Reassessing Semantic Knowledge Encoded in Large Language Models through the Word-in-Context Task

Yoshihiko Hayashi
Perceptual Computing Laboratory, Waseda University
Waseda-machi 27, Shinjuku, Tokyo Japan
yoshihiko.hayashi@gmail.com

Abstract

Despite the remarkable recent advancements in large language models (LLMs), a comprehensive understanding of their inner workings and the depth of their knowledge remains elusive. This study aims to reassess the semantic knowledge encoded in LLMs by utilizing the Word-in-Context (WiC) task, which involves predicting the semantic equivalence of a target word across different contexts, as a probing task. To address this challenge, we start by prompting LLMs, specifically GPT-3 and GPT-4, to generate natural language descriptions that contrast the meanings of the target word in two contextual sentences given in the WiC dataset. Subsequently, we conduct a manual analysis to examine their linguistic attributes. In parallel, we train a text classification model that utilizes the generated descriptions as supervision and assesses their practical effectiveness in the WiC task. The linguistic and empirical findings reveal a consistent provision of valid and valuable descriptions by LLMs, with LLM-generated descriptions significantly improving classification accuracy. Notably, the highest classification result achieved with GPT-3-generated descriptions largely surpassed GPT-3's zero-shot baseline. However, the GPT-4-generated descriptions performed slightly below GPT-4's zero-shot baseline, suggesting that the full potential of the most advanced large language models, such as GPT-4, is yet to be fully revealed.

Keywords: Word-in-Context task, Large language models, Prompting, Supervised text classification.

1. Introduction

Recent advancements in large language models (LLMs) have propelled significant progress, extending their application across various domains including dialogue systems, text generation, translation systems, and beyond. Conversely, efforts have been directed towards evaluating the technical capabilities of language models more broadly, through the development of benchmarking frameworks that aggregate standard NLP/NLU tasks such as Natural Language Inference (NLI) and Question Answering (QA). Notably, SuperGLUE (Wang et al., 2019) has emerged as a prominent example in this domain. While these initiatives underscore the versatility of LLMs, there remains uncertainty about the extent and depth of linguistic and/or real-world knowledge captured in these models.

This study specifically aims to reassess the semantic knowledge encoded in LLMs by employing the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019) as a probing task. This task involves predicting the semantic sameness of a target word across different contexts. We decided to focus on the WiC task because it directly addresses contextualized word meanings, a critical aspect of lexical semantics (Cruse, 1986; Ravin and Leacock, 2000).

In this study, we extend beyond merely evaluating task performances. Instead, we try to reassess semantic knowledge implicitly encoded in an LLM both linguistically and empirically. Initially, we prompt an LLM to articulate relevant semantic knowledge and the rationale behind its decisions in the WiC task. Subsequently, we conduct manual and linguistic investigations of the generated descriptions, enabling us to assess the validity of the semantic knowledge conveyed through the articulation. In parallel, we empirically assess the practical effectiveness of these descriptions. Specifically, we adopt a supervised text classification approach, as depicted in Figure 1, where a `Classifier` is trained using the LLM-generated descriptions. Throughout this study, we specifically utilize GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) as the most effective representative LLMs currently available[1].

After manual investigations of the LLM-generated descriptions and empirical examinations through classification experiments, we achieved the following main implications[2] :

- Analytically, it has been verified that LLMs typically offer valuable and valid information to distinguish or contrast the meanings of a target word in context, drawing upon various linguistic patterns.
- Experimental findings validate that descriptions generated by LLMs generally improve the classification accuracy of the WiC task. How-

---

[1] In the initial stage of this study, we also experimented with some open-source LLMs, but they yielded unstable and suboptimal results.

[2] Relevant codes and data will be available at: `https://github.com/yoshihikohayashi/wic_llm`.

LREC-COLING 2024, pages 13610–13620
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

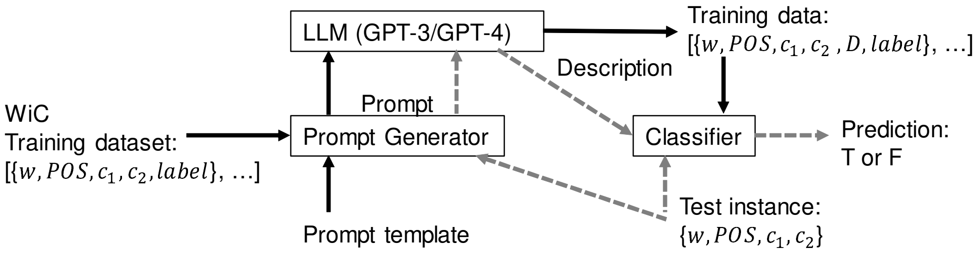

Figure 1: Supervised text classification approach. Solid arrows represent data flow during training, while dotted arrows represent data flow during testing.

ever, the highest performance achieved with GPT-4-generated descriptions falls below GPT-4's zero-shot baseline accuracy, underscoring the untapped potential of the most advanced LLM.

