Traffic Light or Light Traffic? Investigating Phrasal Semantics in Large Language Models

Anonymous ACL submission

Abstract

Phrases are fundamental linguistic units through which humans convey semantics. This study critically examines the capacity of API-based large language models (LLMs) to comprehend phrase semantics, utilizing three human-annotated datasets. We assess the performance of LLMs in executing phrase semantic reasoning tasks guided by natural language instructions and explore the impact of common prompting techniques, including few-shot demonstrations and Chain-of-Thought reasoning. Our findings reveal that LLMs greatly outperform traditional embedding methods across the datasets; however, they do not show a significant advantage over fine-tuned methods. The effectiveness of advanced prompting strategies shows variability. We conduct detailed error analyses to interpret the limitations faced by LLMs in comprehending phrase semantics.

1 Introduction

Understanding phrase semantics presents unique challenges for Artificial Intelligence due to the compositionality and ambiguity inherent in natural languages. The same set of words can yield phrases with vastly different meanings, exemplified by phrase pairs such as “traffic light” versus “light traffic” and “hard drive” versus “drive hard”. Understanding phrase semantics (Korkontzelos et al., 2013; Turney, 2012; Asaadi et al., 2019; Pavlick et al., 2015) involves reasoning the structure and meaning of the combined words and aligning these symbol with real-world concepts commonly understood by humans. Common tasks used to assess a model’s ability to comprehend phrase semantics include identifying similar phrases or estimating the semantic similarity between phrase pairs.

Traditional methods for representing phrase semantics generally rely on either embedding-based strategies or fine-tuning approaches. Embedding-based methods (Wang et al., 2021; Joshi et al., 2020; Reimers and Gurevych, 2019; Gao et al., 2021; Yu and Dredze, 2015) generate vector representations for phrases, using vector similarity as a proxy for semantic relatedness. However, this approach often oversimplifies the complex relationships between phrases. On the other hand, fine-tuning methods (Wang et al., 2021; Pham et al., 2023) adapt a pretrained model (like BERT) to a specific task, such as classification or regression, but require extensive task-specific data.

Large language models (LLMs) excel by generalizing to new tasks with minimal instructions, an advantage in data-scarce situations. Innovations such as in-context few-shot learning (Dong et al., 2022; Cahyawijaya et al., 2024) and Chain-of-Thought prompting (Wei et al., 2022b; Nguyen et al., 2023; Ling et al., 2024) have shown promise in transferring knowledge to novel tasks and enhancing reasoning skills. The potential of LLMs in phrase semantic understanding remains underexplored and warrants investigation.

This study evaluates the performance of LLMs in phrase semantics tasks by focusing on their instruction-following abilities. Our contributions are threefold: (1) We analyze LLMs’ processing of phrase semantics, particularly their alignment with human linguistic understanding – a crucial yet underexplored aspect; (2) We benchmark LLMs against a range of traditional methods to determine if LLMs can effectively perform semantic reasoning tasks solely based on instruction-following and various prompting techniques; (3) We provide a detailed error analysis to elucidate the reasons behind the failures of current LLMs in specific scenarios.

2 Methods

We compare three types of models to evaluate their performance in understanding phrase semantics:

Embedding similarity based methods: these models encode phrases into dense vectors and mea-

Task	Prompt	Ground-truth	Output
Turney +CoT	Given 5 word candidates (delimited by |), you are tasked to answer which word is most semantically similar to a given phrase. Please respond with the word only, without any explanation. Among 5 words "proverb | sabre | stockpiling | moonlighter | jigsaw", which word is most similar to the phrase "saber saw"? Let's think step by step, briefly explain the meaning of each word/phrase, and end the response with a new line that only contains the answer.	jigsaw	- Sabre: A type of sword with a curved blade. - Stockpiling: Accumulating a large amount of goods or materials. - Jigsaw: A tool used for cutting precise curves in wood, metal, or other materials. - Moonlighter: A person who has a second job in addition to their main one, typically done at night. - Proverb: A short, well-known saying that expresses a general truth or piece of advice. Jigsaw
BiRD +CoT	Given two phrases, you are tasked to predict the semantic relatedness of them. Now, given two "access road" and "road access", what is their semantic relatedness? Please only output a number within the range [0, 100]. Let's think step by step, briefly explain the meaning of each phrase, and end the response with a new line that only contains the estimated number of semantic relatedness.	0.684	"Access road" typically refers to a road that allows people to reach a main road or destination. "Road access" generally means the ability to reach or use a road. It implies that there is a way to get onto the road from a property, area, or another road. Given the close relationship between the physical entity of an access road and the concept of having road access, the relatedness is high. 95.
PiC +CoT	Given a pair of phrases sharing the same context, you are tasked to answer whether the two phrases have the same semantic meaning. Now here is the test: - On the local circuit, Faiz emerged champion in 3 events and obtained medals in both the Foil and Sabre U-12 events. - On the community competitions, Faiz emerged champion in 3 events and obtained medals in both the Foil and Sabre U-12 events. Please answer whether the phrase pair "local circuit" and "community competitions" has the same semantic meaning. Let's think step by step, briefly explain the similarity/difference between the two phrases, and end the response with only Yes or No).	yes	The phrase "local circuit" typically refers to a series of competitions that are held within a specific local area. It implies a sequence of events that participants travel around to compete in. On the other hand, "community competitions" suggests contests or events that are organized within a community, which could be a town, city, or a smaller locality. This term emphasizes the community aspect and may not necessarily imply a series of events or the same structured nature as a circuit. While both phrases indicate events that are local and not of a larger national or international scale, "local circuit" suggests a more organized series of events, whereas "community competitions" could be less structured and not part of a circuit. No.

