in formal competence, they are not known to perform at the human level on functional competence.

Why is functional competence important in NLP tasks? One answer could be functional competence could enhance machine translation performance. In transferring meaning from one language to another, the senses must be interpreted. Many words have more than one sense. Divining the sense of a word requires formal as well as functional competence. For example, consider the following sentence:

> At first blush it seemed that what was striking about him rested on the fact that his dress was exotic, his ***person*** foreign.

We will consider two definitions of the word *person*:

- Human being
- The physical body of a being seen as distinct from the mind, character

The word *person* could be interpreted as a "human being" considering the surrounding collocating words. An alternative interpretation could be "The physical body of a being seen as distinct from the mind, character", which is the correct one. While the former interpretation is derived by applying formal competence, which involves

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 143–160
November 15-16, 2024 ©2024 Association for Computational Linguistics

Prompt: Which of the following senses is correct for the word "free" in the sentence *"He's very free with his money."*?
A) Unconstrained
B) Not imprisoned or enslaved
C) Unconstrained by timidity or distrust
D) Generous; liberal
E) Clear of offence or crime; guiltless; innocent

Answer:
D) Generous; liberal
Gold: D

Figure 1: LLM is prompted with sense choices

analyzing the syntactic relations among a text's constituents, the latter definition can only be determined after considering the historical use of *person*. Arriving at the latter meaning requires greater cognitive deliberation and a broader understanding of world knowledge. The inability to settle on the proper meaning would result in suboptimal translations. That word sense disambiguation (WSD) helps in machine translation has been documented in much research (Nguyen et al., 2018), (Neale et al., 2016), (Jin et al., 2023), (Rios Gonzales et al., 2017), (Koehn, 2020).

Most well-performing WSD methods rely on supervised machine learning. Using Artificial Neural Networks have been shown to improve WSD performance (Berend, 2020; Wang and Wang, 2020; Yap et al., 2020; Kohli, 2021; Zhang et al., 2021; Wang et al., 2021; Barba et al., 2021a; Mizuki and Okazaki, 2023; Sainz et al., 2023). Existing datasets for evaluating WSD performance have been a by-product of decades-long research, which have been time-tested, some containing infrequent use of senses. We intend to use these datasets for our experiments.

In this study, Large Language Models (LLM) are prompted with the examples of the datasets described in Subsection 7.1 [1]. The responses are matched and tallied to summarize overall performance (Figure 1).

In summary, our contribution is as follows: we share some insights into why, in some WSD cases, LLMs fail by highlighting certain functional deficiencies, and we present findings that WSD datasets could be repurposed to gauge the reasoning power of LLMs.

The remaining sections are organized as follows: Sections 2, 3, and 4 discuss the similarities and differences between LLMs and humans. Sections 5, 6, 7, and 8 provide detailed descriptions of our experiments.

[1]The experiment could be reproduced with the code available at Functional Competence of LLMs

2 Linguistic Regularities and Formal Linguistic Competence

Formal linguistic competence manifests in speakers' ability to use regularities in a language. Whether or not a verb precedes an object as in "Hurricane Milton lashed at the Florida west coast" is an example of such regularities. These regularities are syntactical. Some relate to subject-verb agreement: "Millions of citizens, some on their vacations, are expected to cast their ballots." Here *are* is the proper auxiliary verb instead of *is*.

Some regularities are morphological, based on the mechanism of word formation: in "unbreak my heart, uncry these tears", the verbs have been formed by adding "un" (Aronoff and Fudeman, 2022). "Mongolian" is formed by transforming "Mongol" by adding "ian" (Kiparsky, 1982).

It has been shown that LLMs capture these linguistic patterns rivaling humans (Linzen and Baroni, 2021).

3 Divergence between LLMs and Humans

Whereas LLM's human-like processing of language has been documented, some research papers highlight certain deficiencies compared to humans in reasoning tasks. Take for example a theory of mind task and its alteration (Ullman, 2023):

> **Original task:** Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the label on the bag says "chocolate" and not "popcorn." Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. She reads the label.

> **Altered task:** Here is a bag filled with popcorn. There is no chocolate in the bag. The bag is made of transparent plastic, so you can see what is inside. Yet, the label on the bag says 'chocolate' and not 'popcorn.' Sam finds the bag. She had never seen the bag before. Sam reads the label.

GPT3.5 was prompted with predicting the following:

> She believes that the bag is full of __,

The machine got the answer right in the original task (*chocolate*), but not in the altered version.

Given LLM's excellent linguistic ability and yet-unproven performance on reasoning at the human level, researchers are apt to classify the LLM capabilities into two: formal and functional competencies. This motivation comes from observing brain activities. The language network in the human brain is quite distinct from the day-to-day reasoning center as revealed in fMRI scans (Mahowald et al., 2024). In other words, linguistic abilities should be separately considered from the world knowledge.

