Rethinking Semantic Parsing for Large Language Models: Enhancing LLM Performance with Semantic Hints

Kaikai An[1,2*], Shuzheng Si[1,2*], Helan Hu[1,2], Haozhe Zhao[1,2], Yuchi Wang[1], Qingyan Guo[3], Baobao Chang[1†]

[1] National Key Laboratory for Multimedia Information Processing, Peking University
[2] School of Software and Microelectronics, Peking University [3] Tsinghua University
`ankaikai@stu.pku.edu.cn, chbb@pku.edu.cn`

Abstract

Semantic Parsing aims to capture the meaning of a sentence and convert it into a logical, structured form. Previous studies show that semantic parsing enhances the performance of smaller models (e.g., BERT) on downstream tasks. However, it remains unclear whether the improvements extend similarly to LLMs. In this paper, our empirical findings reveal that, unlike smaller models, directly adding semantic parsing results into LLMs reduces their performance. To overcome this, we propose SENSE, a novel prompting approach that embeds semantic hints within the prompt. Experiments show that SENSE consistently improves LLMs' performance across various tasks, highlighting the potential of integrating semantic information to improve LLM capabilities.

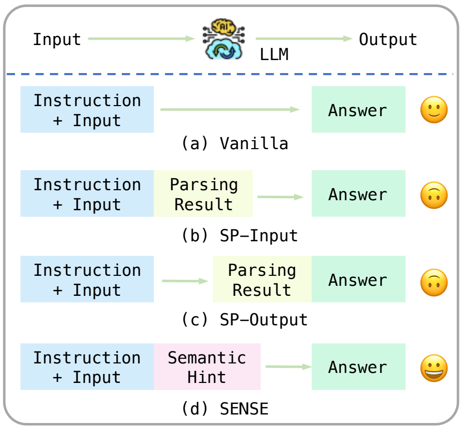

Figure 1: Different ways of evaluating LLMs on downstream tasks. While (a) represents direct prompting models, (b) and (c) add semantic parsing results either from the input or output side. The upside-down face indicates a negative impact. Our method, SENSE, introduces semantic hints without perception of the results.

1 Introduction

Semantic Parsing is a fundamental task in Natural Language Processing, which involves converting a natural language sentence into structured meaning representation. This includes tasks like Semantic Role Labeling (SRL), Frame Semantic Parsing (FSP) and Abstract Meaning Representation (AMR) (Gildea and Jurafsky, 2002; Baker et al., 2007; Banarescu et al., 2013; Palmer et al., 2010; An et al., 2023). Such structured information are applicable across various tasks, like Question Answering (Khashabi et al., 2022), Machine Translation (Rapp, 2022), Dialogue Systems (Xu et al., 2020; Si et al., 2022, 2024) and so on.

Previous work from Bonial et al. (2020); Rapp (2022); Khashabi et al. (2022) demonstrate that integrating semantic parsing results from SRL or AMR parsing into a model's input can effectively enhance its ability to understand illocutionary acts and linguistic abstractions, thereby improving downstream performance. However, these findings are largely limited to smaller models like BERT (Devlin et al., 2019). With the rise of Large Language Models (LLMs), it becomes essential to explore how the integration of semantic parsing could impact. Recently, Jin et al. (2024) investigates the role of semantic representation in LLMs by proposing AMRCOT, a method similar to that depicted in Fig.1 (b). Their findings reveal that introducing AMR results into the input generally harms LLM performance more than it helps, likely because AMR is not yet a representation well-suited for LLMs. However, this analysis remains limited, as it only considers the effects of AMR on several tasks, leaving the broader potential of semantic parsing in LLMs largely unexplored.