Table 1: Examples of instructions with CoT and the corresponding `GPT-4-Turbo` outputs on three tasks.

sure semantic relatedness through vector similarity. For classification, the phrase with the highest vector similarity is selected, while for regression tasks, the similarity score itself is used as the prediction.

Fine-tuning based methods: these approaches involve adding a prediction head to a pretrained language model, fine-tuning it to capture the nuances of specific semantic tasks. In this study, results are reported for a classification dataset, PiC-PS, where a training split is available.

Natural language instruction based methods: Leveraging instruction-tuned models like GPT-3 (Brown et al., 2020b) and FLAN (Chung et al., 2022), these methods guide language models by providing task instructions as part of the input. Techniques such as in-context few-shot demonstration (Brown et al., 2020a) and chain-of-thought prompting (CoT) (Wei et al., 2022c,a) are employed to enhance models' task-specific reasoning capabilities. When CoT prompting is enabled, we guide the model to execute semantic reasoning, i.e. explain the meaning of each phrase and their differences before making predictions, which is different from math or logical reasoning tasks (Cobbe et al., 2021; Srivastava et al., 2023) that care about the soundness of the reasoning procedure.

3 Experiment Setting

To evaluate phrase semantics understanding, three datasets were utilized:

Turney (Turney, 2012): Comprising 2,180 examples derived from WordNet, this dataset pairs one bigram phrase (query phrase) with five unigrams, where one unigram is a synonym of the bigram, and the others are related distractors. As a multiple-choice classification task, it challenges the model to identify the synonym from the unigrams, with performance measured by accuracy.

BiRD (Asaadi et al., 2019): This is a bigram relatedness dataset using the Best–Worst Scaling (BWS) annotation. In contrast to common annotation techniques that use a discrete 0 to 5 scale, BWS is expected to provide more reliable and discriminating annotations. Specifically, it employs comparative annotations, where annotators are given multiple items at a time and asked to select which item is the best and worst. BiRD consists of 3,345 pairs of bigram phrases, each including a pair of phrases and a human rating of similarity between 0 and 1. The Pearson correlation coefficient between the model predictions and human judgments is reported.

PiC-PS (Phrase-in-Context - Phrase Similarity) (Pham et al., 2023): PiC-PS is a binary clas-

Model	Turney (Accuracy)		BiRD (PearsonCC)	
Baselines (Embedding)				
BERT	42.5		0.445	
BERT-Large	42.5		0.476	
SpanBERT	20.8		0.257	
SpanBERT-Large	41.0		0.330	
Phrase-BERT	57.1		0.689	
LLMs (Instruction)		+CoT		+CoT
GPT-3.5T, 0-shot	87.0	83.4	0.719	0.689
GPT-3.5T, 2-shot	87.0	76.0	0.637	0.639
GPT-3.5T, 4-shot	**87.6**	78.6	0.636	0.632
GPT-4T, 0-shot	86.1	86.2	0.750	0.724
GPT-4T, 2-shot	86.8	86.4	0.759	0.733
GPT-4T, 4-shot	87.5	86.6	**0.761**	0.738

Table 2: Model performance on Turney and BiRD. Best score in each group/task is underlined/**boldfaced**.