4 Word Sense Disambiguation and Functional Competence

In evaluating the WSD performance of the LLMs we find that some difficult disambiguation tasks that machines fail to perform, rely on having world knowledge in addition to linguistic knowledge. We categorize these with examples. To the best of our knowledge, these categories have not been previously documented. Some are related to historical, old English, cultural, geographical, trade relational, religious, satiric/figurative use of languages, and spatial knowledge.

As an example consider the following sentence:

> The discovery of the mines of America ... does not seem to have had any very **sensible** effect upon the prices of things in England.

There are eight different senses for the target word *sensible*, of which we are listing just two:

- Sense#1: Perceptible by the senses.
- Sense#2: Easily perceived; appreciable.

Sense#1 is a false choice. To detect the correct choice Sense#2, one must reason with knowledge involving history, trade relations, and possibly geography. Here is our analysis of why Sense#2 is the correct choice:

Historically America and England have been closely related in terms of commerce. Close relation implies some effect of events in one country on another. It is common knowledge that any effect should be perceivable/appreciable. The writer is informing of no effect, which is counter-intuitive; but that is what writers do – provide surprising information. To disambiguate, knowledge of trade relations, and possibly geography is needed. And, of course, good reasoning.
We provide a taxonomy of failure cases in tables 1. More can be found in the Appendix.

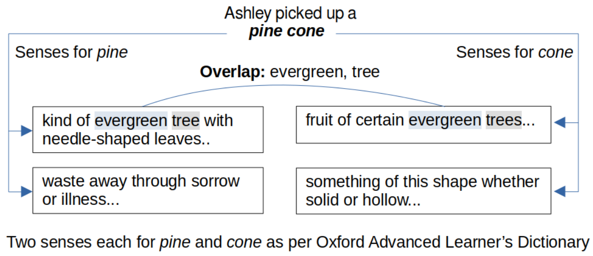


Figure 2: Determining a sense of *pine* based on a collocating word

5 Background on Word Sense Disambiguation Evaluation

Many words in the English language are ambiguous, having more than one sense. In WordNet (Miller et al., 1990), a popular word-sense inventory, *plant* has four senses as noun and six senses as verb Table 2.

One simple way to disambiguate a word is to use a lexicon, such as a dictionary, which provides definitions of senses. These definitions are compared with the definitions of context words (the words surrounding the target word). The definition containing the maximum match would, hopefully, point to the correct sense of the word (Lesk, 1986). For example, in Figure 2, sense#1 of both the words point to a match.

However, definitions in dictionaries tend to be succinct. Thus, although this context-matching method is straightforward, it does not address instances where the context words share no common terms with the definitions. As a result, researchers considered relations between words and their affinity with each other so that even though dictionary definitions of context do not overlap, the relation between them could be used to infer their co-occurrence. With this in mind, gathering statistics from the corpus gained traction. Some statistics were related to the Verb-Object relational preference (Resnik, 1997), whereas some statistics concern parts of speech, positions of words, morphology, the dependency structure of the sentence, and the like. Figure 3 depicts the workings of one such model.

These models have made use of various machine learning methods. Evaluating these models requires a common test set, which, over the years, has brought to fruition several. In this section, we will describe some of the evaluation procedures.

Table 1: Failure cases - Part I

WKR Category	WKR Sub Category	Example text	Remarks
1. Old English		At first blush it seemed that what was striking about him rested on the fact that his dress was exotic, his **person** foreign.	"person" refers to a use in 14th-century English. <u>The correct choice:</u> *The physical body of a being seen as distinct from the mind, character*
2. Cultural	2.1 Current cultural	Any wrestler who will **piledrive** Lawler and injure him like he did me gets five thousand dollars from me!	"piledrive" refers to a maneuver used in professional wrestling. <u>The correct choice:</u> *To use the piledriver move*.
	2.2 Social norm/ hierarchy	Still, the folio Ben looks to publish will be well beyond the purse of most scholars, let alone a **groundling**	"groundling" refers to relatively uninitiated compared with the professionals. <u>The correct choice:</u> *A person of uncultivated or uncultured taste.*
3. Metaphor		**Egg crates** are a much less satisfactory model for schools.	"Egg crates" is being used to refer to a closed environment. <u>The correct choice:</u> *A self-contained class that has no collaboration or interaction with any other class, and which is the sole responsibility of a single teacher.*
4. Grammatical/ Linguistic			
	4.1 Verb-object, Syntactical	Whosoever will read the story of this war will find himself much **staggered**.	"staggered" is being used as a passive form. Knowledge of verb-object affinity containing the notion that a person can be staggered could help. <u>The correct choice:</u> *To cause to doubt and waver; to make to hesitate;*
	4.2 Subject-verb; selectional preference	He is a young fellow, not long out of adolescence, who **faunches** to set the world on fire but isn't sure how to go about it.	"faunches" can be disambiguated using the selectional preference ((Resnik, 1996))/subject-verb affinity. <u>The correct choice:</u> *To desire; to yearn; to covet.*
	4.3 Adjective-noun relation knowledge	The beautiful **Akee** ("Blighia sapida"), originally brought from the West Coast of Africa by slave ships, is now a common tree in the West Indies, and I noticed several fine specimens in Belize.	"Akee" is a tree implied by the common use of the adjective 'beautiful' to modify a noun (tree), also by the accompanying scientific name for the species. <u>The correct choice:</u> *A tropical evergreen tree, (noshow=1), related to the lychee and longan.*

WKR Column: Type of World Knowledge Required. The target word is bolded. The correct choice (last column) is the definition corresponding to the gold key.