In this paper, we systematically investigate the impact of semantic parsing on LLMs to address the question: ***Can Semantic Information Still Contribute to Improve Downstream Tasks on LLMs?*** We empirically compare different paradigms for integrating semantic parsing into LLMs, as shown in Fig.1. These paradigms include approaches com-

[*] Equal contribution
[†] Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1020–1029
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

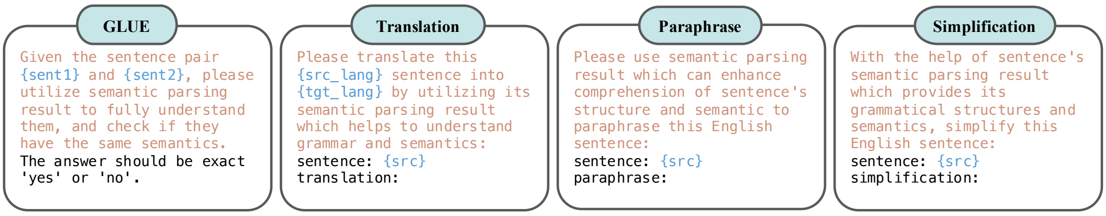

Figure 2: Illustration of SENSE designed for downstream tasks.

monly used for smaller models, such as incorporating semantic parsing results directly on the input side by fine-tuning or integrating them on the output side. However, these methods negatively affect model performance since they limit fixed types of semantic parsing and might introduce erroneous results. Thus, we propose a novel prompting approach, **SENSE**, illustrated in Fig.1 (d). Instead of injecting explicit parsing results, SENSE encourages LLMs to harness their internal semantic parsing capabilities through the addition of semantic hints. These hints are as simple as ***"please use semantic parsing result to enhance comprehension of the sentence's structure and semantics"***. Our comprehensive experiments demonstrate that SENSE promote LLM to focus more on key semantic information, not only achieves superior and consistent performance across various tasks, but also produces more linguistically aligned results, particularly on simplification and paraphrasing tasks, underscoring the effectiveness of semantic parsing for enhancing LLMs' performance.

2 Semantic Information → LLMs

In this section, we delve into answering the question: ***Can Semantic Information Still Contribute to Improve Downstream Tasks on LLMs?***

2.1 Methodology

Previous studies, such as those by Ettinger et al. (2023) and Jin et al. (2024), highlight the difficulty LLMs face in processing the schemes and symbols of explicit semantic parsing results. Their findings suggest that directly integrating these results can degrade model performance. Given that LLMs are already capable of achieving strong results in an end-to-end manner, we propose a novel approach: incorporating semantic parsing hints into the instruction to prompt LLMs to leverage their internal parsing capabilities.

As Fig.2 shows, our SENSE introduces simple semantic hints such as *"utilize semantic parsing result"* to *"fully understand input"* or *"capture grammatical structures and semantics"* to complete downstream tasks. This strategy encourages LLMs to engage in inherent understanding of linguistic structures without requiring explicit semantic parsing results. The workflow outlined in Fig.1 (d) demonstrates how semantic hints are integrated, and SENSE works in an zero-shot manner.

2.2 Datasets and Evaluation

In our experiments, we select seven understanding tasks from GLUE and three representative generation tasks including Machine Translation, Paraphrasing, and Simplification. Specifically, for paraphrasing task, we report three linguistic metrics across lexical, syntactic, and semantic levels, for simplification task, we report SARI and SAMSA which evaluate the predicted simplified sentences from lexical structure and semantic meaning preservation. More details about our experiments can be found in Appendix A.1 and A.2.

3 Experimental Results

3.1 Main Results

Results on Understanding Tasks From Table 1, the results demonstrate that although LLMs currently lag behind smaller models like BERT, the integration of SENSE significantly narrows this gap. Specifically, SENSE improves the average performance of GPT-4o-mini from 79.43% to 81.25%, bringing it closer to BERT's performance of 83.2%. Moreover, SENSE is effective in enhancing the performance of both closed-source models such as GPT-series, and open-source models like LLaMA3. Across all GLUE tasks, SENSE consistently yields performance gains, with notable improvements in MRPC (72.30% to 76.47%), MNLI (73.90% to 78.20%) and CoLA (65.49% to 67.22%). These results highlight SENSE's ability to enhance LLMs' comprehension of input sentences and demonstrate its robustness across diverse tasks.