In addition to these results, our findings indicate that the WiC task and its associated dataset still present challenges regarding the criteria for semantic sameness. This suggests the potential for addressing these issues through dataset revision leveraging advanced LLMs, such as GPT-4.

2. WiC Task and the Dataset

In the WiC task, the system is tasked with determining whether a target word (w), a verb or noun, demonstrates semantic sameness within two contextual sentences (c_1 and c_2). The WiC dataset[3] maintains annotated instances, each labeled to indicate whether the meanings of the target word in the context sentences are identical (positive; labeled T) or distinct (negative; labeled F). In Appendix A, Figure 6 exemplifies class T and F instances, and Table 9 summarizes the dataset configuration. Given the balanced positive/negative ratio, system performance is assessed using the accuracy metric.

Zero-shot baselines by GPT-3 and GPT-4: In this study, we consider the outcomes produced by GPT-3 and GPT-4 when prompted by the zero-shot template shown in Figure 7 (Appendix B) as baseline performances. GPT-3 achieved accuracies of 0.633 on the validation split and 0.619 on the test split of the WiC dataset, which aligned with our expectations. In contrast, the same template led to accuracies of 0.723 and 0.753, respectively, with GPT-4, a result that was surprisingly close to the state-of-the-art performance of approximately 0.77 as reported on the SuperGLUE leaderboard[4].

3. Related Work

The WiC task was initially introduced by Pilehvar and Camacho-Collados (2019), and the shared task campaign SemDeep-5, held in the same year, collected several research works (Loureiro and Jorge, 2019; Garí Soler et al., 2019; Ansell et al., 2019). Since then, it has evolved to encompass tasks beyond the English-only context. Notable examples include SemEval-2021 Task 2 (Martelli et al., 2021), which deals with multilingual and cross-lingual settings, and EVALITA 2023 (Cassotti et al., 2023), which focuses on the task in the Italian language. However, our study remains focused on the original English version of the WiC task at present.

The WiC dataset is now an integral part of the comprehensive NLP/NLU benchmark SuperGLUE (Wang et al., 2019), and its leaderboard features top scores, including those for the WiC task, which typically reach an accuracy of around 0.77. Many state-of-the-art systems that have achieved these top results devise advanced learning methods, including both the pretraining of a language model and its fine-tuning for downstream tasks. For example, the Vega v2 system (Zhong et al., 2022) successfully introduced a self-evolution learning strategy for efficient pretraining and a prompt transfer method based on knowledge distillation to adapt to downstream tasks. However, these papers did not provide any linguistically oriented discussions, regarding any of the subtasks, including WiC.

To the best of the authors' knowledge, there have been few studies exploring the application of LLMs in the WiC task. A study initially conducted by Brown et al. (2020) reported accuracies of 0.494 for the original GPT-3 in a few-shot setting. The authors argue that the model may encounter difficulties when comparing two sentences, as required in the WiC task. More recently, Laskar et al. (2023) reported an accuracy of 0.621 for `gpt-3.5-turbo` in a zero-shot setting, which represents a significant advance from the results of Brown et al. (2020). However, these studies again lack comprehensive linguistic analyses. In this context, our paper may

---

[3] https://pilehvar.github.io/wic/

[4] https://super.gluebenchmark.com/leaderboard

be the first to present not only experimental results but also a linguistic analysis. It is essential to emphasize that the primary aim of this study is not the advancement of learning methods but rather the exploration of LLMs' capabilities in tackling a semantic task, exemplified by the WiC task.

4. LLM-generated Descriptions

What we aim to ascertain about an LLM-generated description includes: whether it effectively offers an explanation of the meaning of the target word in a given context; whether it introduces linguistic expressions that could guide towards a semantic sameness decision, and whether the implied answer is correct or, at the very least, corresponds with the gold label in the dataset. Furthermore, we aim to explore the connections between these factors and their actual empirical performance when the LLM-generated descriptions are employed in solving the WiC task.

4.1. Description Types

We generated two types of textual descriptions, *direct* and *contrast*, by accordingly prompting both GPT-3 and GPT-4, resulting in a total of four textual descriptions for each WiC instance.

Figure 2 presents the prompt template for creating *direct* descriptions, which includes the request for a Yes or No answer along with the rationale. By analyzing this type of description, we can investigate the direct connection between the decision of semantic sameness (answer) and the underlying semantic knowledge that guides that decision (rationale).

```
Identify if the target word "{word}"
in the following sentences correspond
to the identical meanings or not.
Answer Yes or No, and provide one
brief sentence to describe the
rationale behind the decision.
1: {c1}
2: {c2}
```

Figure 2: Prompt template for *direct* descriptions.

On the other hand, Figure 3 presents the template for generating descriptions using the directive verb "contrast," which does not compel an LLM to provide a direct answer in a description. This type of description could enable us to simply explore the detailed semantic knowledge encoded in an LLM without necessarily forcing it to produce a simple Yes or No answer.