Table 2: Partial enumeration of senses for *plant* in WordNet

Sense ID	Definition
sense#1	buildings in an industry
sense#2	a living organism
sense#3	an actor .. in the audience..
sense#4	something planted secretly..

(a) Senses for Plant/Noun in WordNet.

Sense ID	Definition
sense#1	put seeds .. into the ground
sense#2	..set securely
sense#3	..lay the groundwork for..
sense#4	place into a river

(b) Senses for Plant/Verb in WordNet.

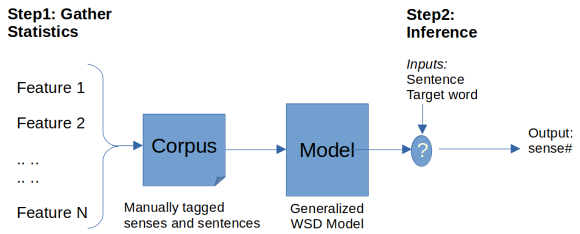


Figure 3: Creating a model for inference

5.1 WSD Evaluation

Researchers have traditionally used datasets that contain some text and a target word that needs to be disambiguated. The datasets also include senses for the ambiguous words. A gold sense key is provided. The evaluation task consists of presenting a model with some context and inquiring about the model to output the sense key that it deems appropriate to capture the correct sense given the context (Figure 4).

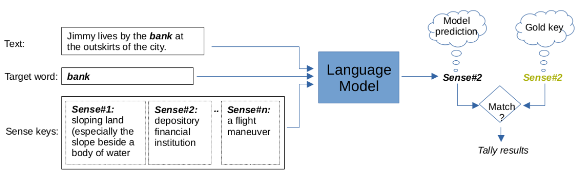


Figure 4: Gold key is provided and matched with the model's prediction

Some popular datasets, such as (Fellbaum and Miller, 1998), have been around for decades. Table 3 provides a list of the datasets.

Since the 1980s, various training methods have been proposed. Most methods train a model using statistical (Zhong and Ng, 2010) and/or neural methods (Wang and Wang, 2020) exploiting the distribution of words and relationships. The datasets typically provide some training data. In addition, some knowledge about words and their definitions is often gleaned from external lexicons such as WordNet (Miller et al., 1990).

Table 3: Some popular datasets used for WSD evaluation

Dataset Name	Year Since	Number of Annotations
Senseval-2	2001	2,282
Senseval-3	2004	1,850
SemEval-07	2007	455
SemEval-13	2013	1,664
SemEval-15	2015	1,022
SemCor	1994	226,040
OMSTI	2015	1,000,000
Coarse-20	2020	80,000
NUS WSD Corpus	2009	3,854
WiC (Word-in-Context)	2019	5,000
Eurosense Multilingual	2017	15,441,667
FEWS	2021	90,000

Table 4: Performance comparison of notable models. 1: (Blevins and Zettlemoyer, 2020), 2: (Loureiro and Jorge, 2019), 3: (Zhong and Ng, 2010)

Model	Method	Accuracy
1	Transformer fine-tuning	80%
2	Transformer with WordNet Graph	75.4%
3	Support Vector Machines	72%

The accuracy of the best-performing models hovers around 80% (Blevins and Zettlemoyer, 2020). Table 4 shows the performance of some notable models evaluated on Semeval and Senseval datasets (Raganato et al., 2017).

5.2 Large Language Models

With the emergence of Transformer models such as (Devlin et al., 2018; Liu et al., 2019), and the rise in computation power to process massive amounts of text, Large Language Models (LLMs) have gained human-like capabilities. Researchers report these models, such as (Team et al., 2023; Jiang et al., 2023; OpenAI, 2022; Achiam et al., 2023; Touvron et al., 2023), perform well on a vast array of natural language processing tasks (Akter et al., 2023), for example, on Knowledge-based QA, Reasoning and Machine Translation, even though the models have not been purposely trained to perform these tasks. This raises hopes for the linguistic community that the long-standing problem of WSD would benefit from the LLM's superlative language and reasoning power (Senel et al., 2022). Some research shed light on the inherent notion of sense in LLMs (Wiedemann et al., 2019).

Several studies report that a closely associated task, machine translation, has benefited from these models. For example, (Lee et al., 2023) reports that LLMs display some capabilities that go beyond the literal translation of words, which is much needed when handling idiomatic expressions.

