# Language Models as Knowledge Bases for Visual Word Sense Disambiguation

Anastasia Kritharoula[1,*], Maria Lymperaiou[1,*] and Giorgos Stamou[1]

[1]*Artificial Intelligence and Learning Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens*

## Abstract

Visual Word Sense Disambiguation (VWSD) is a novel challenging task that lies between linguistic sense disambiguation and fine-grained multimodal retrieval. The recent advancements in the development of visiolinguistic (VL) transformers suggest some off-the-self implementations with encouraging results, which however we argue that can be further improved. To this end, we propose some knowledge-enhancement techniques towards improving the retrieval performance of VL transformers via the usage of Large Language Models (LLMs) as Knowledge Bases. More specifically, knowledge stored in LLMs is retrieved with the help of appropriate prompts in a zero-shot manner, achieving performance advancements. Moreover, we convert VWSD to a purely textual question-answering (QA) problem by considering generated image captions as multiple-choice candidate answers. Zero-shot and few-shot prompting strategies are leveraged to explore the potential of such a transformation, while Chain-of-Thought (CoT) prompting in the zero-shot setting is able to reveal the internal reasoning steps an LLM follows to select the appropriate candidate. In total, our presented approach is the first one to analyze the merits of exploiting knowledge stored in LLMs in different ways to solve WVSD.

## Keywords

Visual Word Sense Disambiguation, Multimodal Retrieval, Large Language Models, Language Models as Knowledge Bases, Reasoning in Large Language Models

## 1. Introduction

Visual Word Sense Disambiguation (VWSD) is a newly introduced task, where the most appropriate candidate image has to be retrieved given an ambiguous target word within a specific context [1]. For example, given the ambiguous target word "andromeda" within the context "tree", the phrase "andromeda tree" is formulated. The word "andromeda" itself may have different meanings, referring to a constellation, flower, reptile or fish, and this ambiguity is mitigated via the given context "tree". However, the presence of other tree-related candidate images complicates the choice of the golden "andromeda tree" image.

VWSD can be viewed as a multimodal extension of textual word sense disambiguation, which has received some considerable attention in recent years [2]. Resolving polysemy has been approached in several ways, with one of the most prominent being knowledge-based

disambiguation. According to it, knowledge graphs representing linguistic relationships such as WordNet [3] for English and BabelNet [4] for multilingual disambiguation are exploited to frame the meaning of an ambiguous word based on its context, alleviating the need for training during the disambiguation process. More specifically, both WordNet and BabelNet knowledge graphs utilize synsets, i.e. sets of synonyms which express different meanings within different contexts; hence, this structure is inherently related to the problem of sense disambiguation, since several possible meanings are already explicitly encoded within each knowledge graph.

Nevertheless, given the ever-increasing amounts of linguistic information that need to be stored in a knowledge base, structures such as WordNet and BabelNet naturally face scalability limitations: novel words, contexts or languages demand high-quality manual engineering of the related knowledge graphs to accommodate tasks such as word sense disambiguation in a viable way. For example, given the example phrase "andromeda tree", a knowledge base lacking the target word "andromeda" or the context "tree" (with respect to "andromeda") would inevitably lead to disambiguation failure. To this end, the introduction of the *Language Models as Knowledge Bases (LM-as-KB)* paradigm [5] opened a whole new world of possibilities for knowledge representation. The expressive power of language models allows the extraction of rich relational knowledge, similar to knowledge obtained when querying a structured knowledge graph. Despite recent research advancements, the real-world utilization of the LM-as-KB paradigm is still in its infancy, with several challenges and limitations arising [6]. These challenges interrogate the adoption of the LM-as-KB framework in the visiolinguistic (VL) setting [7], even though it is already favored in some tasks, such as Visual Question Answering [8, 9, 10], Visual Commonsense Reasoning [11, 12] and Image Captioning [13].

In this work, we aim to explore the potential of following the LM-as-KB paradigm towards resolving VWSD challenges. Specifically, we focus our contributions on two main directions:

- Ambiguous phrases are enriched with knowledge stored in Large Language Models (LLMs) in a zero-shot fashion to provide more context for multimodal retrieval.

- We explore the conversion of VWSD to textual question-answering (QA) upon which we apply zero-shot and few-shot prompting strategies. We further investigate the usage of Chain-of-Thought (CoT) prompting to obtain explanations for LLM-based decisions.

## 2. Related work

**Language models as Knowledge Bases** is a novel paradigm that harnesses the implicit knowledge stored in neural weights of Language Models (LMs) similar to how explicit knowledge of Knowledge Graphs (KGs) has served related applications [5, 6]. Recent works probe characteristics of different LLM-stored knowledge senses, such as commonsense [14], factual [15] and temporal [16]. Model scale has revealed unprecedented LM capabilities related to knowledge, such as different types of reasoning; there are open research questions related to arithmetic [17], symbolic [18], commonsense [19] and causal reasoning [20] in large LMs, and whether such models merely overfit large amounts of data or are genuinely capable of reasoning similar to humans [21]. More recently, the LM-as-KB paradigm has been favored by the VL community to enhance popular VL tasks [7, 8, 9, 10, 11, 12], as in the case of the current paper.

**Prompting Language Models** Traditionally, LMs are pre-trained on large amounts of text corpora and then fine-tuned on smaller labelled ones to address longstanding challenges in NLP. Scaling-up to several billion parameters (Large Language Models - LLMs) enables emergent model capabilities, which can be communicated via *prompting* [22]. Different prompting strategies have been proposed, such as zero-shot prompting, where the task is explicitly described in natural language, and few-shot prompting, where few exemplars are provided to the LM to guide answer generation [23]. More specifically, in-context learning (ICL) is a novel few-shot paradigm where few sample demonstrations from the dataset are used to retrieve related knowledge stored in the LLM without updating any parameters [24]. This technique is inspired from the way humans learn from analogy [25] and has been successfully applied ever since to serve several NLP tasks [26, 27, 28, 29, 30]. Another human-inspired prompting technique is Chain-of-Thought (CoT), where the LM is encouraged by the prompt phrasing to generate rationales together with the predicted answer [31, 32, 33]. Apart from the aforementioned *discrete* prompt strategies, *soft prompting* sacrifices human-interpretable templates to achieve more advanced knowledge retrieval [34]. In total, prompting can be viewed to be analogous to querying knowledge graphs in the traditional knowledge retrieval scenario.

## 3. Method

The retrieval backbones utilized in all cases are pre-trained VL transformers, which place both images $i$ and textual phrases $t$ in a joint embedding space. The retrieval of the most appropriate image $i$ with respect to a given text phrase $t$ is performed based on a similarity score $score(t, i) = max(sim(t, i))$, which can be obtained using different similarity measures, such as cosine similarity or euclidean/manhattan distance[1]. In the baseline case, no knowledge enhancement is performed, thus image candidates $i$ and original phrases $t$ are passed in the VL transformer of choice. The VL transformers used are CLIP-L (CLIP [35] with ViT [36] large encoder), $CLIP_{LAION}$ (LAION/CLIP ViT-H/14)[37] trained on LAION-2B English subset of LAION-5B, ALIGN [38], and BLIP [39] variants; specifically for BLIP, $BLIP_C$ and $BLIP\text{-}L_C$ refer to variants pre-trained on COCO [40] and using ViT base/ViT large as backbone encoders respectively, while $BLIP_F$ and $BLIP\text{-}L_F$ variants are pre-trained on Flickr30k [41].