Tables 1 and 2 exemplify the *direct* and *contrast* descriptions produced by both GPT-3 and GPT-4 for an instance from the validation split of the WiC dataset, which is illustrated as the second example (the target word: excite/V) shown in Figure 6 (Appendix A).

```
Contrast the core senses of "{word}"
in these two sentences and summarize
the result in a short sentence.
1: {c1}
2: {c2}
```

Figure 3: Prompt template for *contrast* descriptions.

4.2. Global Statistics of the Descriptions

Table 3 provides basic statistics, computed for the validation split of the WiC dataset,[5] concerning the descriptions categorized by description type and the LLM used.

These statistics highlight the overall trend in terms of text length, encompassing the average number of characters, sentences, and tokens per description. It clearly shows that *direct* descriptions typically have a shorter length compared to *contrast* descriptions. Furthermore, the descriptions generated by GPT-4 are generally more concise than those produced by GPT-3. These trends are also observed from the examples displayed in Table 1 and Table 2.

4.3. Discourse Patterns of the Descriptions

The *direct* descriptions and the *contrast* descriptions exhibit different patterns of discourse development, as demonstrated by the examples in Table 1 and Table 2. These patterns are generally shared by both GPT-3 and GPT-4. In the case of *direct* descriptions, they tend to first summarize high-level semantic commonalities and then contrast the specific meanings of the target word in both contextual sentences, followed by a summary sentence. On the other hand, *contrast* descriptions tend to start with a series of sentences, with each sentence describing the specific meaning of the target word in a contextual sentence. They are also followed by a summary sentence.

4.4. Accuracy of the Answers Provided in the *direct* Descriptions

Without any specific formatting request given in the prompt, all *direct* descriptions generated by GPT-4 begin with a Yes or No answer, and a significant number of those generated by GPT-3 (587 out of 638 instances in the validation set) also commence with "Yes" or "No" as the initial word. Therefore,

---

[5] The training and test splits exhibit similar statistics.

LLM	Description
GPT-3	Yes. Rationale: In both sentences, the word "excite" is used to describe a state of stimulation or arousal. In the first sentence, it refers to stimulating the neurons, while in the second sentence, it refers to the fireworks stimulating or arousing the people present. Therefore, the target word "excite" corresponds to identical meanings in both sentences.
GPT-4	Yes, the target word "excite" in both sentences has the same meaning. In both contexts, "excite" is used to describe the act of stimulating or causing a heightened response or activity, whether it's in neurons or people at a festival.

Table 1: Examples of LLM-generated *direct* descriptions. (target word: excite/V, c_1: Excite the neuron, c_2: The fireworks which opened the festivities excited anyone present.)

LLM	Description
GPT-3	In the first sentence, "excite" refers to stimulating or activating the neurons. In the second sentence, "excite" means to arouse or generate enthusiasm in anyone present due to the fireworks. The result is that "excite" has a literal and physical connotation in the first sentence, while it has a figurative and emotional connotation in the second sentence.
GPT-4	In the first sentence, "excite" refers to stimulating or activating neurons, a physical or biological process. In the second sentence, "excite" refers to causing an emotional response of enthusiasm or eagerness in people. Therefore, "excite" can refer to either a physical stimulation or an emotional stimulation.

Table 2: Examples of LLM-generated *contrast* descriptions (for the same WiC data instance of Table 1).

Type/Model	Length	# Sents	# Tokens
contrast/GPT-3	336.0	2.9	66.9
contrast/GPT-4	292.8	2.1	59.9
direct/GPT-3	275.3	4.1	57.9
direct/GPT-4	163.8	1.4	35.1

Table 3: Statistics of LLM-generated descriptions.

accuracy scores can be calculated by comparing these answers with the gold labels. The scores for the validation set instances are 0.649 for GPT-3 and 0.726 for GPT-4 generated descriptions, respectively. These results are comparable to the GPT-3/GPT-4 zero-shot baselines described in section 2. This could be expected since the prompt for generating *direct* descriptions is, in essence, equivalent to the zero-shot prompt.

4.5. Manual Investigation

We randomly selected 100 WiC instances from the validation split of the dataset and conducted a manual investigation of their descriptions. Our main objective is to evaluate the accuracy of LLM-generated description text as an explanation of the contextualized meaning of a target word and to assess its relevance to the semantic sameness judgment provided by the gold label. For each of the sampled WiC instances, we reviewed all four descriptions, each generated by specifying a combination of description type (*direct* or *contrast*) and the LLM used (GPT-3 or GPT-4).

4.5.1. Validity of descriptions

We flagged a description as invalid if it contained information that was either significantly irrelevant or contradicted the contextualized meaning of the target word. Many of these "errors" were found in the concluding summary sentence of the description. An example of this description kind is displayed in Table 4, illustrating the case of the noun "breeze." Although the explanations for both contextualized meanings are adequate, the final sentence fails to accurately summarize them, as "breeze" cannot refer to the temperature of the air.

c_1 c_2	The **breeze** rustled the papers on her desk. The **breeze** was cooled by the lake.
Desc.	In the first sentence, "breeze" refers to a light wind that is causing movement, while in the second sentence, "breeze" refers to a flow of air that is being affected by temperature. In summary, "breeze" can refer to both the movement and *temperature of air*.