Model	PiC-PS Accuracy	
Baselines	**M1: Embedding**	**M2: Fine-tuned**
BERT	64.1	68.9
SpanBERT	64.0	66.9
SpanBERT-Large	66.3	69.3
SentenceBERT	60.3	62.6
PhraseBERT	63.4	66.1
SimCSE	62.5	66.7
LLMs	**M3: w/o CoT**	**M4: w/ CoT**
GPT-3.5T, 0-shot	62.1	66.3
GPT-3.5T, 2-shot	70.8	62.0
GPT-3.5T, 4-shot	69.9	63.4
GPT-3.5T, 8-shot	67.6	63.3
GPT-4T, 0-shot	**73.7**	64.4
GPT-4T, 2-shot	72.8	70.7
GPT-4T, 4-shot	73.5	72.3
GPT-4T, 8-shot	72.7	72.0

Table 3: Model performance on PiC-PS.

sification task and it contains 2,000 examples for testing. The model is tasked to predict whether two noun phrases are semantically similar or not in the same context. The challenge of this task is that, without taking the contextual information into account, very likely the two phrases are interpreted as synonyms. The accuracy score is reported.

For Turney and BiRD, we replicate baselines following (Wang et al., 2021). For PiC-PS, we follow (Pham et al., 2023) and reuse the reported scores. We select two API-based LLMs: `GPT-3.5-Turbo` (*gpt-3.5-turbo-1106*) and `GPT-4-Turbo` (*gpt-4-1106-preview*). Sample data and prompts are detailed in Table 1 and 4. If the OpenAI API fails to produce valid responses (e.g., system refusal messages), we retry up to 10 times or assign default values (-1 for BiRD and *INVALID* for others). The potential of few-shot learning to enhance LLM performance was also considered. Given the absence of a development set for Turney and BiRD, 10 examples from each test split were allocated as a makeshift development set, leaving 2,170 and 3,335 examples for final testing, respectively.

4 Results and Analyses

4.1 Main Results

The results of phrase semantic similarity are presented in Table 2 and 3. The evaluation of phrase semantic similarity across the datasets shows that instruction-based methods employing LLMs set new benchmarks (Turney from 57.1 to 87.6, BiRD from 0.689 to 0.761, PiC-PS from 69.3 to 73.5). However, the highest score on each dataset comes with different models or configurations.

On the Turney benchmark, the previous best accuracy score (57.1) was achieved by an embedding-based method using Phrase-BERT and cosine similarity. Both LLMs under the basic prompting significantly outperform Phrase-BERT (+30 points), primarily due to enhanced language understanding capabilities. Furthermore, more improvements can be attained (GPT-3.5T +0.6, GPT-4T +1.4, no CoT) by providing few-shot examples within the context, allowing the model to glean more task-specific information for refining its predictions.

BiRD directly evaluates the alignment between human annotators and models on phrase similarities, which is measured by Pearson correlation coefficient. It may favor embedding-based methods since they were trained by optimizing the similarities between phrases (Wang et al., 2021). LLMs output similarities in natural language, similar to the way that humans annotate the data (Asaadi et al., 2019). The result shows that LLMs still outrun baselines by a clear margin. Similar to Turney results, few-shot examples show beneficial impacts on GPT-4T but CoT hurts both models. One possible reason is that the longer prompts potentially cause "distractions" to the models. For example, we observe many more bad-format cases for GPT-3.5T when CoT prompting is enabled.

PiC-PS is a context-dependent test. The task is arguably more challenging as the difference between them can be subtle and models need to comprehend the context for support. The previous best score (69.3) is achieved by a fine-tuned SpanBERT-Large with 7k training examples. The best LLM setting only attains a minor improvement (73.7). The few-shot demonstration benefits `GPT-3.5-Turbo` the most when two examples are used, but it worsens the performance of

GPT-4-Turbo. Additionally, the result shows that the CoT prompting leads to degradation in general. The models appear to give reasonable interpretations and only few invalid responses are detected.

4.2 Error Analyses

LLMs demonstrate state-of-the-art performance across three phrase semantics benchmarks without any fine-tuning. However, advanced prompting techniques yield inconsistent results. Our goal here is to conduct thorough analyses to discern the reasons behind these failures.

Failure of CoT: We have observed that semantic CoT prompting generally impacts performance negatively. A benefit of CoT prompting is its ability to make the model's reasoning processes transparent, allowing for detailed examinations of its predictions. In this study, we analyze 50 Turney incorrect predictions made by GPT-4-Turbo, using 4-shot and CoT to investigate the behaviors of LLMs.