Moreover, based on the observations of Dadas [42], there are image candidates $i$ that present high similarity scores with multiple text phrases $t$, thus imposing a majority bias. To this end, we adopt the penalty factor $p(i)$, as introduced in [42], to diminish the similarity scores for such frequently similar candidates $i$, without affecting the similarity scores for the rest of the candidates. In cases $p(i)$ is incorporated in VL retrieval, the similarity score is altered as:

$$score(t, i) = sim(t, i) - p(i) \tag{1}$$

### 3.1. LLM-based phrase enhancement

Following the LM-as-KB paradigm, the original text phrases $t$ are transformed to knowledge-enhanced phrases $t_e$ by harnessing the rich factual knowledge stored in state-of-the-art LLMs.

---

[1]We will be using the term *similarity* to refer to both cosine similarity and euclidean/manhattan distance.

Zero-shot enhancement is achieved via the usage of discrete prompts that follow specific handcrafted templates, as presented in Tab. 1. Those templates are created based on human conversational intuition, encouraging the retrieval of related facts to produce $t_e$.

**Table 1**
Prompt templates used for zero-shot LLM knowledge enhancement of textual phrases $t$.

| Prompt name | Prompt template |
|---|---|
| exact | "<phrase> " |
| what_is | "What is <phrase>?" |
| describe | "Describe <phrase>." |
| meaning_of | "What is the meaning of <phrase>?" |

Emerging capabilities of LLMs, such as multiple types of reasoning, are analogous to model size [33, 43], demonstrating that models beyond a specific scale may contain more advanced knowledge. We examine whether this statement also stands for knowledge enrichment by experimenting with models up to 13B parameters, which correspond to the upper limit our hardware can accommodate, as well as with orders of magnitude larger models (175B parameters) accessible via public APIs (susceptible to pre-defined pricing schemes). More specifically, we use GPT2-XL (1.5B parameters) [44], BLOOMZ-1.7B & 3B [45], OPT-2.7B & 6.7B [46], Galactica 6.7B [47], LLAMA-7B [48] and Vicuna 7B & 13B [49] in the lower-billion scale, as well as the 175B parameter models of GPT-3 [50] and GPT-3.5-turbo[2].

As in the baseline case, we attempt to incorporate the penalty $p(i)$ in the VL retrieval stage, forming the following knowledge-enhanced similarity score for images $i$ and phrases $t_e$:

$$score(t_e, i) = sim(t_e, i) - p(i) \qquad (2)$$

## 3.2. Prompting LLMs with question-answering prompts

We transform VWSD to a question-answering (QA) task by converting the textual phrases $t$ to questions $Q$ that follow handcrafted prompt templates, as presented in Tab. 2. Our experimentation includes both *zero-shot* and *few-shot* prompting. In both scenarios, the LLMs prompted are the Vicuna-13B [49] and the 175B GPT-3.5-turbo. Since the LLMs to be prompted can currently handle textual but not VL inputs, we need to transit to exclusively textual representations for both images $i$ and phrases $t$. Therefore, image captioning techniques are leveraged to achieve this transformation, providing captions $c_i$ for each image candidate $i$. We experiment with greedy search, where only one caption $c_i$ per $i$ is returned, as well as with beam-search multinomial sampling using 5 beams, which returns $k$=10 captions $c_i^k$ per image $i$. The models selected for captioning are GiT-L [51] and BLIP-L [39], which are both based on ViT-large [36] encoder, and ViT-GPT2 [52], which uses ViT-base as the encoder and GPT-2 [44] as the decoder. Throughout this VWSD as QA conversion we aim to leverage different aspects of knowledge stored in the selected LLMs by employing the expressiveness of the designed QA prompts, which allow for advanced flexibility compared to the knowledge enhancement setting described in Sec. 3.1.

Reasoning capabilities of LLMs can be unlocked via the so-called Chain-of-Thought (CoT) prompting [33, 32, 31], where the LLM is asked to output a series of intermediate reasoning

---

[2]https://platform.openai.com/docs/models/gpt-3-5

**Table 2**

QA prompts with and without CoT. 'Beam' and 'Greedy' refer to the corresponding captioning strategy.

| Name | QA Prompt template |
|---|---|
| think (greedy) | "Q: What is the most appropriate caption for the <context>? Answer choices: (A) <caption for image 1> (B) <caption for image 2> ... A: Let's think step by step. " |
| think (beam) | "Q: What is the most appropriate group of captions for the <context>? Answer choices: (A) <captions for image 1 (separated with comma)> (B) <captions for image 2> ... A: Let's think step by step. " |
| CoT | "<think_prompt> <response of LLM with think prompt> Therefore, among A through J, the answer is" |
| no_CoT (greedy) | "Q: What is the most appropriate caption for the <context>? Answer choices: (A) <caption for image 1> (B) <caption for image 2> ... A: " |
| no_CoT (beam) | "Q: What is the most appropriate group of captions for the <context>? Answer choices: (A) <captions for image 1> (B) <captions for image 2> ... A: " |
| choose no_CoT (greedy) | You have ten images, (A) to (J), which are given to you in the form of captions.(A) <caption for image 1>...(J) <caption for image 10> You should choose the image, and therefore the caption that could better represent the <phrase>. What image do you choose? |
| choose no_CoT (beam) | You have ten images, (A) to (J), which are given to you in the form of captions.(A) <captions for image 1 (separated with comma)>...(J) <captions for image 10 (separated with comma)> You should choose the image, and therefore the set of captions that could better represent the <phrase>. What image do you choose? |
| choose CoT (greedy) | You have ten images, (A) to (J), which are given to you in the form of captions. (A) <caption for image 1> ... (J) <caption for image 10> You should choose the image, and therefore the caption that could better represent the <phrase>.<br>Use the following format: Question: What image do you choose? Thought: you should always think about what you choose. Result: the result of your thought. Final Answer: the image that you choose.<br>Begin! Question: What image do you choose? |
| choose CoT (beam) | You have ten images, (A) to (J), which are given to you in the form of a set of captions. (A) captions for image 1 (separated with comma) ... (J) captions for image 10 (separated with comma) You should choose the image, and therefore the set of captions that could better represent the <phrase>.<br>Use the following format: Question: What image do you choose? Thought: you should always think about what you choose. Result: the result of your thought. Final Answer: the image that you choose<br>Begin! Question: What image do you choose? |

steps that logically lead to its answer. Even though CoT prompting has mostly been exploited for multi-step reasoning tasks, it is also able to provide human-understandable *explanations* regarding the choice of the most appropriate candidate $i$ for each phrase $t$. To this end, the first 5 templates and the prompting pipelines of Tab. 2 are adopted from Kojima et al. [31], where a "reasoning" prompt (*"Let's think step by step"*/*think* prompts of Tab. 2) retrieves the reasoning path stored in the LLM, followed by an "answer" prompt (*"Therefore, [...] the answer is"*/*CoT* prompt of Tab. 2) that returns the final answer in an appropriate format. The rest of the templates ("choose" prompt names) are inspired from LangChain prompts [53], following an instructive and descriptive template to encourage the selection of the correct image candidate.

### 3.2.1. Zero-shot prompting

In the *zero-shot* setting, we input a selected prompt from Tab. 2 to the LLM (Vicuna-13B/GPT-3.5-turbo), which generates the answer **A**. The generated **A** can be one of the caption options A-J or a statement that an answer cannot be defined; in any case, **A** is compared with the ground truth caption to determine the success or failure of the zero-shot prompting strategy under

investigation. CoT prompts are accompanied with producing an explanation for choosing **A**.

### 3.2.2. Few-shot prompting

Additionally, we experiment with *few-shot prompting* in place of the previously described *zero-shot prompting*. In this case, we select *k no_CoT/choose no_CoT (greedy)* prompts (Tab. 2), accompanied by their ground truth **A**, thus forming *QA in-context samples*. The number of in-context samples *k* is defined by the user. We design three different ways of selecting the *k* in-context samples. In the *baseline* case (*random*), the *k* samples are randomly selected from the dataset. Nevertheless, since the relevance of selected samples with respect to a chosen sample is significant [54], as well as the sample ordering [55], we design two similarity-based sample selection methods, namely *top* and *inverse-top*. Both selection methods exploit embedding representations of full phrases *t*, which are obtained using ALIGN [38]. The *k* nearest embeddings to a given phrase embedding are efficiently retrieved with the help of cosine similarity. Then, the *top* ordering strategy places the top-1 most similar QA sample first (which corresponds to the most similar full phrase embedding), followed by the 2nd most similar QA sample, until the k-th most similar in the k-th position. On the other hand, the *inverse-top* strategy reverses this order by placing the top-1 most similar QA example in the k-th position.