Table 4: Example of deemed invalid description.

The summary of the validity results is shown in Table 5, where each number represents the count of valid descriptions among the hundred samples. Notably, a significant number of descriptions are considered valid, and both LLMs show similar levels of validity in the generated descriptions. We note that GPT-4 often attempts to provide richer information, such as grammatical notices but seems to fail occasionally, resulting in relatively lower counts.

Description type	GPT-3	GPT-4
direct	96	93
contrast	96	90

Table 5: Number of valid descriptions out of 100 WiC data instances.

4.5.2. Relevance to the gold label

We also examined each description to determine if it contained an expression explicitly indicating the answer (Yes or No) and whether it aligned with the gold label (T or F) provided in the dataset. This analysis specifically concentrated on words or expressions recognized as immediately conveying the decision, including terms like "same," "identical," "different," "vary," and similar ones.

The results are displayed in Table 6: the M-column and UM-column respectively count the number of descriptions that contain an immediate expression that matched and did not match with the gold labels, while the N-column indicates the number of descriptions lacking immediate expressions. For example, the GPT-3-generated `direct` description, as seen in Table 1, falls under the M-category due to the presence of the immediate expression "corresponds to identical meanings" in its final sentence. This demonstrates alignment with the gold label (T), as shown in Figure 6.

	GPT-3			GPT-4		
	M	UM	N	M	UM	N
direct	40	17	43	23	18	59
contrast	5	5	90	6	1	92

Table 6: Relevance of descriptions to the gold labels. The M column represents the count of descriptions that contain an exact match with the gold labels, while the UM column represents those that do not match. The N-column indicates the number of descriptions lacking immediate expressions.

Note that when counting the numbers, we did not consider the specific presence of "Yes" or "No" at the beginning of a *direct* description. That is, the 41 $(= 23 + 18)$ *direct* descriptions generated by GPT-4 contain direct expressions, regardless of whether the descriptions start with "Yes" or "No." The remaining 59 descriptions commence with "Yes" or "No," but they do not include direct expressions within the description. The ratio of matched descriptions for the *direct* descriptions is much higher for GPT-3 $(0.70 = 40/(40+17))$ compared to GPT-4 $(0.56 = 23/(23 + 18))$, whereas the ratio by GPT-4 is significantly higher for the *contrast* descriptions. While we cannot definitively conclude based on this limited-scale investigation, these results may suggest the difficulty of fully identifying the knowledge inherent in an LLM from its result of articulation. One short note: the high number of gold-matched descriptions does not necessarily correlate with the quality of descriptions. This is due to the potential unreliability of the gold labels, as discussed in the later part of this paper.

4.6. Elements that Could Lead to the Semantic Sameness Decision

In lexical semantics, lexical ambiguity is categorized into homonymy and polysemy (Pustejovsky and Batiukova, 2019). Detecting homonymy and answering 'F' in the WiC task should be relatively straightforward for an LLM, as it may possess sufficient knowledge to discern *unrelated* meanings.

On the other hand, polysemy poses a greater challenge. Polysemy refers to the situation where a word has multiple *related* meanings. In this context, the concern is focused on the extent and specific categories of related meanings that can be classified as "identical" within the WiC task. There are no predefined and stringent criteria in place for making this determination.

Nonetheless, LLM-generated descriptions often provide clues by presenting contrasting pairs of expressions, including examples like "personal" and "collective," "individual" and "general," "metaphorical" and "literal," "tangible" and "intangible," "physical" and "emotional," among others. While these expressions can enhance the understanding of contextual nuances in some cases, they do not always lead to answers that align with the gold labels.

5. Supervised Text Classification Approach

A supervised text classification approach is performed to evaluate the practical utility of the LLM-generated descriptions in the WiC task.

5.1. Overview

Figure 1 provides an overview of the setup for the supervised text classification approach. In the diagram, the `Prompt Generator` is a built-in Python method used for formatting string data, the `LLM` is either GPT-3 or GPT-4, and the `Classifier` consists of a BERT text encoder with a binary classification head on top of it. Solid arrows depict the data flow during the training, while dotted arrows signify the flow during testing. With this supervised learning setup for the WiC task, experiments were primarily conducted to verify and compare the semantic knowledge captured by the LLMs.

As detailed in the previous section, descriptions (either of the *direct* or *contrast* type) that elucidate the meanings of the target word in the contextual sentences are generated by prompting an LLM