Our analysis reveals three primary categories of errors: **1. Erroneous selection despite correct analysis (27/50)**: The model correctly explains individual phrases – like "inkpot" as a small container for ink and "vessel" as a container for holding liquids—but fails to select the most relevant phrase. For instance, despite explaining both terms accurately, the model erroneously selects "vessel" as most similar to "ink bottle". **2. Unknown concepts (13/50)**: For example, it does not understand the phrase "magic eye" as referring to a photoelectric cell or a tuning indicator in radios. Thus the model mistakenly associates it with "thaumaturgy" (magic spells) instead of "photocell". **3. Label ambiguity (7/50)**: Phrases with multiple meanings pose significant challenges. "Small beer" can denote either a minor issue or a type of beer. The model struggles to differentiate whether "brew" or "trivia" is more appropriate, given both could contextually align with "small beer". Additional error types account for the remaining cases (3/50).

Alignment to human semantic preference: We analyze the distribution of phrase-pair similarities as assessed by human annotators and two LLMs, depicted in Figure 1. The similarities assigned by human annotators typically follow a normal distribution. GPT-3.5-Turbo, on the other hand, frequently assigns a similarity value of 65.0 – observed 1,725 times – which may suggest a training bias. This atypical tendency is also noted under other test conditions. GPT-4-Turbo demonstrates a broader range of similarity assignments, yet both models tend to assign higher similarity scores than those given by human annotators. The distributions observed in CoT variants are similarly skewed. These findings indicate that LLMs are not yet fully aligned with human preferences regarding phrase similarity assessments.

Failure to follow instructions: In our tests, there were relatively few instances (less than 10 per test) where failures were due to invalid output formats. However, when Chain-of-Thought (CoT) prompting was enabled, the performance of GPT-3.5-Turbo on the Turney dataset deteriorated significantly. Specifically, the incidence of invalid outputs increased dramatically from 0/5/7 to 825/1,354/1,224 under the 0/2/4-shot settings. Moreover, in 37/159/136 of these cases, the model merely repeated the query phrases. The model often produced a sentence (e.g., "the word most similar to the phrase 'elephant bird' is: aepyornis") instead of the expected single candidate phrase, necessitating post-processing (results shown in Table 2). Despite changes in prompt settings, GPT-4-Turbo consistently generated valid outputs.

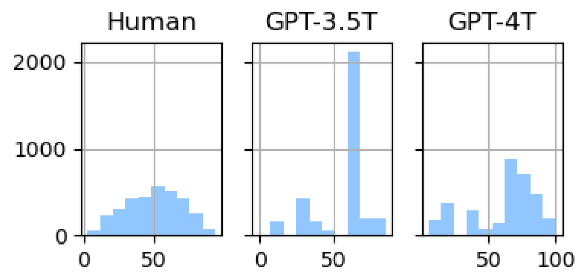

Figure 1: Histograms of BiRD phrase-pair similarities assigned by human annotators, GPT-3.5-turbo and GPT-4-turbo (4-shot demonstration, no CoT prompting).

5 Conclusion

We investigated the performance of LLMs on phrase semantics to assess their alignment with human preferences. Despite strong performance across benchmarks, our findings reveal significant opportunities for improvement. Key areas include ensuring consistent adherence to instructions, crafting better prompts to enhance semantic reasoning, and integrating external knowledge to resolve ambiguities in phrase interpretation. This study inspects the alignment of LLMs with human understanding through the lens of phrase semantics, shedding light on the specific challenges and limitations these models face. It also inspires further exploration into diverse perspectives to assess the cognitive differences between AI agents and humans.

6 Limitations

Model Dependency: The research only utilizes two API-based large language models: `GPT-3.5-Turbo` and `GPT-4-Turbo`. They represent a specific subset of available LLMs, and their findings may not generalize to other models or newer versions.

Dataset and Evaluation Constraints: Our experiments were conducted using three specific datasets, which may not encompass the full spectrum of phrase semantics diversity. The inherent ambiguity in phrase semantics can render benchmarking both challenging and debatable. Additionally, the BiRD dataset relies on human ratings of phrase pairs, introducing a degree of subjectivity and potential biases in the annotations.

Prompt Dependence: The study heavily relies on the effectiveness of prompting techniques, including few-shot demonstrations and CoT prompting. Models may be sensitive to the exact wording and structure of prompts, which can lead to variability in model performance.

References

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516.

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

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. 2020b. Language models are few-shot learners. NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2024. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36.