## 4. Experiments

### 4.1. LLM-based phrase enhancement results

In Tab. 3 we present results regarding LLM-based phrase enhancement using different LLMs-as-KB and the prompts of Tab. 1. More results are provided in Appendix A.

We can easily observe that LLM-based enhancement helps trespassing baseline accuracy and MRR scores when appropriate prompting is used, irrespectively of the usage of VL penalty $p(i)$; "exact" prompt seems to be rather weak towards triggering the necessary knowledge to further drive VL retrieval, as in several cases metrics corresponding to phrases enhanced using the "exact" prompt fall below the baseline performance. On the other hand, "meaning_of" is the more powerful prompt attempted, resulting in non-negligible performance advancements compared to the baselines in most cases. Overall, the combination of GPT-3 phrase enrichment together with $\text{CLIP}_{LAION}$ (with penalty $p(i)$) as the VL retrieval module lead to optimal results.

An interesting observation is that the Vicuna 7/13B models **perform comparably** to GPT-3/3.5 models despite being orders of magnitude smaller. This is an encouraging result suggesting that LLM-based phrase enrichment *may* be successfully performed with more lightweight LLMs that do not adhere to a limiting pricing plan, which would impede large-scale experimentation. However, in most cases that LLMs in the lower-billion scale are employed for knowledge enhancement (Tab. 8), retrieval results struggle to compete with the knowledge-free baselines, revealing a non-negligible association between scale and knowledge-enhancement capabilities.

### 4.2. Question-answering prompting results

In Tab. 4 we present accuracy scores occurring from transforming VWSD to QA using zero-shot (with and without CoT) and few-shot (without CoT) prompting, denoting best results per prompt

**Table 3**

Results for zero-shot LLM-based enhancement using prompts of Tab. 1. **Colored** instances denote overall best results per metric, while **bold** numbers indicate best results for each LLM.

| | | $\text{CLIP}_{LAION}$ | | CLIP-L | | ALIGN | | $\text{BLIP}_C$ | | $\text{BLIP-L}_C$ | | $\text{BLIP}_F$ | | $\text{BLIP-L}_F$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | acc. | MRR | acc. | MRR | acc. | MRR | acc. | MRR | acc. | MRR | acc. | MRR | acc. | MRR |
| **With penalty** | | | | | | | | | | | | | | | |
| | **Baseline** | 71.06 | 81.50 | 62.85 | 76.24 | 68.90 | 80.00 | 60.90 | 74.33 | 64.58 | 77.51 | 60.47 | 73.87 | 69.76 | 80.42 |
| Vicuna7B | exact | 64.58 | 77.20 | 59.18 | 72.46 | 61.99 | 73.62 | 55.94 | 70.98 | 59.83 | 73.52 | 53.56 | 69.18 | 62.20 | 75.68 |
| | what_is | 69.76 | 80.49 | 69.05 | 77.82 | 68.25 | 80.16 | 66.74 | 78.00 | 69.33 | 80.00 | 61.34 | 74.61 | 70.84 | 81.00 |
| | describe | 72.79 | 82.26 | 68.47 | 78.01 | 70.41 | 79.79 | 68.25 | 79.09 | 70.41 | 80.85 | 62.20 | 75.09 | **73.43** | **82.45** |
| | meaning_of | 70.41 | 81.13 | 67.76 | 77.85 | 69.76 | 79.49 | 65.87 | 77.84 | 66.31 | 78.43 | 62.20 | 74.73 | 68.03 | 79.40 |
| Vicuna13B | exact | 67.60 | 79.48 | 62.12 | 74.49 | 64.15 | 75.60 | 60.69 | 73.98 | 65.01 | 77.46 | 54.86 | 69.91 | 66.74 | 78.74 |
| | what_is | 72.14 | 81.69 | 69.98 | 77.63 | 70.84 | 80.58 | 67.82 | 79.01 | 69.11 | 79.87 | 59.18 | 73.48 | **73.43** | **82.91** |
| | describe | 69.98 | 80.70 | 63.28 | 77.01 | 66.74 | 76.85 | 61.12 | 74.73 | 66.95 | 78.57 | 56.16 | 71.10 | 67.17 | 79.11 |
| | meaning_of | 70.63 | 81.33 | 68.26 | 78.25 | 70.63 | 79.83 | 67.82 | 78.70 | 69.76 | 80.66 | 61.99 | 75.08 | 71.71 | 81.80 |
| GPT-3.5 | exact | 64.36 | 75.38 | 60.18 | 72.73 | 62.42 | 74.43 | 57.02 | 70.78 | 59.18 | 72.32 | 52.92 | 67.40 | 63.07 | 74.65 |
| | what_is | 70.63 | 81.46 | 69.35 | 80.51 | 70.41 | 81.42 | 67.60 | 78.56 | 68.47 | 79.67 | 60.91 | 74.30 | 71.71 | 82.02 |
| | describe | 73.22 | 82.50 | 69.28 | 80.31 | 73.22 | 82.73 | 69.33 | 79.90 | 70.41 | 80.80 | 59.83 | 73.65 | 70.63 | 81.29 |
| | meaning_of | 73.65 | 82.71 | 69.06 | 80.55 | 70.41 | 81.38 | 66.52 | 78.59 | 66.52 | 79.16 | 58.53 | 73.31 | 69.98 | 81.46 |
| GPT-3 | exact | 68.03 | 78.41 | 64.07 | 76.58 | 66.52 | 78.37 | 60.48 | 73.99 | 64.15 | 76.58 | 59.61 | 72.91 | 65.23 | 77.06 |
| | what_is | 72.35 | 82.19 | 70.73 | 81.57 | 71.71 | 82.27 | 68.25 | 78.93 | 68.90 | 79.91 | 60.48 | 74.24 | 69.11 | 80.25 |
| | describe | 70.63 | 81.05 | 68.72 | 80.26 | 72.57 | 82.52 | 64.58 | 76.75 | 68.25 | 79.35 | 61.34 | 74.03 | 69.33 | 80.47 |
| | meaning_of | 73.65 | 83.52 | 69.84 | 81.56 | 74.95 | 84.09 | 66.74 | 78.37 | 71.71 | 81.55 | 62.63 | 75.55 | 72.35 | 82.28 |
| **Without penalty** | | | | | | | | | | | | | | | |
| | **Baseline** | 67.82 | 79.50 | 60.69 | 74.42 | 65.66 | 77.48 | 57.24 | 72.07 | 61.34 | 75.88 | 57.67 | 71.96 | 65.01 | 77.86 |
| Vicuna7B | exact | 61.99 | 75.30 | 57.45 | 70.07 | 58.96 | 71.92 | 52.70 | 68.00 | 56.16 | 71.28 | 50.32 | 66.75 | 60.04 | 74.04 |
| | what_is | 68.47 | 79.28 | 65.58 | 76.70 | 64.58 | 77.93 | 64.79 | 76.50 | 66.52 | 78.11 | 57.88 | 72.77 | 67.17 | 78.98 |
| | describe | **71.06** | **81.17** | 65.87 | 76.16 | 68.25 | 77.91 | 66.09 | 77.78 | 68.90 | 79.63 | 60.69 | 73.97 | 70.19 | 80.69 |
| | meaning_of | 69.11 | 80.13 | 65.80 | 75.77 | 68.25 | 78.11 | 63.28 | 75.79 | 63.07 | 75.96 | 59.18 | 72.88 | 64.36 | 77.09 |
| Vicuna13B | exact | 65.23 | 77.73 | 58.44 | 71.67 | 61.56 | 73.10 | 56.80 | 71.48 | 60.69 | 74.43 | 51.19 | 66.95 | 63.28 | 76.68 |
| | what_is | **70.63** | **80.71** | 66.74 | 75.51 | 68.68 | 78.55 | 65.44 | 77.22 | 68.47 | 79.14 | 58.96 | 73.22 | 68.90 | 80.27 |
| | describe | 67.82 | 79.14 | 60.69 | 75.75 | 64.58 | 74.88 | 57.67 | 72.47 | 62.20 | 75.75 | 52.27 | 68.72 | 63.50 | 76.66 |
| | meaning_of | 68.90 | 80.18 | 65.87 | 76.64 | 67.60 | 78.38 | 65.66 | 77.62 | 66.31 | 78.31 | 58.96 | 73.19 | 68.03 | 79.64 |
| GPT-3.5 | exact | 62.20 | 73.38 | 57.11 | 70.36 | 60.48 | 72.15 | 54.43 | 68.33 | 56.80 | 70.42 | 51.40 | 65.68 | 58.32 | 71.11 |
| | what_is | 69.55 | 80.51 | 65.87 | 78.11 | 67.82 | 79.52 | 64.15 | 75.91 | 65.87 | 77.78 | 58.10 | 72.32 | 68.03 | 79.36 |
| | describe | 72.57 | 81.78 | 66.67 | 78.42 | 70.84 | 81.16 | 65.44 | 77.57 | 69.11 | 80.20 | 58.96 | 72.66 | 67.60 | 79.47 |
| | meaning_of | 72.57 | 82.05 | 67.10 | 79.07 | 68.47 | 79.87 | 63.93 | 77.05 | 65.66 | 78.33 | 63.93 | 72.23 | 68.25 | 80.17 |
| GPT-3 | exact | 66.74 | 77.22 | 61.68 | 74.91 | 64.79 | 76.27 | 58.96 | 71.92 | 60.48 | 74.02 | 55.72 | 70.34 | 62.42 | 75.04 |
| | what_is | 71.06 | 81.12 | 68.15 | 79.38 | 69.55 | 80.22 | 63.28 | 75.56 | 65.01 | 77.40 | 56.59 | 71.54 | 67.82 | 79.03 |
| | describe | 69.55 | 80.15 | 68.25 | 79.81 | 71.27 | 81.21 | 63.93 | 75.81 | 66.31 | 77.74 | 58.96 | 72.62 | 67.17 | 78.93 |
| | meaning_of | **73.00** | **82.89** | 68.96 | 80.26 | 72.57 | 82.29 | 65.87 | 77.56 | 69.55 | 80.26 | 60.26 | 74.26 | 70.41 | 81.09 |

