Exploring the Role of Semantic Parsing on Downstream Tasks for Large Language Models

Anonymous EMNLP submission

Abstract

Semantic Parsing focuses on converting sentences into structured forms. While previous studies show its benefits for smaller models, the impact on Large Language Models (LLMs) remains under explored. Our paper explores whether integrating Semantic Parsing can enhance LLMs' performance in downstream tasks. Unlike prior approaches, we propose SENSE , adding semantic parsing hint instead results into prompt and find that this approach consistently improves performance across tasks, highlighting the potential of semantic information integration in enhancing LLM capabilities.

1 Introduction

Semantic Parsing is a fundamental and crucial task in Natural Language Processing (NLP), which involves converting a natural language sentence into a logical form, including tasks such as 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). The goal of semantic parsing is to capture the meaning of the sentence in a structured representation that can be used for various tasks such as Question Answering (Khashabi et al., 2022), Machine Translation (Rapp, 2022), Dialogue Systems (Xu et al., 2020; Bonial et al., 2020) and so on.

Previous works like Bonial et al. (2020); Rapp (2022); Khashabi et al. (2022) demonstrate that the introduction of semantic information from SRL or AMR can effectively enhance the ability of the model to grasp illocutionary and linguistic abstractions, and thereby improve the performance of downstream tasks. However, these findings have been predominantly limited to smaller-scale models like BERT (Devlin et al., 2019). With the emergence of Large Language Models (LLMs), researchers are more willing to evaluate the performance of downstream tasks on LLMs. Even though LLMs achieve remarkable performance in an end-to-end manner, it remains an interesting question to explore the potential contribution of integrating Semantic Parsing into LLMs. Ettinger et al. (2023) shows that even though LLMs have acquired sufficient knowledge of AMR parsing and semantic structure for reliable generation of basic AMR format, however, the model are not currently sufficient out-of-the-box to yield reliable and accurate analyses of abstract meaning structure. Furthermore, Jin et al. (2024) investigates the role of semantic representation in the era of LLMs by proposing the AMR-driven chain-of-thought, adhere to in Fig. 1 (a). Consistent with Ettinger et al. (2023), they find that AMRCOT generally hurts the performance more than it helps, and explain that it is caused by AMR is not yet a representation immediately fit for LLMs.

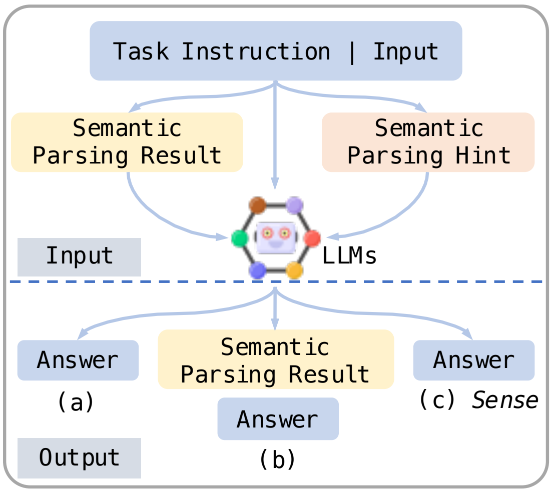

Figure 1: Different methods for introducing Semantic Parsing into LLMs. (a) and (b) directly incorporate semantic parsing results into input or output, while (c), our SENSE, just adds the semantic parsing hint into the prompt and avoids the direct perception of the result.

In our paper, we seek to explore the following question: ***Can Semantic Parsing Still Con-***