Results on Paraphrasing Table 2 indicates that SENSE effectively enhances linguistic diversity in paraphrasing tasks while maintaining high semantic similarity. Notably, SENSE retains the

System	SST-2	MRPC	QQP	MNLI	QNLI	RTE	CoLA	Average
	Acc	Acc	Acc	Acc	Acc	Acc	Mcc	
BERT$_{\text{LARGE}}$ (2019)	93.20	88.00	91.30	86.60	92.30	70.40	60.60	83.20
RoBERTa$_{\text{LARGE}}$ (2019)	96.40	90.90	92.20	90.20	94.70	86.60	68.00	88.43
LLaMA3-70B	**95.64**	73.52	74.60	71.90	91.30	84.48	63.90	79.34
+ SENSE	95.18	74.04	76.50	73.10	92.80	85.56	65.53	80.25
GPT-3.5-turbo	91.86	73.28	73.40	61.80	82.40	81.81	63.50	75.44
+ SENSE	92.20	75.49	**77.20**	64.60	83.20	84.12	64.57	77.34
GPT-4o-mini	91.63	72.30	73.00	73.90	92.30	87.36	65.49	79.43
+ SENSE	92.08	**76.47**	73.00	**78.20**	**93.30**	**88.45**	**67.22**	**81.25**

Table 1: Experimental results on GLUE benchmark.

System	Prediction–Source		
	Semantic Similarity ↑	Lexical Overlap ↓	Syntactic Diversity ↑
LLaMA3-70B	83.71	30.00	10.85
+ SENSE	**84.02**	**29.00**	**11.51**
GPT-3.5-turbo	85.79	46.37	8.76
+ SENSE	**85.79**	**25.33**	**10.24**
GPT-4o-mini	89.71	39.00	7.25
+ SENSE	**90.26**	**34.00**	**8.08**

Table 2: Experimental results on Paraphrasing. We report linguistic metrics between source and prediction.

semantic similarity score at 90.26 but significantly reduces lexical overlap from 39.00 to 34.00 and increases syntactic diversity from 7.25 to 8.08. This indicates that the semantic hints introduced by SENSE lead to more diverse syntactic structures and reduced lexical repetition while preserving the core meaning of the source sentence, which validates the effectiveness of SENSE in generating paraphrases that are not only semantically faithful but also exhibit greater lexical and syntactic variety.

System	BLEU ↑	SARI ↑	SAMSA ↑
	TrukCorpus		
GPT-3.5-turbo	58.16	42.25	31.42
+ SENSE	**63.42**	**42.42**	**37.03**
	GoogleComp		
GPT-3.5-turbo	13.12	35.53	28.14
+ SENSE	**14.31**	**35.67**	**30.52**

Table 3: Experimental results on Simplification. We add two metrics, SARI and SAMSA to evaluate the semantic structure of the output.

Results on Simplification Table 3 showcases the improved performance of SENSE on two simplification datasets. Compared to the vanilla prompt, SENSE delivers higher BLEU scores of 63.42 on TrukCorpus and 14.31 on GoogleComp, alongside a modest increase in SARI, which evaluates the alignment between the source and target sentences. More importantly, the SAMSA scores, which measure the preservation of syntactic structure, show substantial improvement, reaching 37.03 and 30.52 respectively. These results demonstrate that integrating semantic hints into prompts enhances the model's ability to simplify sentences while preserving their original structure, resulting in more effective overall simplification.

Results on Machine Translation We further conduct experiments on Machine Translation task and present a comparative analysis of GPT-3.5-turbo across the vanilla prompt, our SENSE, and other state-of-the-art systems in Table 8. Results show that SENSE consistently enhances GPT-3.5 across all evaluated metrics and language pairs. For the DE-EN task, SENSE achieves the highest scores: COMET22 (86.44), ChrF (59.08), and BLEU (33.75), outperforming the WMT-Best system. Similarly, in the EN-DE task, SENSE significantly boosts GPT-3.5's performance, reaching COMET22 (86.65), ChrF (62.84), and BLEU (34.18). These improvements highlight the effectiveness of SENSE in enhancing GPT-3.5's ability to handle translation tasks across different language pairs. The results for ZH-EN and EN-ZH in Table 8 further confirm SENSE's effectiveness.

3.2 Analytical Results

Analysis of Different Paradigms In Table 4, we compare various approaches for incorporating semantic parsing into LLMs. We examine methods that either concatenate pre-generated parsing results using LLM or generate them on output side[1]. The results demonstrate that directly adding semantic parsing results degrades performance, aligning with findings by Jin et al. (2024). This degradation arises from the unfamiliar symbolic representation and the diversity of semantic parsing tasks, integrating specific type, and potentially erroneous results limits LLM's capability. In contrast, SENSE avoids explicit incorporation while consistently outperforming these methods. Such finding un-

[1]We do not specify certain type of semantic parsing during our experiments.

