Retrieval-Augmented Semantic Parsing: Using Large Language Models to Improve Generalization

Anonymous ACL submission

Abstract

Open-domain semantic parsing remains a challenging task, as models often rely on heuristics and struggle to handle unseen concepts. In this paper, we investigate the potential of large language models (LLMs) for this task and introduce Retrieval-Augmented Semantic Parsing (RASP), a simple yet effective approach that integrates external lexical knowledge into the parsing process. Our experiments not only show that LLMs outperform previous encoder-decoder baselines for semantic parsing, but that RASP further enhances their ability to predict unseen concepts, nearly doubling the performance of previous models on out-of-distribution concepts. These findings highlight the promise of leveraging large language models and retrieval mechanisms for robust and open-domain semantic parsing.

1 Introduction

Open-domain semantic parsing involves mapping natural language text to formal meaning representations, capturing the concepts, relations between them, and the contects in which they appear (Oepen and Lønning, 2006; Hajič et al., 2012; Banarescu et al., 2013; Bos et al., 2017; Martínez Lorenzo et al., 2022). Such meaning representations are applied in many downstream applications—ranging from database querying to embodied questions answering—where parsers must handle a vast array of concepts that may not appear in the training data. While neural encoder-decoder architectures have shown impressive performance in semantic parsing tasks, their reliance on training distributions constrains their ability to generalize, especially to out-of-distribution (OOD) concepts.

Most existing semantic parsers struggle to interpret unseen words or rare senses, often defaulting to the most frequent meaning encountered during training. As a result, they fail to adapt to novel linguistic phenomena and remain limited to fixed patterns. Recent work (Zhang et al., 2024a) have attempted to mitigate these limitations by encoding concept representations numerically, forcing models to learn underlying ontologies from resources like WordNet (Fellbaum, 1998). However, these approaches require substantial preprocessing and intricate encodings that can be difficult for models to fully exploit.

In our work, instead, we explore the potential of large language models (LLMs), powerful decoder-only architectures with strong in-context learning capabilities and extensive pretraining, to enhance ability of semantic parsers to generalize. We pose two central research questions:

- **Do large language models outperform traditional encoder-decoder architectures in semantic parsing?** Decoder-only architectures are known to be more scalable and to internalize broader knowledge, potentially leading to stronger generalization and learning abilities. Assessing their performance in semantic parsing tasks can help reveal the architectural advantages of these decoder-only models.

- **How can these large language models be leveraged to improve generalization to out-of-distribution concepts?** Beyond simple architecture comparisons, we investigate whether LLMs can be guided to handle concepts more flexibly, using their ability to interpret and integrate external information.

In Section 2 we provide background on the semantic formalism of our choice, earlier approach to semantic parsing, and the challenge of an important sub-task, word sense disambiguation. Then we propose **Retrieval-Augmented Semantic Parsing** (RASP) in Section 3, a technique that integrates a retrieval mechanism into the parsing process. RASP leverages external lexical knowledge (e.g., WordNet) in the input, enabling the model

to dynamically access and interpret relevant concept information. By incorporating this retrieval step (Section 4), we relax the reliance on lemma-based mappings and allow the model to adapt more naturally to unseen words or senses. Our results show that this approach nearly doubles the performance on predicting OOD concepts compared to previous methods, demonstrating a substantial advancement in handling challenging open-domain data (Section 5).

2 Background and Related Work

2.1 Discourse Representation Structure

Discourse Representation Theory (Kamp and Reyle, 1993, DRT) is a semantic modeling framework. The core component of DRT is the Discourse Representation Structure (DRS), a formal representation that captures the meaning of a discourse, which captures the essence of the text and covering linguistic phenomena like anaphors and temporal expressions. Unlike many other formalisms such as Abstract Meaning Representation (Banarescu et al., 2013, AMR) used for large-scale semantic annotation efforts, DRS covers logical negation, quantification, and discourse relations, has complete word sense disambiguation, and offers a language-neutral meaning representation. A Discourse Representation Structure (DRS) can be coded and visualised in various ways, which are all provided in Parallel Meaning Bank (Abzianidze et al., 2017). In formal semantics they are often pictured in a human-readable box format. The clause notation was introduced to represent DRS in a sequential format suitable for machine learning models (van Noord et al., 2018). To further simplify DRS, Bos (2023) proposed a variable-free format known as Sequence Box Notation (SBN). An example of the three different but logically equivalent formats is shown in Figure 1. Recent trends in using seq2seq models have led to a preference for sequence notation, which is also the format used in this paper.

2.2 Semantic Parsing

Semantic parsing, as a traditional NLP task, remains essential in real-world applications, despite recent progress in natural language understanding shown by large language models. For instance, natural language front-end interfaces to databases require a mapping from text to structured data. Speech interactions with conversational agents that act in the real world (e.g., service robots) require

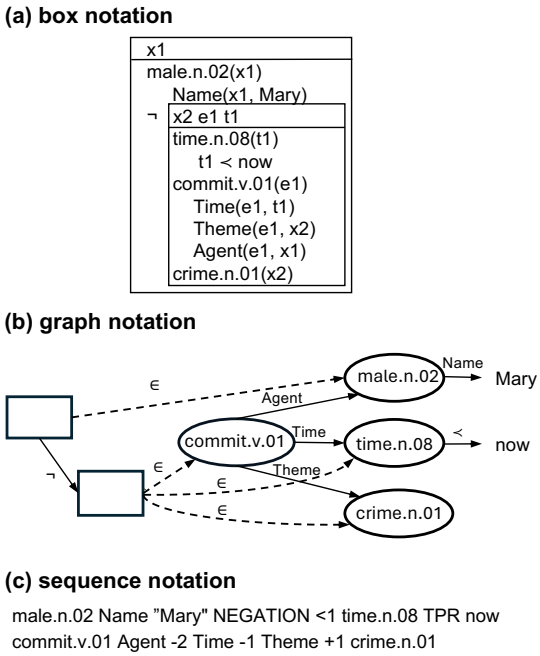

Figure 1: Three formats of Discourse Representation Structure (DRS) for "Mary didn't commit a crime.": the box notation, a directed acyclic graph, and the sequence notation.

situation-sensitive symbol grounding. Hence, advancing the development of more robust and general semantic parsers remains crucial.