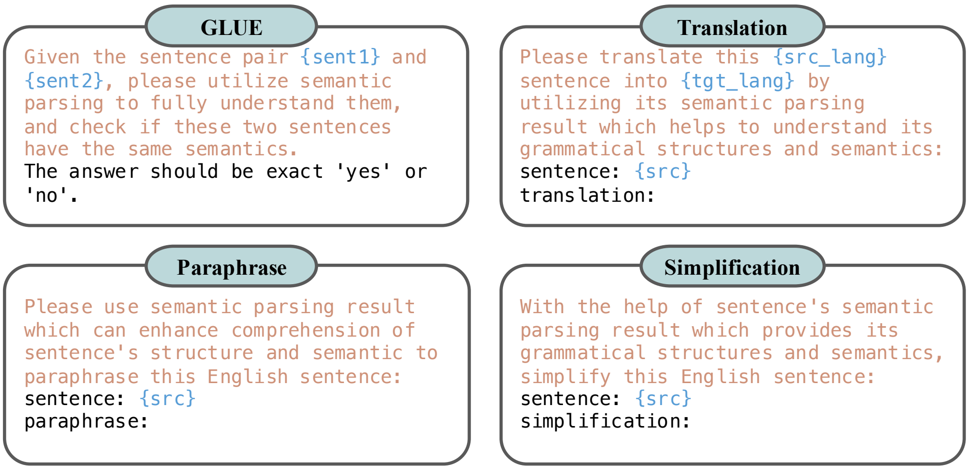

Figure 2: Illustration of SENSE designed for downstream tasks. We list the instruction we use for GLUE (QQP), Machine Translation, Paraphrase and Simplification.

tribute to the Improvement of Downstream Tasks on LLMs? Different from Jin et al. (2024), we propose a novel prompting schema, **SENSE** , just shown in Fig. 1 (c) we do not directly introduce the semantic parsing result into the input or output, instead we only suggesting LLMs should utilize their semantic parsing capabilities to help themselves in downstream tasks. The prompt schema is just as simple as **"*please use semantic parsing result which can enhance comprehension of the sentence's structure and semantic*"**. We evaluate SENSE on both understanding and generation tasks and test the generation task on linguistic metrics. By directly infusing semantic parsing information into the prompt, SENSE consistently yields performance gains and better semantic evaluation metrics. We examine the impact of varying depths of semantic parsing and discover that more comprehensive parsing encapsulates wider sentence information and achieves superior performance. In addition, to thoroughly assess the influence of semantic parsing, we contrast the effects of incorporating parsing results into prompts. Our findings indicate that the direct integration of higher-quality semantic information correlates with degraded task performance.

2 Realted Work

From small language models on, a large of works utilize semantic parsing results to help models better grasp the structure and illocutionary of the text, including Question Answering (Shen and Lapata, 2007; Khashabi et al., 2022), Machine Translation (Bazrafshan and Gildea, 2013; Rapp, 2022), Dialogue Systems (Chen et al., 2013; Xu et al., 2020; Bonial et al., 2020), and gain great performance. With the widespread of LLMs, prompt engineering has received widespread attention. The effectiveness of a language model in performing a task is significantly influenced by how the input prompt is structured and researchers now concentrate on the optimization of discrete prompts, utilizing such as model feedback (Zhou et al., 2022; Pryzant et al., 2023), reinforcement learning (Deng et al., 2022) or evolutionary algorithms (Guo et al., 2023) to search for better prompts. However, while smaller models indicate that semantic parsing can improve model performance, highlighting a significant opportunity in this field, we explore the role of semantic parsing for LLMs. Different from Jin et al. (2024), we do not investigate the role of semantic representations by directly inputting the result of AMR into input, we are investigating the role of semantic parsing in the helpfulness of downstream tasks, as the smaller models do. By incorporating semantic parsing hints into the prompt, our SENSE can achieve consistent improvement on downstream tasks.

3 Semantic Parsing $\rightarrow$ LLMs

In this section, we delve into answering the question: ***Can Semantic Parsing Still Contribute to the Improvement of Downstream Tasks on LLMs?*** We first introduce the methodology of SENSE , then give the experiment details, and at last show the experimental results of our method.

3.1 Methodology

As Ettinger et al. (2023); Jin et al. (2024) shown, it is difficult for LLMs to better grasp the schemes

System	SST-2	MRPC	QQP	MNLI	QNLI	RTE	CoLA	Average
	Acc	Acc	Acc	Acc	Acc	Acc	Mcc	
$\text{BERT}_{\text{LARGE}}$ (2018)	93.20	88.00	91.30	86.60	92.30	70.40	60.60	83.20
$\text{RoBERTa}_{\text{LARGE}}$ (2019)	96.40	90.90	92.20	90.20	94.70	86.60	68.00	88.43
GPT-3.5-turbo	91.86	73.28	73.40	61.80	82.40	81.81	63.50	75.44
+ CoT (2022)	89.11	73.28	77.07	56.20	82.70	82.54	64.32	75.03
+ SENSE	92.20	75.49	77.19	64.60	83.20	84.12	64.57	77.34

Table 1: Experiment results on GLUE benchmark.

and symbols of semantic parsing results. From their conclusions, directly ingesting the semantic parsing result will hurt the model performance. Since LLM itself is able to achieve good performance in an end-to-end manner, we propose to add the semantic parsing hint into the instruction to remind LLM to use its semantic parsing capabilities to complete the tasks.