System	SST-2	MRPC	QQP	MNLI	QNLI	RTE	CoLA
GPT-3.5-turbo	91.86	73.28	73.40	61.80	82.40	81.81	63.50
+ CoT (2022)	89.11$_{-2.75}$	73.28$_{+0.00}$	77.00$_{+3.60}$	56.20$_{-5.60}$	82.70$_{+0.30}$	82.54$_{+0.73}$	64.32$_{+0.82}$
+ SP-Input	87.50$_{-4.36}$	74.26$_{+0.98}$	74.30$_{+0.90}$	50.50$_{-11.30}$	78.40$_{-4.00}$	84.11$_{+2.30}$	58.37$_{-5.13}$
+ SP-Output	89.11$_{-2.75}$	73.52$_{+0.24}$	71.90$_{-1.50}$	62.00$_{+0.20}$	78.40$_{-4.00}$	81.59$_{-0.22}$	64.44$_{+0.94}$
+ SENSE	92.20$_{+0.34}$	75.49$_{+2.21}$	77.20$_{+3.80}$	64.60$_{+2.80}$	83.20$_{+0.80}$	84.12$_{+2.31}$	64.57$_{+1.07}$

Table 4: Analysis of different approaches that introduce semantic parsing into LLMs on GLUE benchmark. Improvements are marked in red and decreases in green, relative to GPT-3.5-turbo.

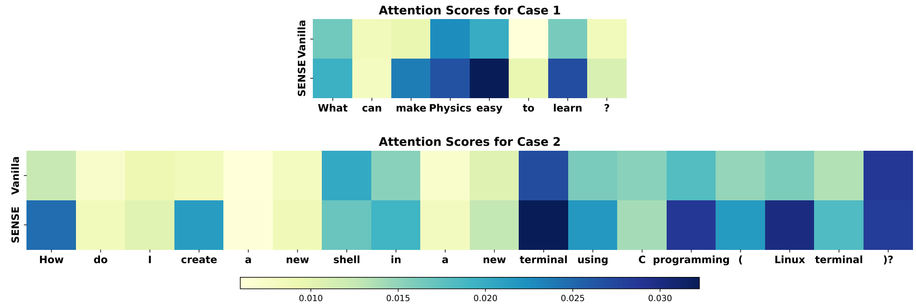

Figure 3: Visualization of attention scores from LLaMA3-70B on the source sentence in the Paraphrasing Task.

derscores SENSE as a more effective strategy for leveraging semantic parsing on LLMs.

Comparison with Chain-of-Thought Since SENSE shares similarities with CoT (Kojima et al., 2022), which works by adding "Let's think step by step", we compare it on GLUE (Table 4) and machine translation task (Table 8). While CoT degrades performance across tasks, as it is better suited for reasoning tasks, SENSE significantly enhances LLM performance by improving the model's ability to understand input sentences, thus yielding better results.

Visualization of Attention Scores We present the distribution of attention scores for paraphrasing task in Fig.3, where we average attention scores for each output token with respect to original sentence. The visualization reveals that, compared to vanilla prompt, SENSE places greater emphasis on key semantic elements, such as important lexical units and core components. This indicates that SENSE more effectively directs attention toward critical semantic information, and thus generates outputs that are more linguistic-aligned. Additionally, we provide case study on such examples in Table 9 and 10. While both vanilla prompt and SENSE successfully capture the paraphrased meaning, SENSE is superior at transforming syntactical structure and utilizing more diverse expressions.

4 Related Work

Semantic parsing has significantly contributed to enhancing the performance of smaller language models. Integrating results from SRL and AMR (Gildea and Jurafsky, 2002; Palmer et al., 2010; Banarescu et al., 2013) has shown to improve model performance on various tasks (Khashabi et al., 2022; Rapp, 2022; Xu et al., 2020; Si et al., 2022, 2024). However, the effectiveness of semantic parsing to LLMs is under-explored. Recent work, such as Jin et al. (2024), explores the use of AMR results with LLMs and finds that direct integration of these results may not always yield positive influences. Unlike approaches focused on optimizing prompts directly (Zhou et al., 2022; Pryzant et al., 2023; Deng et al., 2022; Guo et al., 2024), our work proposes a novel strategy for leveraging semantic parsing in LLMs. Similar to CoT (Wei et al., 2022; Kojima et al., 2022) and DTG (Li et al., 2023), our method involves integrating semantic parsing hints into prompts rather than optimizing the prompts.