Early approaches to semantic parsing primarily relied on rule-based systems (Woods, 1973; Hendrix et al., 1977; Templeton and Burger, 1983). The advent of neural methodologies, coupled with the availability of large semantically annotated datasets (Banarescu et al., 2013; Bos et al., 2017; Abzianidze et al., 2017), marked a significant shift in semantic parsing techniques (Barzdins and Gosko, 2016; van Noord and Bos, 2017; Bevilacqua et al., 2021a). The introduction of pre-trained language models within the sequence-to-sequence framework further improved parsing performance (van Noord et al., 2018, 2020; Ozaki et al., 2020; Samuel and Straka, 2020; Shou and Lin, 2021; Bevilacqua et al., 2021a; Zhou et al., 2021; Martínez Lorenzo et al., 2022; Zhang et al., 2024b; Liu, 2024a,b; Yang et al., 2024). Furthermore, several studies introduced more pre-training tasks specifically designed for semantic parsing (Bai et al., 2022; Wang et al., 2023a). With the rise of large language models, there has been considerable discussion about leveraging these models for semantic parsing, achieving notable results through techniques like prompting and chain-of-thought reasoning (Roy

et al., 2022; Ettinger et al., 2023; Jin et al., 2024). However, there is currently no work that leverages the knowledge and understanding capabilities of large language models to address the generalization problem in semantic parsing.

2.3 Word Sense Disambiguation

The generalization problem introduced in the previous section can also be understood as word sense disambiguation (WSD) for out-of-distribution concepts, within the context of semantic parsing. For instance, consider the sentence "She had £10,000 in the bank", with the target word "bank". In traditional WSD tasks, a predefined inventory of possible senses (e.g., 1. sloping land; 2. financial institution; 3. a long ridge or pile; 4. ...) would be provided, and the WSD model would classify the word according to one of these senses (Navigli, 2009; Bevilacqua et al., 2021b).

In semantic parsing, WSD can be seen as a subtask (Zhang et al., 2024a), but it is more challenging because the parsing model must generate the correct sense directly without access to an explicitly provided sense inventory. However, traditional knowledge-based WSD offers a potential solution that inspires our approach: by retrieving and presenting all possible concepts as alternatives, we can explicitly provide external information to the model, thereby enhancing its generalization capability. As a consequence, this requires the model to be able to process long contexts, making the LLMs become the preferences, in particular retrieval augmented generation.

2.4 Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is a hybrid approach that combines retrieval mechanisms with generative models to enhance the quality and accuracy of text generation tasks (Zhao et al., 2024; Gao et al., 2024). In RAG, a retrieval component first identifies relevant information from a large external knowledge base or corpus, which is then used as additional context for the generative model. This method allows the model to generate more informed and contextually accurate outputs, particularly in scenarios where the input data alone may not provide sufficient information.

By integrating retrieved knowledge into the generative process, RAG effectively bridges the gap between retrieval and generative models, leading to improved performance in tasks such as question answering (Karpukhin et al., 2020; Lewis et al., 2020; Borgeaud et al., 2021; Guu et al., 2020; Izacard and Grave, 2021; Petroni et al., 2020), commonsense reasoning (Liu et al., 2021; Wan et al., 2024) and other downstream tasks (Lewis et al., 2020; Izacard et al., 2024; Jiang et al., 2023; Guo et al., 2023; Cheng et al., 2023; Li et al., 2023). While RAG was initially employed in a wide scope of applications, its popularity can be attributed to the advent of large language models and their strong capabilities. Consequently, we will concentrate on the application of RAG in the context of LLMs.

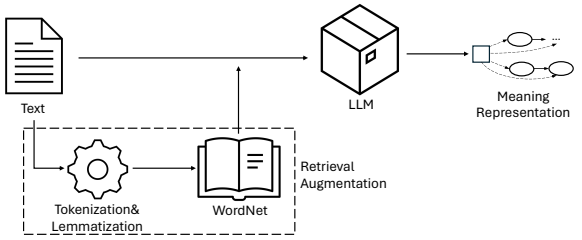

Figure 2: Global overview of RASP (Retrieval-Augmented Semantic Parsing). Both the training and testing phases adhere to this pipeline.

3 Retrieval-Augmented Semantic Parsing

We propose a new method that combines retrieval-augmented generation with semantic interpretation: Retrieval-Augmented Semantic Parsing (RASP), a framework that is outlined in Figure 2. It comprises two key components: retrieval and parsing.