with **color**. In general, there is an obvious discrepancy between the performance of GPT-3.5-turbo and Vicuna-13B, denoting that **model scale does matter** in the VWSD as QA scenario, contrary to the phrase enhancement case (Sec. 4.1): smaller models do not possess the necessary knowledge or reasoning abilities to infer the correct answer from captions in the QA setting, no matter the prompt template deployed or the choice between zero-shot/few-shot strategy. This

finding agrees with the observation of Kojima et al. [31] that LLM reasoning capabilities emerge *at scale.* Moreover, there is no clear pattern whether beam or greedy captioning is more effective towards triggering the necessary knowledge: in the case of GiT-L and BLIP-L there is a clear preference towards greedy decoding in conjunction with CoT & no_CoT prompting. Especially for Vicuna-13B this preference is very distinctive, demonstrating significant performance drops when beam decoding is employed in place of greedy decoding. However, the opposite holds for ViT-GPT2, even though the overall performance of ViT-GPT2 is lower in comparison to the other captioners. The performance becomes even worse when beam decoding is employed with few-shot prompting. At the same time, "choose" prompts do not demonstrate a distinct pattern with respect to the decoding strategy (greedy/beam) as well. Overall, GiT-L (greedy) exhibits the most promising captioning capabilities in the majority of Tab. 4 results, while BLIP-L (greedy) appears more capable in the few-shot prompting setting of Vicuna-13B (**colored** cells).

**Table 4**
Accuracy scores (%) for VWSD as a QA problem with and without CoT prompting.

| Captioner | Zero-shot | | | | Few-shot (random) | Few-shot (top) | Few-shot (inv. top) |
|---|---|---|---|---|---|---|---|
| | no_CoT | CoT | choose no_CoT | choose CoT | no_CoT | no_CoT | no_CoT |
| **GPT-3.5-turbo** | | | | | | | |
| GiT-L (greedy) | 44.49 | **47.30** | **51.84** | **52.27** | **51.19** | **51.40** | **53.56** |
| GiT-L (beam) | 40.82 | 36.50 | 50.54 | 49.68 | 46.12 | 47.83 | 45.61 |
| BLIP-L (greedy) | **47.95** | 43.84 | 49.46 | 44.06 | 48.16 | 48.81 | 50.32 |
| BLIP-L (beam) | 38.01 | 34.13 | 50.97 | 50.97 | 40.91 | 40.49 | 40.49 |
| ViT-GPT2 (greedy) | 28.94 | 25.05 | 32.40 | 29.81 | 31.32 | 31.45 | 28.91 |
| ViT-GPT2 (beam) | 30.24 | 25.92 | 32.83 | 33.05 | 32.03 | 28.73 | 23.64 |
| **Vicuna-13B** | | | | | | | |
| GiT-L (greedy) | **34.34** | **27.65** | 20.52 | 20.52 | 31.89 | 33.63 | 36.30 |
| GiT-L (beam) | 11.02 | 7.91 | 19.44 | 11.23 | < 2 | < 2 | < 2 |
| BLIP-L (greedy) | 30.02 | 23.76 | **20.95** | **21.81** | **35.56** | **36.08** | **36.48** |
| BLIP-L (beam) | 9.41 | 6.27 | 12.74 | 8.64 | < 2 | < 2 | < 2 |
| ViT-GPT2 (greedy) | 21.60 | 21.17 | 17.49 | 15.33 | 24.83 | 24.94 | 26.11 |
| ViT-GPT2 (beam) | 11.45 | 6.91 | 16.85 | 12.74 | 2.81 | 3.89 | 4.75 |

### 4.2.1. Zero-shot prompting and CoT reasoning

According to Tab. 4, accuracy results with zero-shot QA prompting are not always encouraging, especially when compared to LLM-based enrichment results of Tab. 3, which approach state-of-the-art performance. Apart from the fact that proper LLM reasoning is still an open problem [56, 57], one possible fundamental reason for these lower accuracy scores can be the conversion from images to text via captioning: this intra-modality conversion may induce errors and information loss that impacts the final performance. Of course, lower performance will also affect the quality of the produced explanations, when CoT is employed. We delve into this scenario by presenting an example where the usage of CoT leads to an erroneous answer, while the no_CoT case succeeds. In Fig. 1 the candidates for the phrase "tender embrace" are showcased. By employing GiT-L (greedy) as the image captioner, we form the question Q as:

*Q: What is the most appropriate caption for the tender embrace? Answer Choices: (A) a small boat sitting on top of a dock. (B) a group of people walking on a green hill. (C) a student gets a hug from a student. (D) a large fly laying on a rock in the water. (E) the bus stop at the station (F) a train is parked at a station. (G) a crowd of people watching a concert. (H) a train station with a sign on the side of it. (I) a black and red train on a track. (J) a man laying in the sand on top of a surfboard.*

**Figure 1:** Candidate images for the phrase *t* "tender embrace".

In Tab. 5 the results of GPT-3.5-turbo with and without *zero-shot* CoT are demonstrated.