As Fig. 2 shows, our SENSE directly adds hints like "utilize semantic parsing result" to "fully understand the input" or "capture the grammatical structures and semantics" to complete downstream tasks. We propose the flow in Fig. 1 (c) to utilize semantic parsing to improve the performance of downstream tasks.

3.2 Datasets and Evaluation

In our experiments, we select 7 understanding tasks from GLUE and 3 representative generation tasks including Machine Translation, Paraphrase, and Simplification. We summarize the details of each dataset, including source, number, and metrics for each task in Table 5 and test our SENSE on GPT-3.5 (OpenAI, 2023) with temperature of 0 and top_p of 1.

GLUE We test on seven tasks from GLUE [1] benchmark 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.

Paraphrase We evaluate on the Quora Question Pairs (QQP) [3] dataset. To validate that semantic parsing helps the model output, we follow Huang et al. (2024) 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 [4] (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.

3.3 Experimental Results

Results on Understanding Tasks From Table 1, we can see that GPT-3.5 falls behind the small models. When enhanced with our proposed SENSE , it shows a significant improvement, achieving an average accuracy of 77.34%, which is a notable gain over the vanilla GPT-3.5 of 75.44% and also higher than GPT-3.5 with CoT (75.03Specifically, SENSE consistently enhances performance in several tasks, such as MNLI (from 61.80% to 64.60%), RTE (from 81.81% to 84.12%), and so on. This demonstrates the effectiveness of SENSE in improving the model's ability to understand sentences. While CoT might degrade the performance on SST-2 and MNLI, we find that CoT tends to generate ambiguous or unsure answers at that time.

Results on Machine Translation We compare the performance of GPT-3.5 with vanilla prompting, our SENSE , and other state-of-the-art (SoTA) systems in Table 2. The results indicate that our SENSE consistently improves the performance of GPT-3.5 across all evaluated metrics and language pairs. For DE-EN, SENSE achieves the highest scores: COMET22 (86.44), ChrF (59.08), and

[1] https://gluebenchmark.com/

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

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

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

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
+ SENSE	**86.44**	**59.08**	**33.75**	86.65	**62.84**	**34.18**

Table 2: Experiment results on WMT22.

System	Prediction–Source		
	Semantic Similarity ↑	Lexical Overlap ↓	Syntactic Diversity ↑
GPT-3.5-turbo	85.79	46.37	8.76
+ SENSE	**85.79**	**25.33**	**10.24**

Table 3: Experiment results on Paraphrase. We evaluate the linguistic metrics between the source and prediction to validate the advantage of utilizing semantic parsing.

System	BLEU ↑	SARI ↑	SAMSA ↑
TrukCorpus			
MUSS (2020)	**63.76**	40.85	-
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 4: Experiment results on Simplification. We add two metrics, SARI and SAMSA to evaluate the semantic structure of the output.

BLEU (33.75), outperforming the WMT-Best system and other baselines. Similarly, in the EN-DE task, SENSE enhances GPT-3.5, yielding scores close to the WMT-Best system: COMET22 (86.65), ChrF (62.84), and BLEU (34.18). And we present the results of ZH-EN and EN-ZH in Table 6. The consistent improvements across different language pairs highlight the effectiveness of SENSE .

Results on Paraphrase Table 3 indicates that our SENSE can generate more linguistic paraphrases compared with the source sentence. We can see that while SENSE retains the semantic similarity at 85.79, it significantly reduces the lexical diversity from 46.37 to 25.33 and enhances the syntactic diversity from 8.76 to 10.24, suggesting the semantic parsing hint helps to improve more lexical variety and better syntactic variation. These improvements demonstrate the effectiveness of SENSE in enhancing paraphrase by keeping the semantic information while diverse lexical and syntactic structures.