Different from the Dense Passage Retrieval (Karpukhin et al., 2020) method, which is commonly employed in question-answering tasks, our retrieval process is designed to be more straightforward and tailored to the needs of semantic parsing. The process begins with tokenizing and lemmatizing[1] the source text. Following these, we perform a search for relevant concept synsets in an external knowledge base, specifically WordNet. For example, in the sentence "Mary went for birdwatching. She saw a harrier, a golden eagle, and a hobby", the retrieval process would identify multiple synsets for "go", "birdwatch", "see", "harrier", "golden eagle" and "hobby", as illustrated in Table 1. Additionally, to ensure comprehensive coverage of multi-word expressions, which are critical in capturing the correct semantic meaning, we employ a hierarchical n-gram search strategy. This strategy involves sequential searches using 4-gram, 3-gram, 2-gram, and 1-gram patterns, thereby ensuring that

[1] https://www.nltk.org/api/nltk.stem.wordnet.html

Source Text	Mary went for birdwatching. She saw a harrier, a golden eagle, and a hobby.	
Concepts	golden_eagle.n.01:	large eagle of mountainous regions of ... having a golden-brown head and neck
	birdwatch.v.01:	watch and study birds in their natural habitat
	...	...
	harrier.n.01:	a persistent attacker
	harrier.n.02:	a hound that resembles a foxhound but is smaller
	harrier.n.03:	hawks that hunt over meadows and marshes and prey on small terrestrial animals ...
	...	...
	hobby.n.01	an auxiliary activity
	hobby.n.02	a child's plaything consisting of an imitation horse mounted on rockers ...
	hobby.n.03	small Old World falcon formerly trained and flown at small birds
Prompts	Normal prompt:	Text to parse: {*Source Text*}
	RASP prompt:	Considering the concepts with glosses: {*Concepts*}. Text to parse: {*Source Text*}
Gold DRS	female.n.02 Name "Mary" time.n.08 TPR now birdwatch.v.01 Agent -2 Time -1 ELABORATION <1 female.n.02 ANA -3 see.v.01 Experiencer -1 Time +1 Stimulus +3 time.n.08 TPR now harrier.n.03 golden_eagle.n.01 entity.n.01 Sub -2 Sub -1 Sub +1 hobby.n.03	

Table 1: An example illustrating the workflow of RASP. We omit some senses and words for the retrieved concepts to save space. The distinction between prompts for semantic parsing with and without RASP are shown in the *Prompts* row. Some examples of complete prompts can be found in Appendix A.

no multi-word expressions (such as "golden eagle") are overlooked.

The parsing process for a decoder-only model[2] is guided by the probability distribution of possible output sequences given an input sequence. The model generates an output sequence by predicting each token iteratively, based on the input text and previously generated tokens, as shown in (1).

$$p_{decoder-only}(o|x) = \prod_{i=1}^{n} p_\theta(o_i|x, o_{1:i-1}) \quad (1)$$

Here, x is the input text, $o_{1:i-1}$ represents the sequence generated so far, and o_i is the token generated at the current step. θ refers to the model's parameters, and p denotes the likelihood of generating output sequence o given input sequence x.

To enhance this process, retrieval and generation are integrated, leveraging external knowledge to inform output generation. Mathematically, the retrieval step introduces a probability, $p(o'|x)$, which models the likelihood of retrieving relevant concepts o' based on x. This probability is combined multiplicatively with the generation probability, as shown in (2). This combination ensures both components contribute meaningfully, with retrieval acting as a filter to guide the generation process toward relevant concepts.

$$\begin{aligned} P_{RASP}(o|x) &= p(o'|x) p_{decoder-only}(o|x, o') \\ &= p(o'|x) \prod_{i=1}^{N} p_\theta(o_i|x, o', o_{1:i-1}) \end{aligned} \quad (2)$$

By incorporating retrieved concepts, RASP goes beyond relying solely on the input sequence and training data, adding additional context to guide generation. For example, when handling words with multiple meanings, like "hobby," retrieved synsets help the model select the correct interpretation based on glosses and context. This integration sharpens the model's focus on relevant concepts, reducing the likelihood of generating incorrect or overly broad outputs, particularly for out-of-distribution concepts.

4 Experiments

4.1 Datasets

We conduct our experiments on the Parallel Meaning Bank (PMB, version 5.1.0)[3] (Abzianidze et al., 2017; Zhang et al., 2024b). We first use the gold-standard English data of the PMB to evaluate the large language models and their retrieval-augmented version under in-distribution conditions.

To further assess the models' ability to handle out-of-distribution (OOD) concepts, we adopt the

[2]The models we use are all in decoder-only architecture, so we omit the discussion about encoder-decoder architecture.

[3]https://pmb.let.rug.nl/releases

challenge set proposed by Zhang et al. (2024a), which is also derived from the PMB. Neural semantic parsers often default to the first sense of unknown concepts–an approach that can lead to "lucky guesses" without truly understanding new words. The challenge set, consisting of 500 sentences, is deliberately designed to eliminate this shortcut. Each sentence includes at least one concept that does not appear in the training data and does not correspond to the first sense in the ontology. In total, the challenge set contains 410 unknown nouns, 128 verbs, and 65 modifiers (adjectives and adverbs). By evaluating on this set, we measure the true generalization capability of the models, testing whether they can correctly interpret novel concepts rather than relying on heuristic assigning.

Train	Dev	Standard	Challenge
9,560	1,195	1,195	500

Table 2: Dataset statistics for PMB 5.1.0, i.e., number of meaning representations for train, development and two test sets: standard and challenge.

4.2 Experiment Settings

It is crucial to note that large language models, when used in zero-shot or few-shot scenarios, tend to perform poorly on the highly complex graph structures inherent in formal meaning representations such as DRS. Prior work (Ettinger et al., 2023; Zhang et al., 2024a) demonstrates that without fine-tuning, LLMs struggle to match the performance of models specifically optimized for these tasks. Therefore, in our experiments, we fine-tune all large language models.

For RASP, we explore two retrieval-enhanced approaches: (1) Train+Test Retrieval: Incorporate retrieval-derived concepts both during training and inference, thereby familiarizing the model with external lexical knowledge throughout the entire learning process. (2) Test-Only Retrieval: Use retrieval only during inference, training the model on raw DRS structures without external lexical inputs. Our experiments show that the first approach consistently yields better performance. Thus, we focus our primary analysis on the first approach and provide results for the second approach in Appendix C.