**Table 5**

Example of VWSD as QA with & w/o zero-shot CoT prompting. Questions Q are presented truncated. GPT-3.5-turbo is the selected LLM, as it presented best results for CoT/no_CoT prompts of Tab. 4.

| **no_CoT** |
|---|
| Q: What is the most appropriate [...] |
| A: The most appropriate caption for the tender embrace would be: (C) a student gets a hug from a student. ✔ |
| **CoT** |
| Q: What is the most appropriate [...] |
| A: Let's think step by step. |
| First, we need to understand what the tender embrace is referring to. Once we know that, we can look at the answer choices and find the one that best matches the subject of the photo. Therefore, without further information about the photo, it is not possible to choose the most appropriate caption. Could you please provide more context or information about the photo? |
| Therefore, among A through J, the answer is not applicable without more information about the photo" ✗ |

After observing the captions $c_i$ produced by GiT-L, we confirm that they accurately describe the context, thus being appropriate answer options. Nevertheless, when prompted with CoT in a zero-shot manner, GPT-3.5-turbo is unable to define where "tender embrace" refers to without receiving more information. On the contrary, it successfully returns the right answer when no CoT is used. This discrepancy reveals that even though the knowledge regarding the phrase "tender embrace" exists within the LLM, zero-shot CoT prompting is unable to trigger it. This can be viewed as an inherent problem of CoT prompting, at least in the zero-shot setting, since the only factor that differs between successful and unsuccessful reasoning is the prompt itself.

On the other hand, zero-shot CoT prompting can provide valuable insights in several cases, such as the one presented in Fig. 2 with candidates corresponding to the phrase "metal steel". By again using GiT-L as the captioner, the question Q is formed as:

*Q: What is the most appropriate caption for the metal steel? Answer Choices: (A) a chocolate bar with three sides (B) [unused0] and [unused0] at the concert in 2007 (C) a guitar and a guitar are displayed in front of a speaker. (D) frosty patterns on a window (E) gold in the rocks - - (F) a black piece of metal with a large black square in the middle. (G) a jar of honey on a wooden table. (H) a close up of a metal plate with a pattern of lines. (I) a large white quartz rock with a clear base. (J) gold jewelry from the late 19th century.*
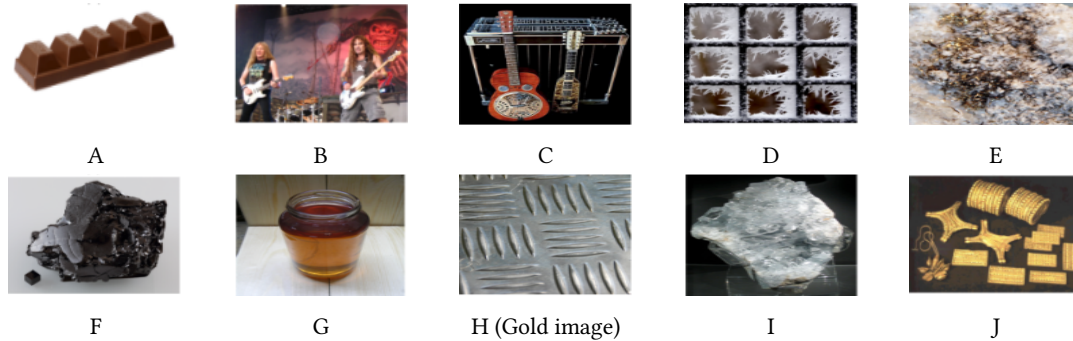


**Figure 2:** Candidate images for the phrase *t* "metal steel".

Then, GPT-3.5-turbo outputs the results presented in Tab. 6, successfully guided by CoT.

**Table 6**
A successful example of VWSD as QA with & w/o zero-shot CoT prompting (GPT-3.5-turbo as the LLM).

| no_CoT |
|---|
| Q: What is the most appropriate [...] |
| A: The most appropriate caption for the metal steel would be (F) a black piece of metal with a large black square in the middle. **X** |

| CoT |
|---|
| Q: What is the most appropriate [...] |
| A: Let's think step by step. |
| First, we need to understand what metal steel is and what its characteristics are. Steel is a hard and strong metal alloy made mainly of iron and carbon. It is often used in construction, machinery, and transportation. Based on this information, the most appropriate caption for metal steel would be (H) a close up of a metal plate with a pattern of lines. This caption describes the texture and appearance of steel, which is often characterized by its distinctive pattern of lines. The other options do not accurately describe steel or its unique qualities. |
| Therefore, among A through J, the answer is (H) a close up of a metal plate with a pattern of lines." ✔ |

This analysis highlights that the current state of CoT reasoning *cannot be entirely trusted*, as the deployment of CoT prompts themselves may drive the LLM towards incorrect reasoning paths. Thus, CoT reasoning for VWSD is not yet mature enough to provide valuable explanations.

Furthermore, regarding the comparison between standalone CoT/no_CoT prompts and "choose" CoT/no_CoT prompts, there is a clear performance advancement when "choose" prompts are leveraged, reaching more than 10% performance gain in the case of BLIP-L (beam) as captioner and GPT-3.5-turbo as the LLM to be prompted. We regard this as an evidence that more descriptive and motivating prompts can better evoke the correct reasoning process of an LLM. Finally, there is no clear indication of whether CoT facilitates performance over no_CoT when combined with the "choose" prompting template, since best results per captioner and LLM of Tab. 4 alternate between "choose" CoT/no_CoT strategies, and are often comparable.

### 4.2.2. Few-shot prompting

In the baseline few-shot setting (few-shot (random)), we randomly select $k$=5 instances to serve as in-context examples. These in-context samples are comprised of $k$ questions Q followed by their ground truth answer choice, as described in Sec. 3.2.2. Results of Tab. 4 denote that few-shot performance using no_CoT prompts is significantly better compared to their zero-shot counterparts, while generally being on par with "choose" prompt results, even though the derived accuracy is close to random choice or even worse in most cases. Therefore, despite the advanced engineering few-shot prompting requires compared to the zero-shot setting, it can be regarded as a more viable choice for retrieving internal reasoning capabilities of LLMs. As for the choice of in-context sample selection and order strategy, the results are not conclusive; similarity-based sample selection (top & inv. top columns of Tab. 4) may result in better (GIT-L greedy), worse (ViT-GPT2 beam) or comparable results to the random baseline accuracy. Moreover, sample ordering (top versus inv. top) presents some variability in accuracy, with each strategy performing better in different cases. Overall, we can state that few-shot prompting calls for more extensive experimentation beyond the purpose of the current work until a standard pattern emerges, while there is a possibility that no pattern can be inferred at all.

We are going to present some qualitative results regarding few-shot prompting. Fig. 3 contains candidates corresponding to the phrase "football goal", while GiT-L (greedy) serves as the captioner. The $k$=5 in-context samples are demonstrated in Tab. 7 followed by the answer **A** generated by GPT-3.5-turbo (in color) corresponding to the given **Q**. In the presented case, in-context prompting achieves in guiding GPT-3.5-turbo to select the correct candidate I.



**Figure 3:** Candidate images for the phrase $t$ "football goal".

## 5. Conclusion

In this work, we explore the conjunction of the Visual Word Sense Disambiguation task and Large Language Models. More specifically, our current paper is the first one to harness the rich knowledge stored in LLMs via different prompting strategies. Following the LLM-as-KB paradigm we managed to boost the performance of baseline visiolinguistic pipelines. Moreover, we examined the potential of unimodal approaches by converting VWSD to a textual question-answer problem, where generated image captions are leveraged as multiple-choice

**Table 7**

Example of few-shot prompting with k=5 in-context samples for disambiguating "football goal" phrase.