5 Conclusion

In this paper, we rethink leveraging semantic parsing to enhance LLMs' performance. Contrary to smaller models, where direct integration of parsing results can be beneficial, we find that this negatively impacts LLMs. With the help of our proposed SENSE, which introduces semantic hints within prompts, LLMs can better comprehend input sentences. Experiments show that SENSE achieves great performance across both understanding and generation tasks, and helps models capture lexical and syntactic structures, producing outputs that align more closely with linguistic metrics.

Limitations

While we validate the effectiveness of SENSE across both understanding and generation tasks, there are limitations that remain for future exploration: Firstly, our validation is restricted to the LLaMA and GPT-series models. Extending SENSE to other LLM architectures will be necessary to confirm its general applicability. Secondly, although SENSE shows promising results on a range of NLP tasks, its performance across more diverse datasets and applications needs further investigation. Our experiments focus on tasks where the benefits of semantic parsing have been established, but broader testing is required to fully assess its potential. Additionally, the underlying mechanism of how semantic parsing influences LLM decision-making remains unclear, as LLMs function largely as black-box systems. Our validation primarily involves comparing methods that directly incorporate semantic parsing results from the input or output sides, and analyzing the outputs in contrast to both the vanilla prompt and SENSE.

Acknowledgements

We thank all reviewers for their great efforts. This work is supported by the National Science Foundation of China under Grant No.61876004 and 61936012.

References

Kaikai An, Ce Zheng, Bofei Gao, Haozhe Zhao, and Baobao Chang. 2023. Coarse-to-fine dual encoders are better frame identification learners. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13455–13466, Singapore. Association for Computational Linguistics.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic. 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. *Linguistic Annotation Workshop,Linguistic Annotation Workshop*.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-amr: abstract meaning representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Daniel Gildea and Dan Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *Preprint*, arXiv:2309.08532.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023. ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Fernando Gonzalez, 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. *arXiv preprint arXiv:2405.01502*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2022. Question answering as global reasoning over semantic abstractions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Bei Li, Rui Wang, Junliang Guo, Kaitao Song, Xu Tan, Hany Hassan, Arul Menezes, Tong Xiao, Jiang Bian, and JingBo Zhu. 2023. Deliberate then generate: Enhanced prompting framework for text generation. *arXiv preprint arXiv:2305.19835*.

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

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`. Accessed: 2023-04-01.

Martha Palmer, Ivan Titov, and Shumin Wu. 2010. Semantic role labeling. *Computational Linguistics*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Reinhard Rapp. 2022. Using semantic role labeling to improve neural machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3079–3083.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2024. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36.

Shuzheng Si, Shuang Zeng, and Baobao Chang. 2022. Mining clues from incomplete utterance: A query-enhanced network for incomplete utterance rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4839–4847, Seattle, United States. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic structural evaluation for text simplification. *arXiv preprint arXiv:1810.05022*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

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

Kun Xu, Haochen Tan, Linfeng Song, Han Wu, Haisong Zhang, Linqi Song, and Dong Yu. 2020. Semantic role labeling guided multi-turn dialogue rewriter. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6632–6639.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Supplementary Details

A.1 Details about Datasets

We list the details of each dataset, including source, number, and metrics for each task in Table 5, and we sample a subset of data if the original dataset is large to reduce the API cost.

Dataset	Num.	Metrics
SST-2	872	Acc
MRPC	408	Acc
QQP	1000	Acc
MNLI	1000	Acc
QNLI	1000	Acc
RTE	277	Acc
CoLA	1053	Mcc
WMT DE-EN	1984	BLEU, COMET22, Chrf
WMT EN-DE	1875	BLEU, COMET22, Chrf
WMT ZH-EN	1875	BLEU, COMET22, Chrf
WMT EN-ZH	1875	BLEU, COMET22, Chrf
QQP	2500	Lexical, Syntactic, Semantic
TurkCorpus	359	BLEU, SARI, SAMSA
GoogleComp	1000	BLEU, SARI, SAMSA

Table 5: Statistics of the dataset we use in our experiment.