Due to computational constraints, we select open-sourced LLMs with model sizes under 10B parameters, including phi3-4B, Mistral-7B, LLaMa3-3B, LLaMa3-8B, Gemma2-2B, Gemma2-9B, Qwen2.5-3B, and Qwen2.5-7B. These models strike a balance between state-of-the-art language understanding and manageable resource requirements. For fine-tuning, we employ Low-Rank Adaptation (Hu et al., 2021, LoRA), a parameter-efficient technique that introduces trainable low-rank matrices (A and B) into the model's layers, greatly reducing computational overhead:

$$W + \Delta W = W + A \cdot B \tag{3}$$

We compare our results against several strong baselines, including BART, T5, byT5, TAX-parser (Zhang et al., 2024b), and AMS-Parser (Yang et al., 2024), all of which were previously fine-tuned on PMB data. We exclude work conducted on earlier versions of PMB or using silver data. Additionally, we do not apply retrieval augmentation to these baseline models due to input length constraints, which limit their ability to incorporate external lexical sources efficiently.

We trained each model for 10 epochs, using a learning rate of 10^{-4}, and fp16 precision. More information on the hyperparameters is provided in Appendix B.

4.3 Evaluation Metrics

We used SMATCH and its variants to evaluate the performance of the models. SMATCH (Cai and Knight, 2013), referred to as Hard-SMatch, strictly matches concepts, where any discrepancy results in a non-match. In contrast, its variant, Soft-SMatch (Opitz et al., 2020), considers concept similarity when matching. Instead of adopting the approach of using word-embedding similarity, we applied the Wu-Palmer similarity (Wu and Palmer, 1994), as introduced by Zhang et al. (2024b). Wu-Palmer similarity provides a precise measure of semantic similarity between concepts based on their positions within the WordNet taxonomy. Unlike embedding-based methods, it does not rely on external training and easily adapts to changes in WordNet's structure or content. The calculation is:

$$\text{WuP} = 2 * \frac{depth(\text{LCS}(s_1, s_2))}{depth(s_1) + depth(s_2)} \tag{4}$$

where s is the concept, LCS refers to the Least Common Subsumer of these concepts, and depth denotes the distance from the concept to the root of the taxonomy.

Model	Size	Input	Graph-level			Node-level
			Hard-SMatch↑	Soft-SMatch↑	IFR↓	F score↑
BART-large	400M	Normal	79.54	82.81	3.92	75.40
T5-large	770M	Normal	84.27	86.44	6.41	79.88
byT5-large	580M	Normal	87.41	89.43	4.78	84.75
AMS-Parser	–	Normal	87.08	89.15	0.00	85.00
TAX-Parser	580M	Normal	86.65	91.80	2.34	80.12
Phi-3	4B	Normal	85.74	87.92	4.94 (59)	81.60
		RASP	85.96 (+0.3%)	88.13 (+0.2%)	4.80 (57)	83.33 (+2.1%)
Mistral	7B	Normal	89.95	92.48	2.00 (24)	83.90
		RASP	90.95 (+1.1%)	93.33 (+0.9%)	1.58 (19)	85.00 (+1.3%)
Qwen2.5	3B	Normal	86.50	88.64	4.69 (56)	82.60
		RASP	88.70 (+2.5%)	90.74 (+2.4%)	3.01 (36)	83.90 (+1.6%)
	7B	Normal	89.88	91.83	2.51 (30)	84.50
		RASP	89.93 (+0.1%)	91.87 (+0.1%)	2.51 (30)	85.50 (+1.2%)
LLama3	3B	Normal	87.30	90.01	3.34 (40)	81.50
		RASP	87.76 (+0.5%)	90.51 (+0.6%)	3.01 (36)	82.30 (+1.0%)
	8B	Normal	89.92	92.46	2.09 (25)	83.90
		RASP	90.65 (+0.8%)	93.10 (+0.7%)	**1.50** (18)	84.72 (+1.0%)
Gemma2	2B	Normal	89.20	91.08	3.01 (36)	84.20
		RASP	89.30 (+0.1%)	91.23 (+0.2%)	3.10 (37)	85.58 (+1.6%)
	9B	Normal	90.72	93.15	1.67 (20)	84.67
		RASP	**91.37** (+0.7%)	**93.65** (+0.5%)	1.58 (19)	**86.11** (+1.7%)

Table 3: Performance of baseline models, large language models (Normal) and their retrieval-augmented variants (RASP) on standard test, with percentage changes in parentheses. Size is the number of model's parameters (B: billion). IFR is Ill-Formed Rate and the number of ill-formed prediction are in parentheses. Note: AMS-Parser (Yang et al., 2024) performs well for IFR for it is a compositional neuro-symbolic system. TAX-Parser (Zhang et al., 2024a) is a neuro-symbolic system, trained with a novel encoded meaning representation.

For the fine-grained evaluation on the challenge set, we applied the metric proposed by Wang et al. (2023b), focusing specifically on concept-node matching scores. When evaluating the results on the challenge set, we directly calculated the Wu-Palmer similarity between the target concepts and the corresponding model-generated results.

5 Results

5.1 Semantic Parsing on Standard Test

Table 3 shows that large language models consistently surpass earlier encoder-decoder baselines, providing direct evidence for our first research question. While BART, T5, and byT5 achieve Hard-SMatch scores up to 87.41, several LLM-based models (e.g., Mistral-7B, Gemma2-9B) exceed 90.0 on the standard test set. This improvement is substantial, with the strongest baseline LLM reaches 90.72 on Hard-SMatch, outperforming the best encoder-decoder model (byT5) by a margin of 3.3 points.