**Q:** What is the most appropriate caption for the vanillin compound? Answer Choices: (A) a fossil fossil with a spiral pattern (B) a blue plate topped with pastries and a piece of bread. (C) the bones of the foot are very thin. (D) a spoon full of sugar on a table. (E) the tower of the building is made of brick. (F) a pair of sunglasses with a white background. (G) a close up of a plant with a yellow flower. (H) turmeric powder in a bowl (I) a man with glasses and a tie sitting at a desk. (J) a close up of a bike tire with a yellow tag on it. **A:** (D) a spoon full of sugar on a table.

**Q:** What is the most appropriate caption for the nankeen fabric? Answer Choices: (A) a close up of a white quilt with a tree design on it. (B) a man in a blue jacket and tan pants holding a sword. (C) a collection of brass trombones (D) a close up of a green plant (E) a pine tree in the forest (F) a row of mannequins in black dresses. (G) two boys are standing in front of a red wall. (H) two apples sitting next to each other on a white surface. (I) a cartoon of two women in fancy hats and a mirror. (J) a close up of a woman's face with a round ball on her forehead. **A:** (B) a man in a blue jacket and tan pants holding a sword.

**Q:** What is the most appropriate caption for the olmec indian? Answer Choices: (A) a large stone head in a garden (B) a woman lighting a candle at a table (C) [ unused0 ] is the new champion of the world (D) a table full of food (E) a group of women holding banners in a protest. (F) a woman in a traditional dress stands in front of a tent. (G) a man in a suit sitting in a chair. (H) a clay pot with a small hole on top of it. (I) a close up of three green leaves of a plant (J) a fern in the forest near the lake **A:** (A) a large stone head in a garden

**Q:** What is the most appropriate caption for the charales order? Answer Choices: (A) a man walking past a store with a sign on it. (B) the road to the island (C) the art of the flower (D) a black snake with white spots on it's body. (E) a bird perched on a branch in a tree. (F) a lion roaring in the wild. (G) a man falling from a skateboard (H) a large, thin, green plant with long thin leaves. (I) a close up of a bush with blue berries (J) a mouse in a hollow log. **A:** (H) a large, thin, green plant with long thin leaves.

**Q:** What is the most appropriate caption for the skink lizard? Answer Choices: (A) two lizards on a log with a log in the background. (B) a man walking past a store with a sign on it. (C) the art of the flower (D) a portrait of [ unused0 ], seated at a desk in front of a telescope. (E) the car is a small car that can be found in the museum. (F) a black snake with white spots on it's body. (G) a bird perched on a branch in a tree. (H) a close up of a bush with blue berries (I) a close up of a metal container with a white label on it. (J) a white door with a glass window **A:** (A) two lizards on a log with a log in the background.

**Q:** What is the most appropriate caption for the football goal? Answer Choices: (A) a large stadium with a large field and a large structure. (B) a woman wearing a blue ski mask (C) [ unused0 ] of spain during the fifa world cup brazil group a match between spain and argentina at the estadio santiago bernabeu on june 25, 2010 in santiago, chile. (D) a woman lifting a barbell in a competition. (E) a man in a suit and tie looking at the camera. (F) person is a player for football team (G) [ unused0 ], the brazilian national team, in the 1960s. (H) the empty stand at the old trafford stadium (I) a soccer field with a goal post in the middle (J) the emblem of the football club of the italian football club of the italian football club of the italian football club of the italian football club of the italian football club of the italian football club of the A: (I) a soccer field with a goal post in the middle ✔

candidates. Finally, Chain-of-Thought prompting highlighted human-interpretable explainability aspects tied to the LLM-based knowledge extraction process. Overall, our analysis reveals the importance of model scale towards performing knowledge-related tasks with LLMs.

# Acknowledgments

# References

[1] A. Raganato, I. Calixto, A. Ushio, J. Camacho-Collados, M. T. Pilehvar, SemEval-2023 Task 1: Visual Word Sense Disambiguation, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023.

[2] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4330–4338. URL: https://doi.org/10.24963/ijcai.2021/593. doi:10.24963/ijcai.2021/593, survey Track.

[3] C. Fellbaum, Wordnet: An electronic lexical database (1998).

[4] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250. URL: https://www.sciencedirect.com/science/article/pii/S0004370212000793. doi:https://doi.org/10.1016/j.artint.2012.07.001.

[5] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, 2019. arXiv:1909.01066.

[6] B. AlKhamissi, M. Li, A. Celikyilmaz, M. Diab, M. Ghazvininejad, A review on language models as knowledge bases, 2022. arXiv:2204.06031.

[7] M. Lymperaiou, G. Stamou, The contribution of knowledge in visiolinguistic learning: A survey on tasks and challenges, 2023. arXiv:2303.02411.

[8] A. M. H. Tiong, J. Li, B. Li, S. Savarese, S. C. H. Hoi, Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training, in: Conference on Empirical Methods in Natural Language Processing, 2022.

[9] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, S. Hoi, From images to textual prompts: Zero-shot vqa with frozen large language models, ArXiv abs/2212.10846 (2022).

[10] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, S. Hoi, From images to textual prompts: Zero-shot vqa with frozen large language models, ArXiv abs/2212.10846 (2022).

[11] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL: https://arxiv.org/abs/2209.09513. doi:10.48550/ARXIV.2209.09513.

[12] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola, Multimodal chain-of-thought reasoning in language models, 2023. URL: https://arxiv.org/abs/2302.00923. doi:10.48550/ARXIV.2302.00923.

[13] R. Mokady, A. Hertz, A. H. Bermano, Clipcap: Clip prefix for image captioning, ArXiv abs/2111.09734 (2021).

[14] X. L. Li, A. Kuncoro, J. Hoffmann, C. de Masson d'Autume, P. Blunsom, A. Nematzadeh, A systematic investigation of commonsense knowledge in large language models, 2022. arXiv:2111.00607.

[15] P. Pezeshkpour, Measuring and modifying factual knowledge in large language models, 2023. arXiv:2306.06264.

[16] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, W. W. Cohen, Time-aware language models as temporal knowledge bases, Transactions of the Association for

Computational Linguistics 10 (2022) 257–273. URL: https://aclanthology.org/2022.tacl-1.15. doi:10.1162/tacl_a_00459.

[17] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, H. Hajishirzi, MathQA: Towards interpretable math word problem solving with operation-based formalisms, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2357–2367. URL: https://aclanthology.org/N19-1245. doi:10.18653/v1/N19-1245.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, ArXiv abs/2201.11903 (2022).

[19] P. Bhargava, V. Ng, Commonsense knowledge reasoning and generation with pre-trained language models: A survey, 2022. URL: https://arxiv.org/abs/2201.12438. doi:10.48550/ARXIV.2201.12438.

[20] J. Li, L. Yu, A. Ettinger, Counterfactual reasoning: Do language models need world knowledge for causal understanding?, 2022. arXiv:2212.03278.

[21] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, ArXiv abs/2212.10403 (2022).

[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. arXiv:2107.13586.

[23] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[24] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, Z. Sui, A survey on in-context learning, 2023. arXiv:2301.00234.

[25] P. H. Winston, Learning and reasoning by analogy, Commun. ACM 23 (1980) 689–703. URL: https://doi.org/10.1145/359038.359042. doi:10.1145/359038.359042.

[26] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, J. Wei, Language models are multilingual chain-of-thought reasoners, 2022. arXiv:2210.03057.