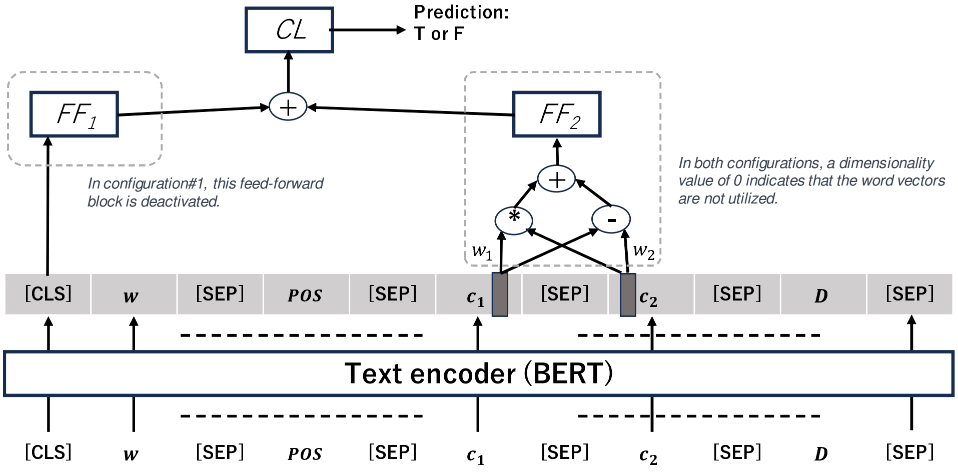
CL
Prediction:
T or F
FF1
FF2
+
*
-
In configuration#1, this feed-forward block is deactivated.
In both configurations, a dimensionality value of 0 indicates that the word vectors are not utilized.
w1
w2
[CLS]
w
[SEP]
POS
[SEP]
c1
[SEP]
c2
[SEP]
D
[SEP]
Text encoder (BERT)

Figure 4: Configuration of the `Classifier`.

(GPT-3 or GPT-4). The descriptive text D, as illustrated in Tables 1 and 2, serves as input to the Classifier when combined with the target word w, its associated part-of-speech POS, and two contextual sentences c_1 and c_2.

Figure 4 zooms up the Classifier, displaying its architecture and the input format. The input data is formatted as indicated in the lowermost part of the diagram and then fed into the BERT encoder. Our primary focus is on the representation obtained from the entire input, so we leverage the encoder output vector of the [CLS] token. This vector is then passed to the feed-forward block labeled as FF_1.

In addition to the [CLS] token vector, we try to enhance the input representation by incorporating the contextualized word vectors w_1 and w_2 derived from the occurrences of the target word w in the contextual sentences c_1 and c_2. These vectors undergo a series of mathematical operations (multiplication marked by '*', subtraction '-', and concatenation '+') before being sent to the feed-forward block FF_2. Note that multiplication is employed to capture the commonalities between the two representations, while subtraction is used to accentuate the differences. Finally, the vectors achieved from the [CLS] token and contextualized representations of the target word are subsequently concatenated and directed to the classification block CL to yield the final prediction.

5.2. Experimental Setup

The original data split in the WiC dataset was preserved, with 5,468 instances used for training, 638 instances for validation, and 1,400 instances for testing. We used accuracy as the primary evaluation metric, consistent with previous WiC studies.

The training processes were managed using PyTorch's `Trainer` with the following hyperparameters: batch size: 16, learning rate: 4e-5, epochs: 20 with early stopping and patience of 5, scheduler: linear, weight decay: 0.1. The experimental runs for each configuration were repeated five times using different random seeds, and we report the averaged accuracy scores in the results section.

We used the API provided by OpenAI[6] to access GPT-3 (`gpt-3.5-turbo-0613`) and GPT-4 (`gpt-4-0613`). We made use of the Hugging Face's transformers library, specifically employing the `bert-base-uncased` pretrained model for the text encoder.[7] The dimensionality of each word vector, including the [CLS] vector, is 768.

All computations were performed on a single machine equipped with two NVIDIA RTX A4000 GPUs. Each training process took, at most, around 10 to 15 minutes, with most completed in fewer than 15 epochs.

5.3. Main Results

Multiple experimental runs were conducted, each involving variations in the classifier model specifications. The main results, as shown in Table 7, were initially categorized into two configurations. In Configuration#1, the FF_1 block in Figure 4 is deactivated, which means that the [CLS] vector was used as is, without passing through the FF_1 block. This FF_1 module is activated in Configuration#2 with a fixed output dimensionality of 512. Additionally, we experimented with different dimensionalities for the output vectors from the FF_2 block, choosing from the following options: [32, 128, 512]. We also

---

[6] https://platform.openai.com/docs/models

[7] We did experiment with alternative encoders, including `bert-large` and `roberta-base`; however, their performance was either comparable or inferior to that of `bert-base`.

	Configuration#1		Configuration#2	
	GPT-3	GPT-4	GPT-3	GPT-4
none	0.671 (512)		0.673 (512)	
direct	0.696 (512)	**0.733** (0)	0.702 (512)	**0.733** (0)
contrast	0.701 (0)	0.718 (512)	0.703 (128)	0.716 (32)

Table 7: Main results of the text classification approach: Accuracy scores, followed by the dimensionality of the word vector in parentheses.

explored special cases where the contextualized word vectors were not employed. These setups enabled us to assess the impact of word vectors on the model's performance.

The main results are summarized in Table 7, with accuracy as the evaluation metric. In the first column of the table, the description type is categorized, where "none" particularly indicates scenarios where no generated descriptions were used. It's crucial to note that we excluded the initial "Yes" or "No" found in the *direct* descriptions, meaning that the classification model cannot reference the predictions made by LLMs and verbalized in the descriptions.