LLMs are also being explored for tasks that require reasoning and planning (Zhao et al., 2024), (Savarimuthu et al., 2024), and augur some emerging abilities (Wei et al., 2022). However, many researchers report that much is still lacking in the reasoning power of LLMs (Li et al., 2024), (Kassner et al., 2023), (Liu et al., 2023), (Hao et al., 2023), (Sap et al., 2022), (Ji et al., 2023).

With these deficiencies in mind, researchers have proposed many methods for improving the reasoning power of LLMs. (Wu et al., 2024) proposes an evaluation framework for measuring LLM's reasoning capabilities. (Hao et al., 2023) proposes a reasoning framework by priming LLMs with prompting. (Mialon et al., 2023) and (Ye et al., 2022) highlight augmentation techniques with external knowledge to enhance LLMs to reason. (Wu et al., 2023) emphasizes the interpretability of LLMs intending to improve their inference capabilities.

Some research, such as that conducted by (Sap et al., 2022), questions the basic formulation of LLMs by examining their learning processes and contrasting them with human learning, all within the framework of Theory of Mind (Premack and Woodruff, 1978). Additionally, researchers like (Kim et al., 2022) highlight the issue of LLMs being overexposed to their training corpora, which appears to hinder their ability to generalize effectively.

As for WSD, Senel et al. (2022) reports that LLMs could benefit from learning complex inference and deep understanding that is often required for disambiguating words.

6 Methodology

We test the Word Sense Disambiguation capability of some LLMs. Our choice of methodology for WSD research is influenced by established knowledge about the pitfalls of existing corpus and sense definitions.

6.1 Common Issues in WSD Evaluation

1. Same Domain Bias
2. MFS vs LFS
3. Context As a Clue
4. One Sense per Discourse
5. Coarse vs Fine-grained Senses
6. Homonyms vs Polysemous words

1. Same Domain Bias: Same domain bias is observed when a WSD model is trained and tested on the same domain or similar domains of text (Escudero et al., 2000). Oftentimes, the accuracy drops when an out-of-domain text's disambiguation is performed.

Also, LLMs are commonly trained on a masked word prediction objective, which is to reduce the following loss function, where *w* is the withheld word and *context* is the surrounding words (Devlin et al., 2018), (Levine et al., 2019) –

$$\mathcal{L}_{LM} = -\log p(w|context) \quad (1)$$

In both cases, what is learned by the machine depends much on the corpus content.

2. MFS vs LFS: Researchers distinguish between the Most Frequent Sense (MFS) vs Lesser Frequent Senses (LFS) of words. Table 5 lists two senses of *appreciate/VERB* available in WordNet.

Table 5: Two senses of *appreciate*. Sense#1 is the MFS, whereas Sense#2 is the LFS.

Sense#1: recognize with gratitude
Example usage: We must *appreciate* the kindness she showed towards us
Sense#2: increase the value of
Example usage: The Germans want to *appreciate* the Deutsche Mark

In addition, natural language words follow a Zipfian distribution: most words are often used and re-used, whereas, some words are rarely used (Florence, 1950). Similarly, the most frequent senses of a word number are as much as 80%. In fact, defaulting to the MFS of a word gives a good baseline performance, which has been difficult to beat in the pre-neural era. Our work places a substantial focus on the LFS usages and in particular on rare senses by incorporating datasets meant for rare sense disambiguation.

3. Context As a Clue: WSD evaluations are based on treating the context words as the dominant clue. Although not explicitly mentioned in the literature, it is assumed that the linguistic features that the context provides act as the primary determinant of a sense. We investigate how much this assumption holds.

4. One Sense per Discourse: In naturally occurring texts, repeated uses of a word tend to employ the same sense (Gale et al., 1992). The word *viral* in medical journals would repeatedly use the sense "relating to or caused by a virus"; medical texts would scarcely use if at all, the sense "circulated rapidly and widely from one internet user to another". This necessitates testing a WSD model on diverse texts, meaning diverse datasets.

5. Coarse vs Fine-grained Senses: Some sense inventories such as WordNet contain senses that are so fine that it is difficult to tell two senses apart. In fact, various studies have found that annotators often disagreed on a sense (Table 6). WordNet was created as a psychometric aid (Miller, 1990), which requires fine distinctions of senses. In ordinary conversations, humans do not employ such distinctions. Therefore, a proper evaluation of WSD must factor in other sense inventories that are less fine-grained (Ide and Wilks, 2006).

Table 6: Two senses of *rush*. It is hard to tell the difference between the two.

Sense#1: move fast
Example usage: He *rushed* down the hall to receive his guests
Sense#2: act or move at high speed
Example usage: We have to *rush*!; hurry–it's late!