Hoang Nguyen, Ye Liu, Chenwei Zhang, Tao Zhang, and S Yu Philip. 2023. Cof-cot: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain nlu tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12109–12119.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430.

Thang Pham, Seunghyun Yoon, Trung Bui, and Anh Nguyen. 2023. Pic: A phrase-in-context dataset for phrase understanding and semantic search. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–26.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of artificial intelligence research*, 44:533–585.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10837–10851.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242.

A Appendix

Examples show the complete collection of prompts used in this study and the corresponding model responses from GPT-4-Turbo.

<table>
<tr><th>Task</th><th>Prompt</th><th>Ground-truth</th><th>Output</th></tr>
<tr><td>Turney</td><td>Given 5 word candidates (delimited by |), you are tasked to answer which word is most semantically similar to a given phrase.
Please respond with the word only, without any explanation.
Among 5 words "device | keypad | nihilism | vestal | machine", which word is most similar to the phrase "computer keyboard"?</td><td>keypad</td><td>keypad</td></tr>
<tr><td>Turney
+ 2-shot</td><td>Given 5 word candidates (delimited by |), you are tasked to answer which word is most semantically similar to a given phrase.
Please respond with the word only, without any explanation.
For example:
- Given phrase "street person" and words "thoroughfare | individual | dosser | sectarianism | paraquat", the answer is: dosser
- Given phrase "good story" and words "advantage | sportsmanship | cloudburst | funny | narrative", the answer is: funny
Among 5 words "device | keypad | nihilism | vestal | machine", which word is most similar to the phrase "computer keyboard"?</td><td>keypad</td><td>keypad</td></tr>
<tr><td>Turney
+ 2-shot
+ CoT</td><td>Given 5 word candidates (delimited by |), you are tasked to answer which word is most semantically similar to a given phrase.
For example:
- Given phrase "street person" and words "thoroughfare | individual | dosser | sectarianism | paraquat", the answer is: dosser
- Given phrase "good story" and words "advantage | sportsmanship | cloudburst | funny | narrative", the answer is: funny
Among 5 words "device | keypad | nihilism | vestal | machine", which word is most similar to the phrase "computer keyboard"?
Let's think step by step, briefly explain the meaning of each word/phrase, and end the response with a new line which only contains the answer.</td><td>keypad</td><td>- Vestal: Referring to or characteristic of a vestal virgin; chaste, pure, or related to the Roman goddess Vesta.
- Machine: An apparatus using mechanical power and having several parts, each with a definite function and together performing a particular task.
- Nihilism: The rejection of all religious and moral principles, often in the belief that life is meaningless.
- Device: A thing made or adapted for a particular purpose, especially a piece of mechanical or electronic equipment.
- Keypad: A set of buttons arranged in a block or "pad" which bear digits, symbols, or alphabetical letters and are used for entering data into a computer system or other device.
keypad</td></tr>
<tr><td>BiRD</td><td>Given two phrases, you are tasked to predict the semantic relatedness of them. Note that you are granted to have the capability to predict semantic relatedness between two specific terms.
Now, given two "access service" and "accessible location", what is their semantic relatedness?
Please only output a number within the range [0, 100].</td><td>0.465</td><td>70</td></tr>
<tr><td>BiRD
+ 2-shot</td><td>Given two phrases, you are tasked to predict the semantic relatedness of them. Note that you are granted to have the capability to predict semantic relatedness between two specific terms.
For example:
- The semantic relatedness of "world class" and "class world" is: 42.0
- The semantic relatedness of "world class" and "premier" is: 36.4
Now, given two "access service" and "accessible location", what is their semantic relatedness?
Please only output a number within the range [0, 100].</td><td>0.465</td><td>58.7</td></tr>
<tr><td>BiRD
+ 2-shot
+ CoT</td><td>Given two phrases, you are tasked to predict the semantic relatedness of them. Note that you are granted to have the capability to predict semantic relatedness between two specific terms.
For example:
- The semantic relatedness of "world class" and "class world" is: 42.0
- The semantic relatedness of "world class" and "premier" is: 36.4
Now, given two "access service" and "accessible location", what is their semantic relatedness?
Please only output a number within the range [0, 100].
Let's think step by step, briefly explain the meaning of each phrase, and end the response with a new line which only contains the estimated number of the semantic relatedness. It does not have to be a precise numerical score. Simply let me know a rough estimate based on your understanding.</td><td>0.465</td><td>The phrase "access service" typically refers to the ability to use or obtain a service, which could be anything from internet access to customer support. It implies the provision of a service that can be used or reached.
On the other hand, "accessible location" generally means a place that can be easily reached or entered, often used in the context of being user-friendly for individuals with disabilities or simply easy to get to for the general public.
While both phrases involve the concept of accessibility, they apply to different contexts one to services, and the other to physical locations. However, there is a thematic overlap in the idea of being able to reach or use a service or location.
Considering this thematic overlap, but also the difference in application (service vs. location), I would estimate the semantic relatedness of the two phrases to be moderate. Here's the estimated number for their semantic relatedness:
55.0</td></tr>
</table>