Results on Simplification Table 4 illustrates that LLM demonstrates better performance across both simplification datasets, surpassing existing methods such as MUSS. Specifically, SENSE significantly improves performance, achieving BLEU scores of 63.42 on TrukCorpus and 14.31 on GoogleComp. SARI scores improve to 42.42 and 35.67, while SAMSA scores show notable improvement to 37.03 and 30.52 respectively, proving that incorporating the semantic parsing hint into the prompt can help the model keep the original structure for simplification task.

Analysis of Directly Digesting Semantic Parsing Result into Input From Table 7, we can see that directly digesting semantic parsing results into input does hurt the model performance with a sharp degradation to 72.48%. The reason exists that directly incorporating specific schemes and symbols of semantic parsing is hard for LLMs to follow, and thus perform worse.

Analysis of Varying Depths of Semantic Parsing Table 8 shows the impact of varying depths of semantic parsing and more comprehensive parsing like FSP encapsulates wider sentence information and achieves superior performance.

4 Conslusion

In our paper, we investigate the potential of Semantic Parsing to enhance Large Language Models in various NLP tasks. Through our proposed SENSE approach, which prompts LLMs to leverage internal semantic parsing capabilities, we have demonstrated consistent performance improvements across understanding and generation tasks. This underscores the value of integrating semantic hints in enhancing LLMs' ability to comprehend and generate language with greater semantic fidelity.

Limitations

As we validate the effectiveness of our SENSE on both understanding and generation tasks, it still has some limitations for future research :

Firstly, the effectiveness of SENSE is validated within the capabilities and constraints of GPT-3.5. Generalizing these findings to other LLMs can further validate our approach.

Secondly, while SENSE demonstrates promising results across a spectrum of NLP tasks, its general ability across diverse datasets and applications requires further exploration. We just test tasks that previous works validate the effectiveness of semantic parsing. More tasks need to be verified.

Moreover, the interpretability of how semantic parsing information influences LLM decisions remains an ongoing issue. Clarifying and controlling these interactions are essential for ensuring transparent and reliable model behavior in practical applications.

References

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*.

Marzieh Bazrafshan and Daniel Gildea. 2013. Semantic roles for string to tree machine translation. *Meeting of the Association for Computational Linguistics,Meeting of the Association for Computational Linguistics*.

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.

Yun-Nung Chen, William Yang Wang, and Alexander I Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 120–125. IEEE.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv 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*.

Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. 2024. Queryagent: A reliable and efficient reasoning framework with environmental feedback based self-correction.

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).

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*.

Louis Martin, Angela Fan, Éric De La Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

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.

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.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.

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

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 about Dataset

Table 5 shows the statistics of the dataset we use, and we sample a subset of data if the original dataset is huge 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, Syntactical, Semantic
TurkCorpus	359	BLEU, SARI, SAMSA
GoogleComp	1000	BLEU, SARI, SAMSA

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

B Supplementary Experimetal Results

B.1 Results on WMT22

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 the baselines.

B.2 Directly Digesting Semantic Parsing Result into Input

Table 7 shows the results of directly giving the semantic parsing results into the input.

B.3 Varying depths of Semantic Parsing

Table 8 shows the results of incorporating varying depths of semantic parsing hints.

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 6: Experiment results on WMT22.

System	SST-2	MRPC	QQP	MNLI	QNLI	RTE	CoLA	Average
	Acc	Acc	Acc	Acc	Acc	Acc	Mcc	
GPT-3.5-turbo	91.86	73.28	73.40	61.80	82.40	81.81	63.50	75.44
+ SP Result	87.50	74.26	74.27	50.50	78.40	84.11	58.37	72.48
+ SENSE	92.20	75.49	77.19	64.60	83.20	84.12	64.57	77.34

Table 7: Extensive experiment results on GLUE benchmark.

Dataset	Not Specific			SRL			FSRL		
	BLEU	SARI	SAMSA	BLEU	SARI	SAMSA	BLEU	SARI	SAMSA
GoogleComp	14.31	35.67	30.52	16.31	36.13	34.00	16.55	35.43	34.94

Table 8: Extensive experiment results of incorporating varying depths of semantic parsing hints into the prompt.