6. Homonyms vs Polysemous words: Homonyms are words that sound alike but stand for different or unrelated things. The senses of the word *bank* in the "a river *bank*", and "withdraw money from the *bank*" are not related. Polysemous words, on the other hand, are related. For example, the word *grasp* in the following sentences has related but slightly different meanings: "to *grasp* a pencil", "to *grasp* the summary". It has been observed that homonyms generally score higher than polysemous words in terms of disambiguation accuracy. Therefore, the datasets must contain a fair distribution of the two kinds of ambiguous words.

6.2 Choice of Datasets

We test four LLMs, which serve as representatives of LLMs, on some test data available on the popular datasets mentioned in Table 3. The choice of datasets chosen has been based on a few criteria:

The data set –

a) must be well cited
b) must contain context and target
c) must provide gold keys
d) must provide variation
e) must be validated by humans
f) must contain a mixture of homonyms and polysemous words

6.3 Procedure for Collecting Results

We prompt the model with context and choices culled from the datasets, and record the response to compare with gold keys (Figure 1). We then tally the results.

6.4 Baselines

Having gone through the existing literature, we select the best-performing models for WSD tasks for comparison in Table 7. In some cases, the authors of a dataset have provided their benchmarks, which we include. To our knowledge, (Barba et al., 2021b) is the best-performing model on WSD. However, Blevins and Zettlemoyer (2020) is a well-performing model known for its strong performance in few-shot and zero-shot settings. This we mention in Table 4.

6.5 Setting up LLMs

We prepare the LLMs for generating appropriate responses by setting some parameters such as "expert" mode, "non-verbose" mode, and "safe" mode. The responses sometimes were found to contain some spurious content. We sanitized the output to collect the response. It took several iterations to arrive at a proper mechanism to capture the response.

6.6 Four LLMs

We experiment on four recent models. These models are recognized for their good performance across various NLP tasks such as Commonsense Reasoning, World Knowledge, and Reading Comprehension (Akter et al., 2023). We opt to choose a mix of open-source and proprietary models. Each model is subtly different in how they were trained.

Here are brief descriptions of these models:

6.6.1 ChatGPT-3.5

OpenAI's ChatGPT-3.5 is demonstrated to perform effectively on NLP tasks (Brown et al., 2020). Since it is a close-sourced model, the model parameters could not be ascertained. However, we experiment with it because of its popularity.

6.6.2 Mistral

We experimented on Mistral 7B, which has 7 billion parameters and is open-source. This model outperforms other open-source models. The Mistral model uses a Sliding Window Attention, which is particularly suited for long text (Beltagy et al., 2020), a feature must desired in disambiguation.

6.6.3 Llama

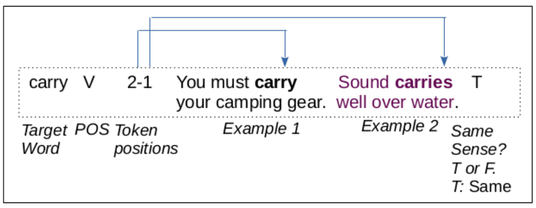
We experiment with the Llama 70 billion parameter model, which is open-sourced. We conduct experiments on it because it has been developed using open and accessible data. It also possesses comparable performance with the state-of-the-art (Touvron et al., 2023).

6.6.4 Gemini Pro

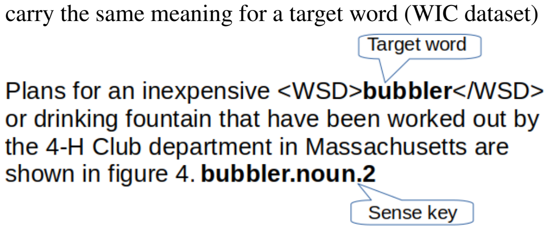
Gemini Pro is Google's latest language model. On various benchmark tests, it shows state-of-the-art performance (Team et al., 2023). Since it is a close-sourced model, the model parameters could not be ascertained.

7 Experimentation

The datasets we use in our experiments contain a variety of sense keys: some use their self-conceived sense keys extracted from popular text sources such as Wikipedia and Wiktionary. Some use WordNet sense keys. Still, others could be found using some other lexicon's keys, for example, BabelNet (Navigli and Ponzetto, 2012). Not all datasets present WSD as a classification task. For example, Word in Context (WIC) dataset (Pilehvar and Camacho-Collados, 2018) presents each evaluation sample as a simple *true* or *false* by giving two sentences, probing the LLM to verify whether the two sentences carry the same meaning of the target word (Figure 5a and Figure 5b).



(a) WSD posed as confirming whether the two sentences carry the same meaning for a target word (WIC dataset)



(b) Most datasets pose WSD as a classification task where a sense key is given as the class

Figure 5: WSD is posed differently in datasets

Data sets are in different formats: some in XML format while others in simple texts. Some datasets contain the sense keys, whereas others refer to senses from external sense inventories. After extracting the sentences and collecting definitions of senses suitable for prompting, we prepare prompts similar to Figure 1.