GLUE We test on seven tasks from GLUE benchmark (Wang et al., 2019) and report the Matthews Correlation Coefficient (MCC) for CoLA and Accuracy (Acc) for the left tasks.

Machine Translation For machine translation, we evaluate our method on the WMT22 [2] dataset, focusing on two language pairs: EN-DE (English to German) EN-ZH (English to Chinese) and report COMET22 (Rei et al., 2022), CHRF, and BLEU scores [3].

Paraphrasing We evaluate on Quora Question Pairs (QQP) [4] dataset. To analyze results professionally, we follow Huang et al. (2023) and report three linguistic evaluation metrics across lexical, syntactic, and semantic levels.

Simplification For text simplification, we evaluate on TurkCorpus and GoogleComp and use BLEU, SARI, and SAMSA as the evaluation metrics. Specifically, SARI [5] (System output Against References and against the Input sentence) is used to compare the predicted simplified sentences against the reference and the source sentences, and SAMSA (Sulem et al., 2018) is a metric specifically designed for text simplification that evaluates structural simplification and meaning preservation.

A.2 Details about Experiment

A.2.1 Experimental Setup

We test our SENSE on GPT-3.5-turbo, GPT-4o-mini (OpenAI, 2023) with the version of 2023-11-06 and 2024-07-18, and LLaMA3-70B-Instruct [6]. The temperature is set to 0 and top_p set to 1.

A.2.2 Prompts used in Experiments

We release the prompts we use during our experiments in Table 6 and Table 7.

A.3 Additional Experimental Results

Results on WMT22 From Table 8, for the ZH-EN translation task, SENSE improves GPT-3.5-turbo's ChrF (58.50) and BLEU (27.04) scores, though the COMET22 score (80.47) is slightly lower than the baseline. In the EN-ZH task, SENSE achieves the highest COMET22 (88.06) and enhances ChrF (39.86) and BLEU (44.40) compared to baselines.

Case Study In Tables 9 and 10, we present case studies on paraphrasing and inference tasks. These demonstrate that SENSE not only excels in altering syntactic structures and employing a broader range of expressions, thereby enhancing the overall quality of paraphrasing, but also better captures sentence semantics.

[2] https://machinetranslate.org/wmt22

[3] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a

[4] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[5] https://huggingface.co/spaces/evaluate-metric/sari