These higher scores are also reflected in Soft-SMatch and node-level F-scores, indicating that LLM-based models not only produce more structurally accurate meaning representations but also more reliably identify concept nodes. Additionally, Ill-Formed Rate (IFR) reductions suggest that these models generate fewer ill-structured outputs. In summary, these improvements highlight that large language models outperforms previous encoder-decoder models.

Beyond confirming the advantages of LLMs, we also examine the impact of retrieval augmentation (RASP) on standard test results. Although the largest gains from retrieval are observed on the challenge set (as discussed in Section 5.2), even here on the in-distribution standard test, RASP provides consistent performance improvements. Most LLMs show an increase of about 0.3% to 2.5% in Hard-SMatch and Soft-SMatch scores when using RASP. Furthermore, the Ill-Formed Rate (IFR) tends to decrease, and the node-level F-score improves by approximately 1.0% to 2.1%. These node-level gains suggest that RASP's improvements stem largely from more accurate concept prediction. While these enhancements are moderate in the standard test scenario, they indicate that retrieval can enhance the model's understanding of concept-level semantics.

Model	Input	Noun	Verb	Modifiers	Overall
BART-large	Normal	26.11	37.34	46.88	30.95
T5-large	Normal	25.48	35.21	41.28	29.45
byT5-large	Normal	27.59	39.14	44.70	32.13
TAX-Parser	Normal	42.15	31.58	43.27	39.68
phi-4B	Normal	35.48	36.91	46.97	37.91
	RASP	62.03 (+74.8%)	46.32 (+25.5%)	63.63 (+35.5%)	58.28 (+53.7%)
Mistral-7B	Normal	38.02	40.61	50.00	39.87
	RASP	72.03 (+89.5%)	59.27 (+46.0%)	67.42 (+34.8%)	68.44 (+71.7%)
Qwen2.5-7B	Normal	38.51	37.52	46.97	39.12
	RASP	66.77 (+73.4%)	56.95 (+51.8%)	64.39 (+37.1%)	64.12 (+63.8%)
LLama3.2-8B	Normal	37.06	34.79	47.73	37.59
	RASP	72.28 (+95.1%)	61.62 (+77.1%)	66.67 (+39.7%)	69.86 (+85.9%)
Gemma2-9B	Normal	39.68	45.01	55.30	42.54
	RASP	**73.93** (+86.3%)	**62.31** (+36.5%)	**69.70** (+26.0%)	**70.41** (+65.6%)

Table 4: Wu-Palmer similarities between unknown concepts and generated concepts across four parts of speech. For the sake of clarity, we exclude the smaller version of the same model.

5.2 Performance on the Challenge Set

Table 4 provides the results on the challenge set, designed specifically to test the models' ability to predict out-of-distribution (OOD) concepts. Here, we report Wu-Palmer similarities for unknown nouns, verbs, and modifiers (adjectives & adverbs). We calculate the Wu-Palmer similarities between the target concepts (out-of-distribution concepts) and the generated concepts (see examples in Table 5).

Among the baselines, TAX-Parser (Zhang et al., 2024a) stands out, achieving an overall similarity score of 39.68. However, some Normal (non-RASP) large language models already exceed this performance on the challenge set. For example, Gemma2-9B (Normal) obtains an overall score of 42.54, indicating that LLMs can yield improvements, even without retrieval augmentation. When retrieval augmentation (RASP) is introduced, these large language models show substantial additional gains. For example, Gemma2-9B (RASP) achieves an overall similarity score of 70.41, compared to the best baseline's 39.68—an increase of over 30 absolute points. These gains are particularly remarkable for noun concepts, with relative improvements of approximately 70% to 95%. Verbs show increases between about 25% and 77%, and modifiers improve by roughly 26% to 43%.

These results directly support our second research question regarding improving out-of-distribution generalization. While model scaling alone can yield moderate improvements, the integration of external lexical knowledge through retrieval allows LLMs to select more accurate concepts in OOD scenarios. In effect, RASP helps the models "look up" relevant information, enhancing their concept selection and producing more semantically appropriate results. In this case, retrieval-augmented LLMs not only outperform strong baselines like TAX-Parser but also set the state-of-the-art for OOD semantic parsing performance.

5.3 Error Analysis on the Challenge Set

We selected a subset of the challenge set and manually checked how the best performing model—Gemma2-9B (Normal) and Gemma2-9B (RASP)—handle the out-of-distribution concepts.

We picked 22 instances, comprising 11 completely perfect predictions (WuP=1.00) and 11 imperfect predictions (WuP<1.00) made by RASP, as presented in Table 5. With respect to the perfect predictions, it is evident that the retrieval significantly enhance the model's ability to interpret most out-of-distribution concepts. For instance, in the text about birdwatching, the word "hobby" clearly refers to a species of bird. The model without RAG defaults to the most frequent sense number, predicts hobby.n.01 (an auxiliary activity). In contrast, retrieval provides the glosses of each sense related to the noun "hobby" and leads the model to pick hobby.n.03 (a falcon), by explicit lexical connections between "falcon" in the gloss of hobby.n.03 and the context provided by "birdwatching".

However, RASP makes imperfect predictions sometimes. We identified three possible causes: (a) similar glosses between WordNet concepts; (b) insufficient textual context; and (c) limitations in the model's linguistic coverage.