7.1 Datasets Considered

a. Eurosense Multilingual WSD Dataset (Bovi et al., 2017)
This dataset is the largest. It also contains multilingual content. However, the dataset lacks proper human evaluation – random samples reveal that it has 67.7% inter-annotator agreement. We do not include this dataset in our experiments.

b. NUS WSD Corpus (Dahlmeier et al., 2009)
This dataset only contains prepositions as the target of disambiguation. Since it does not provide other parts of speeches, we do not include this in our experiments.

c. Unified framework (Raganato et al., 2017)
This dataset contains a collection of datasets that researchers have been using since the 1990s. Since some of the most prominent research cites this dataset as a benchmark, we include this.

d. WiC (Word-in-Context) Dataset (Pilehvar and Camacho-Collados, 2018)
This dataset poses a WSD task in a novel way – that of contrasting two sentences to decide on the sameness of senses in the target word usage. We surmise that this test would be a good test on LLM to evaluate reasoning. Moreover, this dataset has been carefully created using VerbNet (Schuler, 2005), producing verb words as targets of disambiguation. Since Wiktionary has been used to collect data, human evaluation was factored in. Therefore, we include this in our experiments.

e. CoarseWSD-20 (Loureiro et al., 2021)
This dataset has been collected from Wikipedia. Authors report that random samples prove over 90% of the tags are accurate by validating with human annotators. We include this dataset in our experiments.

f. FEWS dataset (Blevins et al., 2021)
This dataset has been created based on the notion of WSD's poor performance on rare senses. In fact, it has been reported that humans outperform the best baseline models on this dataset. The dataset has been created from examples and definitions in Wiktionary, which is human-created. We include this in our experiments.

Table 7: Accuracy (%) found in our experiments. The *COMPARISON* column gives the accuracies obtained by some well-performing models: Sl. 1-5: (Barba et al., 2021b), Sl. 6: (Pilehvar and Camacho-Collados, 2018), Sl. 7: (Loureiro et al., 2021), Sl. 8: (Blevins et al., 2021). *IAA Column*: Inter-Annotator Agreement. *: for Verbs and Nouns, respectively.

Sl.	Dataset	OpenAI	Mistral	Llama	Gemini	COMPARISON	IAA
1	Senseval-2	65.7	65.0	61.0	**71.1**	82.3	-
2	Senseval-3	61.5	58.8	54.5	**70.0**	79.9	72.5
3	Semeval-2007	58.4	55.7	49.1	**65.4**	77.4	72,86*
4	Semeval-2013	70.1	65.9	66.5	**74.1**	83.2	-
5	Semeval-2015	67.3	64.1	63.0	**72.9**	85.2	68
6	WiC (Word-in-Context)	59.4	61.6	55.1	**65.8**	58.0	80
7	CoarseWSD-20	84.1	61.6	33.8	**93.9**	95.0	-
8	FEWS few-shots	63.0	63.7	60.7	**71.0**	66.4	80.2
	zero-shot	59.0	58.7	56.7	**65.0**	-	

Table 8: Pricing per a million tokens. * Llama was accessed through replicate.com.

Language Model	Input	Output
ChatGPT	$0.5	$1.5
Mistral	$4.0	$12.0
Llama*	$0.65	$2.75
Gemini Pro	$0.35	$1.05

7.2 Results

We test nine datasets on each of the four LLMs. Each language model is prompted with a sentence and told to disambiguate a target word. The response of the language model is observed and recorded. Table 7 shows the accuracy found by comparing it with the gold sense key.

8 Discussion

Given that the LLMs have not been fine-tuned, it is understandable from the test results that accuracy is comparable to the state-of-the-art models on WSD. Sometimes a language model fails to accurately identify a sense due to its lack of spatial knowledge; other times it fails because it seems not to be able to put the text in historical context; still other times the lack of application of humans' social relation is to be the reason for failure.

Many disambiguation cases require knowledge from different avenues: political, spatial, cultural, historical, and the like. Many researchers would sometimes club these missing pieces as common-sense knowledge. While investigating the failure cases, we prompted the LLMs to test their world knowledge. We discovered that by using different prompts, it can be confirmed that the LLMs appear to possess much of this knowledge. However, the failure arises when these models do not leverage knowledge across multiple dimensions to integrate it effectively. Much research in the avenue of reasoning is needed to further advancement of Artificial General Intelligence, which concurs with some research findings (Chen et al., 2023).

We stop short of calling our results a benchmark since not all LLMs we considered are open-source and the technology is continuously evolving as a result of which it will be difficult to compare across generations of LLMs.