[6] https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3

Dataset	Method	Prompt
SST-2	Vanilla	Given this sentence: {sentence}, please classify its sentiment as positive or negative. The answer should be exactly 'positive' or 'negative'.
	CoT	Given this sentence: {sentence}, please think step by step, and then classify its sentiment as positive or negative. The answer should be exactly 'positive' or 'negative'.
	SP-Input	Given this sentence: {sentence} and its semantic parsing result {parsing}, please classify the sentence's sentiment as positive or negative. The answer should be exactly 'positive' or 'negative'.
	SP-Output	Given this sentence: {sentence}, please first parse this sentence and then classify the sentence's sentiment as positive or negative. The answer should be exactly 'positive' or 'negative'.
	SENSE	Given this sentence: {sentence}, please use semantic parsing result which can enhance comprehension of the sentence's structure and semantics to classify the sentence's sentiment. The answer should be exactly 'positive' or 'negative'.
MRPC	Vanilla	Given the sentence pair {sentence1} and {sentence2}, please check if these two sentences have the same semantics. The answer should be exactly 'yes' or 'no'.
	CoT	Given the sentence pair {sentence1} and {sentence2}, please think step by step, and then check if these two sentences have the same semantics. The answer should be exactly 'yes' or 'no'.
	SP-Input	Given the sentence pair {sentence1} and {sentence2} and their semantic parsing results {parsing1} and {parsing2}, please check if these two sentences have the same semantics. The answer should be exactly 'yes' or 'no'.
	SP-Output	Given the sentence pair {sentence1} and {sentence2}, please first parse these sentences and then check if these two sentences have the same semantics. The answer should be exactly 'yes' or 'no'.
	SENSE	Given the sentence pair {sentence1} and {sentence2}, please use semantic parsing result which can enhance comprehension of the sentence's structure and semantics to measure if these two sentences have the same semantics. The answer should be exactly 'yes' or 'no'.
MNLI	Vanilla	Given the sentence1 {premise} and sentence2 {hypothesis}, determine whether sentence2 entail, contradict, or is it neutral to sentence1. The answer should be exactly 'entail' or 'contradict' or 'neutral'.
	CoT	Given the sentence1 {premise} and sentence2 {hypothesis}, please think step by step, and then determine whether sentence2 entail, contradict, or is it neutral to sentence1. The answer should be exactly 'entail' or 'contradict' or 'neutral'.
	SP-Input	Given the sentence1 {premise} and sentence2 {hypothesis} and their semantic parsing results {parsing1} and {parsing2}, please determine whether sentence2 entail, contradict, or is it neutral to sentence1. The answer should be exactly 'entail' or 'contradict' or 'neutral'.
	SP-Output	Given the sentence1 {premise} and sentence2 {hypothesis}, please first parse these sentence to fully understand its structure and semantics and then determine whether sentence1 entail, contradict, or is neutral to sentence2. The answer should be exactly 'entail' or 'contradict' or 'neutral'.
	SENSE	Given the sentence1 {premise} and sentence2 {hypothesis}, please use semantic parsing result which can enhance comprehension of the sentence's structure and semantics to determine whether sentence1 entail, contradict, or is neutral to sentence2. The answer should be exactly 'entail' or 'contradict' or 'neutral'.
QNLI	Vanilla	Given the sentence1 {question} and sentence2 {sentence}, please determine if the sentence contains the answer to the question. The answer should be exactly 'entail' or 'not entail'.
	CoT	Given the sentence1 {question} and sentence2 {sentence}, please think step by step, and then determine if the sentence contains the answer to the question. The answer should be exactly 'entail' or 'not entail'.
	SP-Input	Given the sentence1 {question} and sentence2 {sentence} and their semantic parsing results {parsing1} and {parsing2}, please determine if the sentence contains the answer to the question. The answer should be exactly 'entail' or 'not entail'.
	SP-Output	Given the sentence1 {question} and sentence2 {sentence}, please first parse these sentences and then determine if the sentence contains the answer to the question. The answer should be exactly 'entail' or 'not entail'.
	SENSE	Given the sentence1 {question} and sentence2 {sentence}, please use semantic parsing result which can enhance comprehension of the sentence's structure and semantics to determine if the sentence contains the answer to the question. The answer should be exactly 'entail' or 'not entail'.
CoLA	Vanilla	Given the sentence: {sentence}, please check if the sentence is grammatically correct. The answer should be exactly 'yes' or 'no'.
	CoT	Given the sentence: {sentence}, please think step by step, and then check if the sentence is grammatically correct. The answer should be exactly 'yes' or 'no'.
	SP-Input	Given the sentence: {sentence} and its semantic parsing result {parsing}, please check if the sentence is grammatically correct. The answer should be exactly 'yes' or 'no'.
	SP-Output	Given the sentence: {sentence}, please first parse this sentence and then check if the sentence is grammatically correct. The answer should be exactly 'yes' or 'no'.
	SENSE	Given the sentence: {sentence}, please use semantic parsing result which can enhance comprehension of the sentence's structure and semantics to check if the sentence is grammatically correct. The answer should be exactly 'yes' or 'no'.

Table 6: We list the prompts we use during our experiments on GLUE benchmarks and omit QQP and RTE since QQP is similar to MRPC and RTE is similar to MNLI.

Dataset	Method	Prompt
WMT22	Vanilla	Please translate this {src_lang} sentence into {tgt_lang}: sentence: {src} translation:
	SENSE	Please translate this {src_lang} sentence into {tgt_lang} by utilizing its semantic parsing result which helps to understand grammar and semantics: sentence: {src} translation:
Simplification	Vanilla	Please simplify this English sentence: sentence: {src} simplification:
	SENSE	With the help of the sentence's semantic parsing result which provides its grammatical structures and semantics, simplify this English sentence: sentence: {src} simplification:
Paraphrasing	Vanilla	Please paraphrase this English sentence: sentence: {src} paraphrase:
	SENSE	Please use semantic parsing result which can enhance comprehension of sentence's structure and semantic to paraphrase this English sentence: sentence: {src} paraphrase:

Table 7: We list the prompts we use during our experiments on generation tasks.