[27] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. Dai, A. La, A. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur,

B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, B. Orinion, C. Diao, C. Dour, C. Stinson,
C. Argueta, C. F. Ramírez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-
Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel,
C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Free-
man, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito,
D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko,
D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H.
Lee, D. Schrader, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway,
E. Pavlick, E. Rodola, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer,
E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed,
F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Paras-
candolo, G. Mariani, G. Wang, G. Jaimovitch-López, G. Betz, G. Gur-Ari, H. Galijasevic,
H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura,
H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee,
J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocoń, J. Thompson, J. Wingfield,
J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher,
J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller,
J. U. Balis, J. Batchelder, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman,
J. Guerr, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth,
K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. Omondi, K. Math-
ewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds,
L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella,
L. Lam, L. Noble, L. Schmidt, L. He, L. O. Colón, L. Metz, L. K. Şenel, M. Bosma, M. Sap,
M. ter Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. R.
Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen,
M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu,
M. Ivanitskiy, M. Starritt, M. Strube, M. Swędrowski, M. Bevilacqua, M. Yasunaga, M. Kale,
M. Cain, M. Xu, M. Suzgun, M. Walker, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva,
M. Gheini, M. V. T, N. Peng, N. A. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. Roberts,
N. Doiron, N. Martinez, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer,
N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M.
Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang,
P. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli,
Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. Risco, R. Mil-
lière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand,
R. Novak, R. Sitelew, R. LeBras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee,
R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer,
S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous,
S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Ham-
dan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal,
S. Upadhyay, Shyamolima, Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P.
Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin,
S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen,
T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan,

T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. arXiv:2206.04615.

[28] S. Sia, K. Duh, In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models, 2023. arXiv:2305.03573.

[29] M. Pourreza, D. Rafiei, Din-sql: Decomposed in-context learning of text-to-sql with self-correction, 2023. arXiv:2304.11015.

[30] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, E. Chi, Least-to-most prompting enables complex reasoning in large language models, 2023. arXiv:2205.10625.

[31] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. arXiv:2205.11916.

[32] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. arXiv:2203.11171.

[33] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. arXiv:2205.11916.

[34] G. Qin, J. Eisner, Learning how to ask: Querying LMs with mixtures of soft prompts, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5203–5212. URL: https://aclanthology.org/2021.naacl-main.410. doi:10.18653/v1/2021.naacl-main.410.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. arXiv:2103.00020.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[37] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5b: An open large-scale dataset for training next generation image-text models, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022. URL: https://openreview.net/forum?id=M3Y74vmsMcY.

[38] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision (2021). URL: https://arxiv.org/abs/2102.05918. doi:10.48550/ARXIV.2102.05918.

[39] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. arXiv:2201.12086.

[40] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2015. arXiv:1405.0312.

[41] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78. URL: https://aclanthology.org/Q14-1006. doi:10.1162/tacl_a_00166.

[42] S. Dadas, Opi at semeval 2023 task 1: Image-text embeddings and multimodal information retrieval for visual word sense disambiguation, 2023. arXiv:2304.07127.

[43] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[45] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, 2023. arXiv:2211.01786.

[46] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, 2022. arXiv:2205.01068.

[47] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, Galactica: A large language model for science, 2022.

[48] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[49] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. arXiv:2306.05685.

[50] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[51] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, Git: A generative image-to-text transformer for vision and language, 2022. arXiv:2205.14100.

[52] A. Kumar, The illustrated image captioning using transformers, ankur3107.github.io (2022). URL: https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/.

[53] Langchain prompt templates, ???? URL: https://python.langchain.com/docs/modules/model_io/prompts/prompt_templates/.

[54] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3?, in: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Association for Computational Linguistics, Dublin, Ireland and Online, 2022, pp. 100–114.

URL: https://aclanthology.org/2022.deelio-1.10. doi:10.18653/v1/2022.deelio-1.10.

[55] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8086–8098. URL: https://aclanthology.org/2022.acl-long.556. doi:10.18653/v1/2022.acl-long.556.

[56] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. arXiv:2302.04023.

[57] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. arXiv:2305.04388.

## A. Scale significance for knowledge enhancement

As an extension of Tab. 3, we presents results regarding LLM-enhancements using other LLMs (1.7B to 7B parameters) in Tab. 8.

Compared to Tab. 3, smaller models present more mediocre knowledge enhancement capabilities, frequently scoring below the knowledge-free VL retrieval baseline. Such a result indicates that knowledge enhancement is not a "free-lunch" solution to boost results, but requires tuning several aspects of the process in order to benefit the downstream task. Therefore, despite the encouraging performance of Vicuna7B/13B models, other LLMs in the lower-billion range do not possess the necessary knowledge capabilities to effectively perform knowledge enhancement for VWSD. Therefore, we can conclude that similarly to the QA prompting that demonstrated the need for scale, also prompting for knowledge enhancement is strongly associated with model size. To this end, we verify that smaller models with competitive capabilities, similar to the Vicuna series, need to be develop in order to facilitate applications related -but not limited-to LLM-based knowledge-enhancement.

## B. Reproducibility

Our code can be found in GitHub: https://github.com/anastasiakrith/llm-for-vwsd/.