Input Text	Gold	Normal	RASP
He bought the painting for a **song** on a flea market.	song.n.05	n.03 (0.22)	n.05 (1.00)
The detective **planted** a bug in the suspect's office to to gather evidence.	plant.v.05	v.02 (0.22)	v.05 (1.00)
Scientist examines the insect's **antennae**.	antenna.n.03	n.01 (0.24)	n.03 (1.00)
I've seen a short **extract** from the film.	extract.n.02	n.01 (0.25)	n.02 (1.00)
She prepared a three **course** meal.	course.n.07	n.03 (0.27)	n.07 (1.00)
The music student practiced the **fugue**.	fugue.n.03	n.02 (0.28)	n.03 (1.00)
Johanna went birdwatching. She saw a harrier, a kite, and a **hobby**.	hobby.n.03	n.02 (0.38)	n.03 (1.00)
A harrier is a **muscular** dog with a hard coat.	muscular.a.02	a.01 (0.50)	a.02 (1.00)
The hiker spotted an **adder** sunbathing on a rock.	adder.n.03	n.01 (0.50)	n.03 (1.00)
A tiny **wren** was hiding in the shrubs.	wren.n.02	n.01 (0.55)	n.02 (1.00)
Hungarian is a challenging language with 18 cases.	hungarian.n.02	n.01 (0.11)	n.02 (1.00)
The moon is **waxing**.	wax.v.03	v.03 (1.00)	v.02 (0.75)
The function **ordered** the strings alphabetically.	order.v.05	v.02 (0.17)	v.06 (0.75)
The elephant's trunk is an **extended** nose.	extended.a.03	a.01 (0.50)	a.01 (0.50)
A **tripper** helps control the flow of materials on a conveyor.	tripper.n.04	n.02 (0.40)	n.02 (0.40)
We saw a **kite** gliding in the sky during the walking.	kite.n.04	n.03 (0.40)	n.03 (0.40)
The elegant **pen** glided gracefully across the tranquil lake.	pen.n.05	n.01 (0.36)	n.01 (0.36)
The immature sparrows are **feathering** already.	feather.v.05	v.03 (0.20)	v.02 (0.29)
The visitors can **observe** various species of ray in the quarium.	observe.v.02	v.01 (0.25)	v.01 (0.25)
She **hobbled** the horse. It freaked out.	hobble.v.03	v.01 (0.18)	v.02 (0.18)
The gardener noticed the **growth** on the rose after the rain.	growth.n.04	n.01 (0.18)	n.01 (0.18)
The surge **alarmed** the town's residents.	alarm.v.02	v.01 (0.15)	v.01 (0.15)

Table 5: Twenty instances of the challenge set with content words with out-of-distribution concepts in bold face, and the concepts generated by the Gemma2-9B (Normal) and retrieval-augmented Gemma2-9B (RASP). The scores in brackets are the Wu-Palmer Similarity between the predicted concept and gold concept.

The verb "alarm" in Table 5 is an instance of the similarity problem. The challenge arises because some of its senses have similar glosses, such as alarm.v.01 (fill with apprehension or alarm) and alarm.v.02 (warn or arouse to a sense of danger). Similar issues occur with the verbs "wax", "order", "observe" and "hobble". Although glosses were carefully crafted by lexicographers, they don't always show a clear difference in meaning (Mihalcea and Moldovan, 2001; Navigli, 2006).

In cases of insufficient textual context, such as with the noun "kite" in the sentence "We saw a kite gliding in the sky", the sense annotators chose kite.n.04 (a bird of prey). However, kite.n.03 (a plaything) could perhaps also be appropriate given the limited context provided by this sentence. Similar issues can be raised in the sentences with the noun "tripper" and the verb "feathering".

The third cause can be attributed to the model's linguistic coverage. A case in point is "pen": the meanings of pen.n.01 (a writing implement) and pen.n.05 (a female swan) are quite different, but the latter is the correct one in the text "Jane saw two swans. The elegant pen glided gracefully across the tranquil lake". However, the model fails to distinguish them, likely because "pen" is rarely used to refer to "swan" in available corpora. As a result, the models may not have encountered this sense during training, making it challenging for them to predict a meaning they have not been exposed to. In sum, while retrieval drastically improves concept prediction, there are still some difficulties that can pose challenges for the models.

6 Conclusion

This paper demonstrates that large language models (LLMs), even without retrieval augmentation, outperform previous encoder-decoder approaches in semantic parsing for Discourse Representation Structures, thereby answering our first research question in the affirmative, setting a new state of the art. We have also shown that our proposed Retrieval-Augmented Semantic Parsing (RASP) framework, which integrates external lexical knowledge, further enhances the performance of LLMs. Notably, RASP nearly doubles the accuracy on out-of-distribution concepts, providing a direct answer to our second research question and confirming its robust generalization ability in open-domain scenarios. Our experiments show that by simply appending relevant lexical information to the model input, the RASP approach offers a practical and intuitive approach that can be easily applied to other meaning representations used in natural language processing, such as AMR (Banarescu et al., 2013) and BMR (Martínez Lorenzo et al., 2022).

7 Limitations

We think the limitations of this work mainly come from two aspects: the language models used in RASP and the retrieval source (i.e., WordNet).

The retrieval is proven to provide more information and knowledge to the models. However, retrieval will significantly increase the input length of the model, making it (only) adoptable for the large language models with strong context understanding and long text processing capabilities. Therefore, the RASP framework cannot be directly used to improve previous parsers that rely on other methods, which is also why we only provided results of retrieval-augmented LLMs.

Another limitation is the retrieval source. RASP is based on WordNet, so if a sense is not in WordNet, it will never be guessed. For example, "velvet scooter" (a bird) is not in WordNet, nor is Cobb salad (a dish). Hence, RASP will never make a perfect prediction for such cases. Moreover, the glosses in WordNet, even though carefully crafted by lexicographers in most cases, are sometimes concise, lacking information to separate them from other senses. This makes it difficult for the models to accurately distinguish between different meanings (see Section 5.3).

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW@ACL*.