System	DE-EN			EN-DE		
	COMET22 ↑	ChrF ↑	BLEU ↑	COMET22 ↑	Chrf ↑	BLEU ↑
WMT-Best	85.00	58.50	33.40	87.20	64.60	38.40
GPT EVAL (2023)	84.80	58.30	33.40	84.20	59.60	30.90
DTG 5-shot (2023)	85.40	58.20	33.20	86.30	61.60	33.40
BayLing (2023)	85.47	58.65	32.94	**86.93**	62.76	34.12
GPT-3.5-turbo	85.71	58.19	33.15	84.60	60.48	33.42
+ CoT	84.99	57.74	31.46	84.95	61.17	29.70
+ SENSE	**86.44**	**59.08**	**33.75**	86.65	**62.84**	**34.18**
System	**ZH-EN**			**EN-ZH**		
	COMET22 ↑	ChrF ↑	BLEU ↑	COMET22 ↑	Chrf ↑	BLEU ↑
WMTBest	81.00	61.10	33.50	86.70	41.10	44.80
GPT EVAL (2023)	81.20	56.00	25.90	84.40	36.00	40.30
DTG 5-shot (2023)	81.70	55.90	25.20	86.60	39.40	43.50
BayLing (2023)	**82.64**	57.90	26.13	86.81	**40.32**	**44.99**
GPT-3.5-turbo	80.60	58.40	26.93	81.48	37.80	42.85
+ SENSE	80.47	**58.50**	**27.04**	**88.06**	39.86	44.40

Table 8: Experimental results on WMT22.

Source: What can make Physics easy to learn?
Target: How can you make physics easy to learn?
Vanilla Prediction: What makes learning Physics easier?
SENSE Prediction: What factors can help simplify the learning of Physics?

Source: How do I create a new shell in a new terminal using C programming (Linux terminal)?
Target: How do I create a new Terminal and new shell in Linux using C programming?
Vanilla Prediction: How do I open a new shell in a new terminal window using C programming in a Linux environment?
SENSE Prediction: What steps should I follow to launch a new shell in a separate terminal using C programming on a Linux system?

Table 9: Case study comparing SENSE and Vanilla prompt predictions for the paraphrasing task.

Premise: My article does not say or imply that real earnings growth only reflects retentions and that dividend growth must be zero or that all valuation techniques are out the window for firms that don't pay dividends. **Hypothesis:** My article doesn't say or imply that real earnings growth reflects only retentions and that dividend growth must be zero or that valuation techniques are unused for firms which don't pay dividends, it implies that growth is a concrete and calculable statistic. **Ground Truth:** Neutral **Vanilla Prediction:** Contradict **SENSE Prediction:** The semantic parsing result of both sentences indicates that they are expressing the same idea, with some minor differences in wording. Therefore, the two sentences are neutral to each other. **Analysis:** In this case, SENSE correctly identifies that both sentences express the same idea with minor wording differences, thus labeling them as neutral. The baseline prediction incorrectly identifies a contradiction.
Premise: Is there a favorite story or storybook or anything that you had when you were little that somebody read to you? **Hypothesis:** You don't remember your childhood, do you? **Ground Truth:** Contradiction **Vanilla Prediction:** Sentence1 entails sentence2. If someone is asking about a favorite story or storybook from childhood, it implies that they believe the person has memories of being read to as a child. **SENSE Prediction:** Based on the semantic parsing result, sentence1 is neutral to sentence2. The first sentence is asking about a favorite story from childhood, while the second sentence is questioning the person's memory of their childhood. There is no direct contradiction or entailment between the two sentences. **Analysis:** SENSE detects that the first sentence is asking about a favorite story from childhood, while the second sentence is questioning the person's memory of their childhood. Since it regards the second sentence as a question without an answer, it believes that there is no direct contraction or entailment, but regardless, this example should not be the implied relationship as entailment predicted by the baseline.

Table 10: Case study comparing SENSE and Vanilla prompt predictions for the Natural Language Inference task.