Table 4: Prompts used in this study for Turney and BiRD and the corresponding outputs from `GPT-4-Turbo`.

Task	Prompt	Ground-truth	Output
PiC	Given a pair of phrases sharing the same context, you are tasked to answer whether the two phrases have the same semantic meaning. Now here is the test, in the two sentences below: - However, James Alfred was not convicted in this or in a prior case of a similar nature. - However, James Alfred was not convicted in this or in a preceding game of a similar nature. The phrase pair "prior case" and "preceding game" has the same semantic meaning is (only output Yes or No):	no	no
PiC + 2-shot	Given a pair of phrases sharing the same context, you are tasked to answer whether the two phrases have the same semantic meaning. Now here is the test, in the two sentences below: Here are 2 examples: Example#1, in the two sentences below: - Carty attended a Saturday morning acting class at Phildene Stage School from the age of four, which had an attached agency. - Carty attended a Saturday morning acting class at Phildene Stage School from the age of four, which had an accompanying business that represents actors. phrases describe a business connected to the Stage School, with the primary function of representing actors.The phrase pair "attached agency" and "accompanying business that represents actors" has the same semantic meaning is: Yes Example#2, in the two sentences below: - after theo's apparent death, she decides to leave first colony and ends up traveling with the apostles. - after theo's apparent death, she decides to leave original settlement and ends up traveling with the apostles. Both phrases refer to an initial establishment or habitation, but "first colony" carries specific connotations of being part of a series of colonies or a larger colonization effort, whereas "original settlement" is a more neutral term for any first habitation.The phrase pair "first colony" and "original settlement" has the same semantic meaning is: No Now here is the test, in the two sentences below: - However, James Alfred was not convicted in this or in a prior case of a similar nature. - However, James Alfred was not convicted in this or in a preceding game of a similar nature. Please answer whether the phrase pair "prior case" and "preceding game" has the same semantic meaning? Please only respond with Yes or No.	no	no
PiC + 2-shot + CoT	Given a pair of phrases sharing the same context, you are tasked to answer whether the two phrases have the same semantic meaning. Now here is the test, in the two sentences below: Here are 2 examples: Example#1, in the two sentences below: - Carty attended a Saturday morning acting class at Phildene Stage School from the age of four, which had an attached agency. - Carty attended a Saturday morning acting class at Phildene Stage School from the age of four, which had an accompanying business that represents actors. Whether the phrase pair "attached agency" and "accompanying business that represents actors" has the same semantic meaning? Let's think step by step: Both phrases describe a business connected to the Stage School, with the primary function of representing actors. So the answer is: Yes Example#2, in the two sentences below: - after theo's apparent death, she decides to leave first colony and ends up traveling with the apostles. - after theo's apparent death, she decides to leave original settlement and ends up traveling with the apostles. Whether the phrase pair "first colony" and "original settlement" has the same semantic meaning? Let's think step by step: Both phrases refer to an initial establishment or habitation, but "first colony " carries specific connotations of being part of a series of colonies or a larger colonization effort, whereas "original settlement" is a more neutral term for any first habitation. So the answer is: No Now here is the test, in the two sentences below: - However, James Alfred was not convicted in this or in a prior case of a similar nature. - However, James Alfred was not convicted in this or in a preceding game of a similar nature. Please answer whether the phrase pair "prior case" and "preceding game" has the same semantic meaning. Let's think step by step, briefly explain the similarity/difference between the two phrases, and end the response with only Yes or No).	no	The phrase "prior case" refers to a legal matter or situation that has occurred before the current one, suggesting a context of legal proceedings or court cases. On the other hand, "preceding game" implies a sports or recreational activity that took place before another event. The key difference lies in the context implied by "case", which is legal, and "game", which is recreational or competitive. Therefore, the two phrases do not have the same semantic meaning. No.

Table 5: Prompts used for calling LLMs on PiC-PS and the corresponding outputs from `GPT-4-Turbo`.