9 Conclusion

In this research, we demonstrate that WSD involves not just the knowledge of language but world knowledge and the capability of piecing together facts from multiple sources — in other words, functional competence. Our findings also suggest that WSD could be used to verify the reasoning power of LLMs. WSD datasets are aplenty, and some have been human-validated. We conclude that it is worth paying heed to improving the WSD capabilities of LLMs and using these datasets in a novel way to probe. We also release a taxonomy of failure cases requiring world knowledge for WSD, which could further research in this direction.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Mark Aronoff and Kirsten Fudeman. 2022. *What is morphology?* John Wiley & Sons.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. Esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Gábor Berend. 2020. Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8498–8508, Online. Association for Computational Linguistics.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. Fews: Large-scale, low-shot word sense disambiguation with the dictionary. *arXiv preprint arXiv:2102.07983*.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say what you mean! large language models speak too positively about negative commonsense knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9890–9908, Toronto, Canada. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Tanja Schultz. 2009. Joint learning of preposition senses and semantic roles of prepositional phrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 450–458, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gerard Escudero, Lluis Marquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word disambiguation systems. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 172–180.

Christiane Fellbaum and George Miller. 1998. Building semantic concordances.

P Sargant Florence. 1950. Human behaviour and the principle of least effort. *The Economic Journal*, 60(240):808–810.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. *Word sense disambiguation*, pages 47–73.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore. Association for Computational Linguistics.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. *arXiv preprint arXiv:2212.10769*.

Paul Kiparsky. 1982. Word formation and the lexicon. In *1982 Mid America Linguistics Conference Papers/Dept. of Ling., Univ. of Kansas*.

Philipp Koehn. 2020. *Neural machine translation*. Cambridge University Press.

Harsh Kohli. 2021. Training bi-encoders for word sense disambiguation. In *International Conference on Document Analysis and Recognition*, pages 823–837. Springer.

Jaechan Lee, Alisa Liu, Orevaoghene Ahia, Hila Gonen, and Noah Smith. 2023. That was the last straw, we need more: Are translation systems sensitive to disambiguating context? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4555–4569, Singapore. Association for Computational Linguistics.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-wei Lin, and Yang Liu. 2024. Llms for relational reasoning: How far are we? *arXiv preprint arXiv:2401.09042*.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023. Evaluate what you can't evaluate: Unassessable generated responses quality. *arXiv preprint arXiv:2305.14658*.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. *arXiv preprint arXiv:1906.10007*.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

George A Miller. 1990. Nouns in wordnet: a lexical inheritance system. *International journal of Lexicography*, 3(4):245–264.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Sakae Mizuki and Naoaki Okazaki. 2023. Semantic specialization for knowledge-based word sense disambiguation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3457–3470, Dubrovnik, Croatia. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Steven Neale, Luís Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. 2016. Word sense-aware machine translation: Including senses as contextual features for improved translation models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2777–2783, Portorož, Slovenia. European Language Resources Association (ELRA).

Quang-Phuoc Nguyen, Anh-Dung Vo, Joon-Choul Shin, and Cheol-Young Ock. 2018. Effect of word sense disambiguation on neural machine translation: A case study in korean. *IEEE Access*, 6:38512–38523.

OpenAI. 2022. Openai: Introducing chatgpt.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Alessandro Raganato, Jose Camacho-Collados, Roberto Navigli, et al. 2017. Word sense disambiguation: a uinified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 99–110.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Eneko Agirre, and German Rigau. 2023. What do language models know about word senses? zero-shot wsd with language models and domain inventories. *arXiv preprint arXiv:2302.03353*.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bastin Tony Roy Savarimuthu, Surangika Ranathunga, and Stephen Cranefield. 2024. Harnessing the power of llms for normative reasoning in mass.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Lütfi Kerem Senel, Timo Schick, and Hinrich Schütze. 2022. Coda21: Evaluating language understanding capabilities of nlp models with context-definition alignment. *arXiv preprint arXiv:2203.06228*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. arxiv.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Ming Wang, Jianzhang Zhang, and Yinglin Wang. 2021. Enhancing the context representation in similarity-based word sense disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8965–8973, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Zhengxuan Wu, Christopher D Manning, and Christopher Potts. 2023. Recogs: How incidental details of a logical form overshadow an evaluation of semantic interpretation. *Transactions of the Association for Computational Linguistics*, 11:1719–1733.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting bert for word sense disambiguation with gloss selection objective and example sentences. *arXiv preprint arXiv:2009.11795*.

S Ye, Y Xie, D Chen, Y Xu, L Yuan, C Zhu, and J Liao. 2022. Improving commonsense in vision-language models via knowledge graph riddles (2022). *Computing Research Repository*, 10.

Chun-Xiang Zhang, Rui Liu, Xue-Yao Gao, and Bo Yu. 2021. Graph convolutional network for word sense disambiguation. *Discrete Dynamics in Nature and Society*, 2021:1–12.

Yuhan Zhang, Edward Gibson, and Forrest Davis. 2023. Can language models be tricked by language illusions? easier with syntax, harder with semantics. *arXiv preprint arXiv:2311.01386*.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

A Appendix: A Taxonomy of the Failure Cases

Table 1 and 9 show a categorization of primary world knowledge required to decide on a sense. Reference sentences are given as examples.

B Appendix: Details of the Prompts

In Table 10, 11, 12, and 13 we list the prompts used to query the language models.