Guntis Barzdins and Didzis Gosko. 2016. RIGA at SemEval-2016 task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1143–1147, San Diego, California. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021a. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In *AAAI Conference on Artificial Intelligence*.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021b. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. 2021. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*.

Johan Bos. 2023. The sequence notation: Catching complex meanings in simple graphs. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS 2023)*, pages 1–14, Nancy, France.

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337, Singapore. Association for Computational Linguistics.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "you are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. 2023. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10896–10912, Toronto, Canada. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association (ELRA).

Gary G. Hendrix, Earl D. Sacerdoti, Daniel Sagalowicz, and Jonathan Slocum. 1977. Developing a natural language interface to complex data. In *TODS*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Fernando Gonzalez Adauto, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, and Mona Diab. 2024. Analyzing the role of semantic representations in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3781–3798, Mexico City, Mexico. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. From discourse to logic - introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. In *Studies in Linguistics and Philosophy*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. 2023. Structure-aware language model pretraining improves dense retrieval on structured data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11560–11574, Toronto, Canada. Association for Computational Linguistics.

Jiangming Liu. 2024a. Model-agnostic cross-lingual training for discourse representation structure parsing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11486–11497, Torino, Italia. ELRA and ICCL.

Jiangming Liu. 2024b. Soft well-formed semantic parsing with score-based selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15037–15043, Torino, Italia. ELRA and ICCL.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.

Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741, Dublin, Ireland. Association for Computational Linguistics.

Rada Mihalcea and Dan I. Moldovan. 2001. Automatic generation of a coarse grained wordnet. In *NAACL Workshop on WordNet and Other Lexical Resources*.

Roberto Navigli. 2006. Reducing the granularity of a computational lexicon via an automatic mapping to a coarse-grained sense inventory. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 841–844, Genoa, Italy. European Language Resources Association (ELRA).

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

Stephan Oepen and Jan Tore Lønning. 2006. Discriminant-based MRS banking. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Juri Opitz, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8:522–538.

Hiroaki Ozaki, Gaku Morio, Yuta Koreeda, Terufumi Morishita, and Toshinori Miyoshi. 2020. Hitachi at MRP 2020: Text-to-graph-notation transducer. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 40–52, Online. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *AKBC 2020*.

Subhro Roy, Sam Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, Benjamin Van Durme, Microsoft Semantic Machines, and Scaled Cognition. 2022. Benchclamp: A benchmark for evaluating language models on syntactic and semantic parsing. In *Neural Information Processing Systems*.

David Samuel and Milan Straka. 2020. ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online. Association for Computational Linguistics.

Ziyi Shou and Fangzhen Lin. 2021. Incorporating EDS graph for AMR parsing. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 202–211, Online. Association for Computational Linguistics.

Marjorie Templeton and John Burger. 1983. Problems in natural-language interface to DBMS with examples from EUFID. In *First Conference on Applied Natural Language Processing*, pages 3–16, Santa Monica, California, USA. Association for Computational Linguistics.

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

Rik van Noord and Johan Bos. 2017. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *Computational Linguistics in the Netherlands Journal*, 7:93–108.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Chunliu Wang, Huiyuan Lai, Malvina Nissim, and Johan Bos. 2023a. Pre-trained language-meaning models for multilingual parsing and generation. In *Findings of the Association for Computational Linguistics*, page 5586–5600. Association for Computational Linguistics (ACL).

Chunliu Wang, Xiao Zhang, and Johan Bos. 2023b. Discourse representation structure parsing for Chinese. In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 62–74, Nancy, France. Association for Computational Linguistics.

William A Woods. 1973. Progress in natural language understanding: an application to lunar geology. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, pages 441–450.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Xiulin Yang, Jonas Groschwitz, Alexander Koller, and Johan Bos. 2024. Scope-enhanced compositional semantic parsing for DRT. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19602–19616, Miami, Florida, USA. Association for Computational Linguistics.

Xiao Zhang, Gosse Bouma, and Johan Bos. 2024a. Neural Semantic Parsing with Extremely Rich Symbolic Meaning Representations. *Computational Linguistics*, pages 1–40.

Xiao Zhang, Chunliu Wang, Rik van Noord, and Johan Bos. 2024b. Gaining more insight into neural semantic parsing with challenging benchmarks. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 162–175, Torino, Italia. ELRA and ICCL.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *Preprint*, arXiv:2402.19473.

Jiawei Zhou, Tahira Naseem, Ramón Fernández Astudillo, and Radu Florian. 2021. Amr parsing with action-pointer transformer. In *North American Chapter of the Association for Computational Linguistics*.

A Prompt

The following is a complete example of the prompts we use for the LLMs. Since the models we use are all instruction-based versions, the prompt is structured in a dialogue format.

RASP Prompt

```
user:
Please parse the following text into Discourse Representation Structure, considering using the concepts based on the following glosses:
- birdwatch.v.01: watch and study birds in their natural habitat
- saw.n.01: a condensed but memorable saying embodying some important fact of experience that is taken as true by many people
- saw.n.02: hand tool having a toothed blade for cutting
- saw.n.03: a power tool for cutting wood
- saw.v.01: cut with a saw
- harrier.n.01: a persistent attacker
- harrier.n.02: a hound that resembles a foxhound but is smaller; used to hunt rabbits
- harrier.n.03: hawks that hunt over meadows and marshes and prey on small terrestrial animals
- kite.n.01: a bank check that has been fraudulently altered to increase its face value
- kite.n.02: a bank check drawn on insufficient funds at another bank in order to take advantage of the float
- kite.n.03: plaything consisting of a light frame covered with tissue paper; flown in wind at end of a string
- kite.n.04: any of several small graceful hawks of the family Accipitridae having long pointed wings and feeding on insects and small animals
- kite.v.01: increase the amount (of a check) fraudulently
- kite.v.02: get credit or money by using a bad check
- kite.v.03: soar or fly like a kite
- kite.v.04: fly a kite
- hobby.n.01: an auxiliary activity
- hobby.n.02: a child's plaything consisting of an imitation horse mounted on rockers; the child straddles it and pretends to ride
- hobby.n.03: small Old World falcon formerly trained and flown at small birds

Text to parse: Johanna went birdwatching. She saw a harrier, a kite, and a hobby.

model:
female.n.02 Name "Johanna" time.n.08 TPR now birdwatch.v.01 Agent -2 Time -1 ELABORATION <1 female.n.02 ANA -3 see.v.01 Experiencer -1 Time +1 Stimulus +3 time.n.08 TPR now harrier.n.03 kite.n.04 entity.n.01 Sub -2 Sub -1 Sub +1 hobby.n.03
```