**Table 8**
Additional results for zero-shot LLM-based enhancement.

| | | CLIP$_{LAION}$ | | CLIP-L | | ALIGN | | BLIP$_C$ | | BLIP-L$_C$ | | BLIP$_F$ | | BLIP-L$_F$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | acc | MRR | acc | MRR | acc | MRR | acc | MRR | acc | MRR | acc | MRR | acc | MRR |
| **With penalty** | | | | | | | | | | | | | | | |
| | **Baseline** | 71.06 | 81.50 | 62.85 | 76.24 | 68.90 | 80.00 | 60.90 | 74.33 | 64.58 | 77.51 | 60.47 | 73.87 | 69.76 | 80.42 |
| BLOOMZ1.7B | exact | 66.52 | 78.50 | 64.92 | 74.50 | 65.87 | 77.42 | 64.58 | 76.28 | 65.66 | 77.13 | 59.18 | 72.70 | 67.39 | 78.67 |
| BLOOMZ1.7B | what_is | 69.98 | 80.67 | 66.74 | 76.41 | 65.44 | 78.92 | 63.28 | 75.95 | 65.23 | 77.68 | 58.32 | 72.06 | 66.52 | 78.30 |
| BLOOMZ1.7B | describe | 73.65 | 83.52 | 69.84 | 80.08 | 74.95 | 81.56 | 66.74 | 78.37 | 71.71 | 81.55 | 62.63 | 75.55 | 72.35 | 82.28 |
| BLOOMZ1.7B | meaning_of | 69.33 | 80.69 | 65.01 | 76.38 | 66.74 | 78.17 | 63.50 | 76.44 | 65.44 | 78.29 | 58.53 | 72.50 | 68.25 | 79.74 |
| OPT2.7B | exact | 71.06 | 81.46 | 62.85 | 76.00 | 68.68 | 75.93 | 61.12 | 74.46 | 64.58 | 77.41 | 60.26 | 73.73 | 69.76 | 80.36 |
| OPT2.7B | what_is | 66.95 | 78.99 | 66.30 | 74.85 | 63.28 | 78.10 | 60.91 | 74.43 | 66.31 | 77.86 | 57.24 | 71.15 | 67.60 | 78.58 |
| OPT2.7B | describe | 68.03 | 79.89 | 66.08 | 74.75 | 64.79 | 78.14 | 61.77 | 74.73 | 66.31 | 77.57 | 57.67 | 71.48 | 68.03 | 79.03 |
| OPT2.7B | meaning_of | 69.11 | 80.56 | 65.25 | 75.60 | 65.66 | 77.45 | 61.99 | 75.35 | 63.93 | 76.88 | 58.32 | 71.65 | 65.44 | 77.69 |
| BLOOMZ3B | exact | 67.82 | 79.31 | 62.99 | 74.59 | 66.52 | 76.18 | 60.48 | 73.13 | 63.28 | 76.00 | 57.02 | 71.23 | 65.66 | 77.49 |
| BLOOMZ3B | what_is | 71.92 | 81.78 | 68.25 | 76.82 | 67.39 | 79.82 | 61.34 | 74.94 | 66.95 | 78.47 | 59.61 | 73.35 | 68.47 | 79.58 |
| BLOOMZ3B | describe | 70.84 | 81.11 | 65.28 | 75.38 | 66.09 | 78.07 | 62.85 | 75.65 | 67.39 | 78.71 | 57.24 | 71.72 | 67.82 | 79.20 |
| BLOOMZ3B | meaning_of | 70.63 | 81.47 | 67.32 | 77.96 | 68.47 | 78.76 | 63.71 | 76.52 | 66.31 | 78.55 | 59.40 | 73.60 | 68.03 | 79.26 |
| OPT6.7B | exact | 62.63 | 75.84 | 62.20 | 75.54 | 67.82 | 79.24 | 60.91 | 74.23 | 64.79 | 77.58 | 59.83 | 73.40 | 69.11 | 79.94 |
| OPT6.7B | what_is | 61.79 | 75.70 | 64.63 | 77.68 | 64.79 | 77.23 | 61.77 | 75.01 | 63.07 | 76.16 | 57.88 | 71.79 | 65.87 | 77.77 |
| OPT6.7B | describe | 64.43 | 76.91 | 65.73 | 78.24 | 65.23 | 77.89 | 61.12 | 74.67 | 63.93 | 77.07 | 56.16 | 71.30 | 66.09 | 78.38 |
| OPT6.7B | meaning_of | 62.17 | 75.84 | 63.61 | 77.19 | 66.74 | 78.47 | 63.28 | 75.93 | 65.44 | 77.43 | 59.83 | 72.96 | 68.03 | 78.75 |
| LLAMA7B | exact | 62.20 | 76.04 | 57.45 | 70.70 | 60.48 | 72.23 | 52.92 | 68.76 | 54.00 | 69.51 | 48.81 | 65.54 | 58.10 | 72.33 |
| LLAMA7B | what_is | 68.90 | 80.48 | 67.10 | 77.26 | 66.95 | 78.93 | 68.03 | 68.03 | 66.95 | 78.48 | 58.75 | 73.29 | 70.84 | 80.84 |
| LLAMA7B | describe | 72.14 | 82.02 | 66.81 | 76.80 | 69.76 | 79.04 | 64.79 | 76.58 | 66.09 | 77.99 | 58.53 | 72.44 | 69.11 | 80.15 |
| LLAMA7B | meaning_of | 67.82 | 80.11 | 62.90 | 76.12 | 65.66 | 76.53 | 60.26 | 74.40 | 62.42 | 75.71 | 54.00 | 69.84 | 65.87 | 78.19 |
| **Without penalty** | | | | | | | | | | | | | | | |
| | **Baseline** | 67.82 | 79.50 | 60.69 | 74.42 | 65.66 | 77.48 | 57.24 | 72.07 | 61.34 | 75.88 | 57.67 | 71.96 | 65.01 | 77.86 |
| BLOOMZ 1.7B | exact | 63.71 | 76.53 | 61.66 | 72.23 | 63.28 | 75.05 | 59.18 | 72.96 | 62.85 | 74.99 | 56.37 | 70.60 | 63.50 | 76.20 |
| BLOOMZ 1.7B | what_is | 67.60 | 79.04 | 65.01 | 75.30 | 63.07 | 77.33 | 59.40 | 73.50 | 62.85 | 76.00 | 56.16 | 70.54 | 65.23 | 77.26 |
| BLOOMZ 1.7B | describe | 67.60 | 79.54 | 62.99 | 74.68 | 66.52 | 76.05 | 59.83 | 74.59 | 62.63 | 76.47 | 53.56 | 69.84 | 63.71 | 77.06 |
| BLOOMZ 1.7B | meaning_of | 66.52 | 78.70 | 64.15 | 73.78 | 64.36 | 77.03 | 60.48 | 74.34 | 61.99 | 76.13 | 56.16 | 70.61 | 65.01 | 77.89 |
| OPT2.7B | exact | 67.82 | 79.47 | 60.26 | 72.77 | 65.66 | 74.15 | 57.45 | 72.19 | 61.12 | 75.77 | 57.24 | 71.68 | 65.01 | 77.90 |
| OPT2.7B | what_is | 65.44 | 77.60 | 62.75 | 72.91 | 61.12 | 75.47 | 59.83 | 73.13 | 61.12 | 74.54 | 53.35 | 68.71 | 63.50 | 76.22 |
| OPT2.7B | describe | 65.87 | 78.09 | 63.89 | 72.95 | 62.20 | 76.31 | 59.83 | 73.28 | 62.20 | 75.17 | 54.43 | 69.86 | 63.28 | 76.28 |
| OPT2.7B | meaning_of | 67.60 | 79.28 | 62.99 | 72.97 | 64.58 | 75.79 | 59.18 | 73.38 | 60.26 | 74.70 | 54.86 | 69.43 | 62.42 | 75.86 |
| BLOOMZ3B | exact | 64.15 | 76.98 | 59.52 | 71.53 | 63.93 | 73.78 | 58.10 | 71.77 | 59.61 | 74.06 | 54.86 | 69.66 | 61.12 | 74.99 |
| BLOOMZ3B | what_is | 69.11 | 80.03 | 65.66 | 75.39 | 62.85 | 77.88 | 61.34 | 74.35 | 65.01 | 77.32 | 57.24 | 71.85 | 68.03 | 79.12 |
| BLOOMZ3B | describe | 68.68 | 79.47 | 62.88 | 73.83 | 63.50 | 76.11 | 60.48 | 73.87 | 62.85 | 76.06 | 54.86 | 70.48 | 65.66 | 77.64 |
| BLOOMZ3B | meaning_of | 68.25 | 79.73 | 64.94 | 75.51 | 66.31 | 77.17 | 61.77 | 74.92 | 62.42 | 76.27 | 57.02 | 71.79 | 65.23 | 77.21 |
| OPT 6.7B | exact | 58.75 | 72.63 | 59.61 | 73.86 | 64.15 | 76.57 | 57.24 | 71.96 | 61.12 | 75.83 | 56.80 | 71.40 | 64.79 | 77.66 |
| OPT 6.7B | what_is | 60.48 | 74.10 | 62.45 | 75.89 | 61.77 | 75.18 | 57.88 | 72.27 | 61.77 | 74.89 | 52.92 | 68.83 | 61.99 | 75.23 |
| OPT 6.7B | describe | 60.74 | 74.28 | 63.12 | 76.19 | 63.28 | 76.26 | 59.40 | 73.03 | 58.96 | 73.86 | 52.92 | 69.18 | 62.63 | 76.13 |
| OPT 6.7B | meaning_of | 59.28 | 73.77 | 62.17 | 76.04 | 63.71 | 76.31 | 52.92 | 74.37 | 61.99 | 75.47 | 55.94 | 70.67 | 65.01 | 77.27 |
| LLAMA7B | exact | 60.91 | 74.66 | 56.16 | 68.98 | 56.80 | 70.76 | 52.05 | 67.33 | 50.54 | 66.81 | 47.95 | 64.17 | 54.86 | 69.79 |
| LLAMA7B | what_is | 67.60 | 79.19 | 64.71 | 75.98 | 66.31 | 77.31 | 64.36 | 76.30 | 64.36 | 77.00 | 56.37 | 71.90 | 68.47 | 79.43 |
| LLAMA7B | describe | 70.41 | 81.20 | 63.77 | 74.97 | 66.74 | 76.99 | 64.15 | 76.25 | 63.50 | 76.40 | 58.10 | 72.58 | 68.25 | 79.36 |
| LLAMA7B | meaning_of | 66.09 | 78.60 | 60.41 | 72.84 | 62.85 | 74.79 | 58.75 | 72.70 | 58.53 | 73.12 | 52.70 | 68.03 | 62.42 | 76.07 |