Table 9: Failure cases - Part II

WKR Category	WKR Sub Category	Example text	Remarks
5. Commonsense			
	5.1. Knowledge of Geography, Trade relations, Reasoning	The discovery of the mines of America ... does not seem to have had any very **sensible** effect upon the prices of things in England.	"sensible" is being used to provide counter-intuitive information against the expectation that America's affairs could have a perceivable impact on that of England. <u>The correct choice:</u> *Easily perceived; appreciable*
	5.2. Subject/Domain knowledge	The iron content of these growth habits varies as follows: plates and rosettes honeycomb **cabbagehead**.	"cabbagehead" is being used to refer to a composition of minerals. <u>The correct choice:</u> *A roughly spherical aggregation of a mineral*
6. Satire		his lordship was out of humor. That was the way Chollacombe described as knaggy an old gager as ever Charles had had the ill-**fortune** to serve.	"fortune" carries a sense of inevitability. <u>The correct choice:</u> *Destiny, especially favorable*
7. Figurative		One ambassador sent word to the duke's son that his visit should be **retaliated**.	"retaliated" is being used to mean a reciprocal action. <u>The correct choice:</u> *To repay or requite by an act of the same kind.*
8. Religious writing		How **impertinent** that grief was which served no end!	"impertinent" is found in a religious text where the word carries the meaning of lack of patience. <u>The correct choice:</u> *insolent, ill-mannered*
9. World knowledge		Dr. Bertrand tells us that the first patient he ever **magnetized**, being attacked by a disease of a hysterical character, became subject to convulsions of so long duration and so violent in character, that he had never, in all his practice, seen the like ...	"magnetized" is being used to alleviate hysteria. <u>The correct choice:</u> *To hypnotize using mesmerism*

WKR Column: Type of World Knowledge Required. The target word is bolded. The correct choice (last column) is the definition corresponding to the gold key.

Table 10: Prompts for GPT-3.5-Turbo-0125. We use the same prompt template for both 0-shot and few-shot test splits for the FEWS dataset. Also, we explicitly instruct the model not to provide any explanations to prevent it from generating verbose texts.

Dataset Name	**Prompt**
Unified Framework	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Do not provide explanations. Just output the choice.
CoarseWSD-20	Which of the following sense choices is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Do not provide explanations.
WiC	Is the sense of [TGT] same in the following two sentences, say Yes or No: sentence1: [SEN 1] sentence2: [SEN 2] Please do not provide explanations.
FEWS	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Print a choice. Do not provide explanations. Just output the choice.
	Acronyms: *SENSEDEFN: Sense definition;* *SEN: Sentence;* *TGT: Target word to be disambiguated;*

Table 11: Prompts for Mistral 7B. We use the same prompt template for both 0-shot and few-shot test splits for the FEWS dataset. Also, we explicitly instruct the model not to provide any explanations to prevent it from generating verbose texts.

Dataset Name	**Prompt**
Unified Framework	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Do not provide explanations.
CoarseWSD-20	Which of the following sense choices is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Do not provide explanations.
WiC	Is the sense of [TGT] same in the following two sentences, say Yes or No: sentence1: [SEN 1] sentence2: [SEN 2] Please do not provide explanations.
FEWS	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Print a choice. Do not provide explanations.
	Acronyms: *SENSEDEFN: Sense definition;* *SEN: Sentence;* *TGT: Target word to be disambiguated;*

Table 12: Prompts for Llama-2-70b-chat. We use the same prompt template for both 0-shot and few-shot test splits for the FEWS dataset. Also, we explicitly instruct the model not to provide any explanations to prevent it from generating verbose texts.

Dataset Name	**Prompt**
Unified Framework	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Do not provide explanations. Just output the choice.
CoarseWSD-20	Which of the following sense choices is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Do not provide explanations.
WiC	Is the sense of [TGT] same in the following two sentences, say Yes or No: sentence1: [SEN 1] sentence2: [SEN 2] Please do not provide explanations.
FEWS	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3] Print a choice. Do not provide explanations. Just output the choice.
	Acronyms: *SENSEDEFN: Sense definition;* *SEN: Sentence;* *TGT: Target word to be disambiguated;*

Table 13: Prompts for Gemini Pro. We use the same prompt template for both 0-shot and few-shot test splits for the FEWS dataset.

Dataset Name	Prompt
Unified Framework	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3]
CoarseWSD-20	Which of the following sense choices is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3]
WiC	Is the sense of [TGT] same in the following two sentences, say Yes or No: sentence1: [SEN 1] sentence2: [SEN 2]
FEWS	Which of the following senses is correct for the word [TGT] in the following text: [SEN] I) [SENSEDEF 1] II) [SENSEDEF 2] III) [SENSEDEF 3]
	Acronyms: *SENSEDEFN: Sense definition;* *SEN: Sentence;* *TGT: Target word to be disambiguated;*