Each cell in Table 7 displays two numbers: the first represents the average accuracy computed from five trials with different random seeds, while the second, shown in parenthesis, indicates the dimensionality of the output vector from the FF_2 block shown in Figure 4. This dimensionality value corresponds to one of the variations [0, 32, 128, 512] that produced the best result under the condition defined by the model configuration and the language model used. Note that a value of 0 specifically indicates that no word vectors are utilized. This means that only the vector originated from the [CLS] token, whether through FF_1 or not, is used.

The significant findings identified in Table 7 are organized as follows:

1. The use of LLM-generated descriptions has led to a 2 to 6-point increase in accuracy compared to cases where no descriptions were used. These results confirm the effectiveness of using LLM-generated descriptions in the WiC task.

2. GPT-4-generated descriptions generally produced better results than those generated by GPT-3. Specifically, the *direct* descriptions generated by GPT-4 attained the highest accuracy of 0.733 in both of the experimented model configurations, demonstrating the effective utilization of textual descriptions by the classification model. While this accuracy is slightly lower than the zero-shot accuracy of 0.753, it can still be considered relatively good.

3. Interestingly, while both types of descriptions generated by GPT-4 yielded almost comparable results, the GPT-3-generated *contrast* descriptions led to slightly better results than its *direct* descriptions. However, both types outperformed the results obtained with the GPT-3 zero-shot baseline, which had an accuracy of 0.619 on the test split of the WiC dataset.

4. Using contextualized word vectors improved accuracy scores in some cases, though the improvements were not significant. The GPT-4-generated *direct* description yielded the most favorable outcomes, even without relying on word vectors, indicating that GPT-4 proficiently encapsulated the description content.

5.4. Analysis

The results of the manual investigation presented in Tables 5 and 6 indicate that the quality of GPT-4-generated descriptions is not significantly better than those generated by GPT-3. However, the classification results in Table 7 show that GPT-4 yielded significantly better accuracies. This apparent contradiction raises interesting questions about the effectiveness of the generated descriptions versus the classification model's ability to utilize them effectively. While it is difficult to prove quantitatively, the authors posit that the improved performance of GPT-4 may be attributed to the textual features of the descriptions it generates. To be more specific, the descriptions generated by GPT-4 tend to be more concise and coherent in style. Moreover, GPT-4 often exhibits strong abstraction capabilities, which could aid in the effective processing of the descriptions.

To further investigate the insights outlined in the second and third bullets in the above list, we perform pairwise comparisons of the `Classifiers` in Configuration#2, as summarized in Table 8. In the table, columns A and B denote the configuration of the classifiers being compared. For instance, c-3 represents that the classifier is trained with the GPT-3 generated *contrast* descriptions. The figures in the table are derived from results obtained by a classifier with the dimensionality of contextualized word vectors as shown in Table 7, and the optimal random seed selected from among the five random seeds. The columns Both-OK and Both-NG show the instances correctly and incorrectly predicted by

	A	B	Both-OK↑	Only-A	Only-B	Both-NG↓	And-Acc↑	Or-Acc↑
contrast vs. *direct*	c-3	c-4	847	164	158	231	0.605	0.835
	d-3	d-4	**857**	172	144	227	**0.612**	0.838
GPT-3 vs. GPT-4	c-3	d-3	838	167	163	232	0.599	0.834
	c-4	d-4	838	173	191	198	0.599	0.859
Other combinations	c-3	d-4	823	**182***	**206****	**189**	0.588	**0.865**
	c-4	d-3	826	185	175	214	0.590	0.847

Table 8: Pairwise comparison of the `Classifiers`. Columns A and B indicate the configuration of the classifiers under comparison. For example, c-3 signifies that the classifier is trained with GPT-3 generated *contrast* descriptions. The number 182 marked with * signifies that GPT-3-generated *contrast* descriptions yield the highest number of Only-A instances, while the number 206 marked with ** represents that GPT-4-generated *direct* descriptions yield the highest number of Only-B instances.

both Classifiers A and B, respectively. The columns Only-A and Only-B display the instances exclusively correctly predicted by Classifiers A and B, respectively. In addition to the columns provided, the And-Acc column reflects the accuracy score when both predictions made by Classifiers A and B are correct, while the Or-Acc column indicates the accuracy score when at least one of the two classifiers' predictions is correct. Consequently, the latter demonstrates the upper bound performance achievable when the two classifiers are ideally combined. Potentially deeper insights gleaned from the results can be summarized as follows:

***contrast* vs. *direct*:** The first two rows present the results comparing the *contrast* and *direct* descriptions. In summary, the *direct* descriptions yield superior results compared to the *contrast* descriptions, which may be somewhat expected. Moreover, the combination of d-3 and d-4 classifiers achieves the highest And-Acc score. This result suggests that prompting an LLM, especially GPT-4, to generate more task-directed descriptions is more beneficial.