Normal Prompt

```
user:
Text to parse: Johanna went birdwatching. She saw a harrier, a kite, and a hobby.

model:
female.n.02 Name "Johanna" time.n.08 TPR now birdwatch.v.01 Agent -2 Time -1 ELABORATION <1 female.n.02 ANA -3 see.v.01 Experiencer -1 Time +1 Stimulus +3 time.n.08 TPR now harrier.n.03 kite.n.04 entity.n.01 Sub -2 Sub -1 Sub +1 hobby.n.03
```

B Experiment Settings

Table 6 and 7 provide the basic details of the experiments and models.

Category	Details	Category	Details
Stage	SFT/inference	Precision	fp16
Fine-tuning	LoRA	Batch Size	1
Cutoff Length	1024	GPU Number	4
Learning Rate	10^{-4}	GPU	H100
Epochs	10	lr scheduler	cosine

Table 6: Configurations for large language models Fine-Tuning and Inference.

Model	Details
BART-large	facebook/bart-large
T5-large	google-t5/t5-large
byT5-large	google/byt5-large
Phi-3-4B	microsoft/Phi-3.5-mini-instruct
Qwen2.5-3B	Qwen/Qwen2.5-3B-Instruct
Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct
LLama3.2-3B	meta-llama/Llama-3.2-3B-Instruct
LLama3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Gemma2-2B	google/gemma-2-2b-it
Gemma2-9B	google/gemma-2-9b-it

Table 7: Details of Models.

C Additional Experiments

We present the results of fine-tuning on Normal data and testing by RASP prompt, as shown in Tables 8 and 9. This approach involves providing retrieval information during inference but using only text-to-DRS data during training. From the results, it is evident that this training method adversely affects the model's performance, particularly on the standard test. We believe that fine-tuning reduces the models' ability of in-context learning, which limits the models from effectively utilizing the additional information provided by retrieval.

Model	Size	Input	Graph-level			Node-level
			Hard-SMatch↑	Soft-SMatch↑	IFR↓	F score↑
Phi-3	4B	Normal	85.74	87.92	4.94 (59)	81.60
		RASP	66.78 (–22.1%)	70.88 (–19.4%)	14.9 (178)	63.50 (–22.2%)
Mistral	7B	Normal	89.95	92.48	2.00 (24)	83.90
		RASP	83.22 (–7.5%)	85.90 (–7.1%)	3.58 (43)	80.10 (–4.5%)
Qwen2.5	3B	Normal	86.50	88.64	4.69 (56)	82.60
		RASP	84.32 (–2.5%)	87.44 (–1.4%)	5.00 (60)	81.90 (–0.8%)
	7B	Normal	89.88	91.83	2.51 (30)	84.50
		RASP	86.23 (–4.1%)	90.78 (–1.1%)	2.57 (33)	83.40 (–1.3%)
LLama3	3B	Normal	87.30	90.01	3.34 (40)	81.50
		RASP	85.90 (–1.6%)	86.91 (–3.4%)	4.10 (49)	77.59 (–4.8%)
	8B	Normal	89.92	92.46	2.09 (25)	83.90
		RASP	88.65 (–1.4%)	91.30 (–1.3%)	2.50 (30)	82.11 (–2.1%)
Gemma2	2B	Normal	89.20	91.08	3.01 (36)	84.20
		RASP	84.40 (–5.4%)	89.93 (–1.3%)	3.01 (36)	80.11 (–4.9%)
	9B	Normal	90.72	93.15	1.67 (20)	84.67
		RASP	91.11 (+0.4%)	93.35 (+0.2%)	1.79 (21)	83.10 (–1.9%)

Table 8: Performance on standard test.

Model	Input	Noun	Verb	Modifiers	Overall
phi-4b	Normal	35.48	36.91	46.97	37.91
	RASP	40.03 (+12.8%)	36.32 (–1.6%)	49.13 (+4.6%)	40.03 (+5.6%)
Mistral-7b	Normal	38.02	40.61	50.00	39.87
	RASP	40.90 (+7.6%)	49.27 (+21.3%)	50.00 (–0.0%)	43.61 (+9.4%)
Qwen2.5-7b	Normal	38.51	37.52	46.97	39.12
	RASP	40.11 (+4.2%)	43.54 (+16.1%)	50.00 (+6.5%)	41.94 (+7.2%)
LLama3.2-8b	Normal	37.06	34.79	47.73	37.59
	RASP	42.10 (+13.6%)	38.88 (+11.8%)	49.00 (+2.7%)	42.00 (+11.7%)
Gemma2-9b	Normal	39.68	45.01	55.30	42.54
	RASP	**45.93** (+15.8%)	**50.11** (+11.3%)	**59.70** (+8.0%)	**48.34** (+13.6%)

Table 9: Performance on the challenge set.