GPT-3 vs. GPT-4: The next two rows compare the results obtained with descriptions generated by GPT-3 and GPT-4. Interestingly, while the Or-Acc score for GPT-4-generated descriptions is better than that for GPT-3-generated descriptions, the And-Acc score is comparable. This suggests that combining classifiers trained with descriptions generated by different LLMs could be effective in improving task performance.

The best combination: The last two rows display the results for other potential combinations of classifiers. Notably, the best Or-Acc score is achieved by the combination of c-3 and d-4 classifiers. Significantly, this combination also yielded the highest number of instances for Only-A and Only-B: with the GPT-3-generated *contrast* descriptions (182 marked with *) and the GPT-4-generated *direct* descriptions (206 marked with **). This result consistently aligns with the findings presented in Table 7, endorsing the effectiveness of GPT-3-generated *contrast* descriptions and GPT-4-generated *direct* descriptions. Appendix C further compares descriptions generated by GPT-3 and GPT-4.

6. Discussion

The WiC task effectively circumvents certain challenges related to Word Sense Disambiguation (WSD)(Agirre and Edmonds, 2006; Navigli, 2009), such as the granularity of sense distinctions in sense inventories(Ide and Wilks, 2006), due to its relatively straightforward setup. However, the task encounters another inherent challenge owing to its binary classification format. As discussed in lexical semantics, the concept of semantic sameness is better understood as gradational rather than strictly binary. Nevertheless, WiC instances necessitate a binary decision, notwithstanding occasional brevity and vagueness in the provided contextual clues. This may give rise to instances where the gold labels assigned are occasionally dubious, potentially impacting the overall quality of the dataset.

To address this concern, we examined 115 test instances from the test split of the WiC dataset. These instances, consisting of 75 noun instances and 40 verb instances—slightly skewed towards nouns—were predicted incorrectly by any combination of the classifiers shown in Table 8. Upon initial investigation of these 115 instances, we suspect that the annotated labels of nearly half of them should be corrected[8]. Two illustrative examples are presented in Figure 5. In the first example, c_1 depicts a horse racing scene using the idiomatic expression, while c_2 refers to a human body part; it is crucial to differentiate their meanings and hence should be labeled F. In the second example, the verb "wash" denotes activities of cleaning something in both contextual sentences; therefore, it should be labeled T. We observe that in 78 of the 115 instances, the assigned label is F, suggesting a possibility that cases with the same meaning tend

---

[8] The proportion may change after a thorough inspection by a native English speaker.

to be mislabeled as distinct meanings.

```
w/POS: neck/N
c1: The horse won by a neck.
c2: He admired her long graceful neck.
label: T

w/POS: wash/V
c1: Dishwashers wash dishes much more
efficiently than most humans.
c2: Wash the towels, please!
label: F
```

Figure 5: Examples of WiC instances with a questionable gold label.

We further suspect that another source of these problematic annotations could be attributed to the dataset construction process, which involved automatic sense mapping across different lexical resources (Pilehvar and Camacho-Collados, 2019), even though the criteria for sense distinction may vary depending on the specific resources in use. However, these issues should, at least partly, be addressed by employing an advanced LLM like GPT-4, which has the potential to enable better sense mapping and provide more reliable labels when guided with carefully curated prompts.

7. Conclusion

This paper aimed to reassess the latent semantic knowledge captured in LLMs by utilizing the WiC task as a probe. We prompted GPT-3 and GPT-4 to generate contrasting descriptions for a target word's meanings in various contexts. Our manual analysis confirmed the validity of these descriptions in explaining these meanings across different contexts and their relevance to the semantic sameness decisions in the WiC dataset. Additionally, we developed a text classification model for the WiC task, demonstrating the practical value of these descriptions in predicting semantic sameness.

It is, however, suggested that a considerable amount of untapped semantic knowledge exists within LLMs, as evidenced by the exceptional performance of the zero-shot GPT-4 baseline. We believe that further exploration and utilization of such an advanced LLM could yield deeper insights into the semantic knowledge it encodes, thereby enhancing semantic task performances such as the persistently challenging WiC task.

One practical application of predicting the semantic sameness of a polysemous word in context is detecting semantic gaps in a knowledge graph pathway, where the nodes haven't been disambiguated beforehand. By identifying a node where two edges can't be chained because of the semantic disparity of the node in each edge, it would be possible to prevent invalid inference chains (Hayashi, 2022).

Limitations

Below are several limitations that we acknowledge:

- In our evaluation, we have exclusively assessed proprietary LLMs such as GPT-3 and GPT-4. One potential concern associated with the utilization of proprietary LLMs is the opacity surrounding their training data, which could potentially affect experimental outcomes when employing publicly available datasets. In this study, the zero-shot baseline performances might be influenced by this characteristic, but the generated descriptions would remain unaffected. Consequently, the results of the supervised text classification approach would remain valid.
- We have not yet conducted a cost analysis of using OpenAI's API.
- We have not yet conducted an exhaustive hyperparameter tuning during the training process of the classifier, as the primary focus of this study is not geared toward achieving state-of-the-art performance levels.
- Almost for the same reason, we haven't optimized the prompts used in our experiments. The quest for an optimal prompt could be a promising direction to achieve better performances.
- Our manual investigation of generated descriptions has been confined to small-sized samples. A more extensive examination of a larger set of such texts would provide deeper insights into the verbalization patterns of the LLM's internal semantic knowledge.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22K12723.

Bibliographical References

Eneko Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, the Netherlands.

Alan Ansell, Felipe Bravo-Marquez, and Bernhard Pfahringer. 2019. An ELMo-inspired approach to SemDeep-5's word-in-context task. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 21–25, Macau, China. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Pierluigi Cassotti, Lucia Siciliani, Lucia C. Passaro, Maristella Gatto, and Pierpaolo Basile. 2023. Wic-ita at evalita2023: Overview of the evalita2023 word-in-context for italian task. In *Proceedings of EVALITA 2023*, Parma, Italy.

D.A. Cruse. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC challenge: Context representations for word usage similarity estimation. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 6–11, Macau, China. Association for Computational Linguistics.

Yoshihiko Hayashi. 2022. Towards the detection of a semantic gap in the chain of commonsense knowledge triples. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3984–3993, Marseille, France. European Language Resources Association.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In Edmonds Philip Agirre, Eneko, editor, *Word Sense Disambiguation: Algorithms and Applications*, pages 47–73. Springer Netherlands, Dordrecht.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Daniel Loureiro and Alípio Jorge. 2019. LIAAD at SemDeep-5 challenge: Word-in-context (WiC). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 1–5, Macau, China. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

OpenAI. 2023. Gpt-4 technical report.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

James Pustejovsky and Olga Batiukova. 2019. *The Lexicon*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Yael Ravin and Claudia Leacock, editors. 2000. *Polysemy: Theoretical and Computational Approaches*. Oxford University Press, Oxford.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, Xinbo Gao, Chunyan Miao, Xiaoou Tang, and Dacheng Tao. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue.

A. Details on the WiC dataset

Figure 6 illustrates two instances from the validation split of the WiC dataset. In the first example, the target noun "operation" in both c_1 and c_2 refers to a process of operating machinery, resulting in a positive label (T). In contrast, in the second example, the target verb "excite" denotes a physiological activity in c_1, while representing a mental reaction in c_2, justifying the negative label (F).

Table 9 provides an overview of the dataset configuration based on split and part-of-speech.

```
w/POS: operation/N
c1: The plane’s operation in high
winds.
c2: The power of its engine
determines its operation.
label: T

w/POS: excite/V
c1: Excite the neurons.
c2: The fireworks which opened the
festivities excited anyone present.
label: F
```

Figure 6: Examples of the WiC instances.

Split	# of instances	Nouns	Verbs
Training	5,428	49%	51%
Validation	638	62%	38%
Test	1,400	59%	41%

Table 9: Overview of the WiC dataset.

B. Zero-shot Template

The following template was employed to gauge the zero-shot baseline performances of GPT-3 and GPT-4.

```
Your task is to identify if the
meanings of the target word "{word}"
in the following c1 and c2 sentences
correspond to identical meanings or
not.
That is, it is the Word-in-Context
task.

Please simply answer T,
if the meanings correspond to
identical meanings.
Otherwise, simply answer F.
[Question]
Target word: {word}
c1: {c1}
c2: {c2}
Answer:
```

Figure 7: Template for zero-shot prompting.

C. Other Sample Descriptions

Below are two pairs of description examples: one generated by GPT-3 in a contrast style and the other by GPT-4 in a direct style.

C.1. Target word: induce/V

In this example, the description generated by GPT-3 successfully captures the nuanced differences presented in the context sentences, while GPT-4 erroneously focuses on a higher-level commonality.

```
w/POS: induce/V
c1: Induce a crisis.
c2: The ads induced me to buy a VCR.
label: F
```

Figure 8: WiC instance for induce/V.

LLM	Description
GPT-3	In the first sentence, "induce" implies causing or bringing about a crisis. In the second sentence, "induce" suggests persuading or influencing someone to buy a VCR. The core senses of "induce" in these two sentences are causing and persuading. Overall, "induce" can mean causing or persuading depending on the context.
GPT-4	(Yes,) the word "induce" in both sentences carries the same meaning. In both contexts, "induce" is used to mean "cause" or "bring about".

Table 10: Descriptions generated by GPT-3 and GPT-4 for induce/V.

C.2. Target word: wall/N

This is an example of the opposite case: GPT-4 led to a correct prediction, while GPT-3 wrongly interpreted the meaning of "wall" in c_2.

```
w/POS: wall/N
c1: He ducked behind the garden wall
and waited.
c2: The wall followed the road.
label: T
```

Figure 9: WiC instance for wall/N.

LLM	Description
GPT-3	In the first sentence, "wall" is used to describe a physical structure that provides cover. In the second sentence, "wall" is used metaphorically to describe something that moves alongside the road. The contrast between the two senses of "wall" highlights its versatility in conveying both physical and abstract concepts.
GPT-4	(Yes,) the word "wall" in both sentences refers to a physical barrier or partition.

Table 11: Descriptions generated by GPT-3 and GPT-4 for wall/N.