HW-TSC at SemEval-2024 Task 5: Self-Eval? A Confident LLM System for Auto Prediction and Evaluation for the Legal Argument Reasoning Task

Xiaofeng Zhao[1], **Xiaosong Qiao**[1], **Min Zhang, Chang Su, Yuang Li, Yinglu Li, Yilun Liu Feiyu Yao, Xiaowei Liang, Shimin Tao, Hao Yang, Yanfei Jiang, Yunfei Lu, Dandan Tu**
Huawei Translation Services Center, Beijing, China
{zhaoxiaofeng14, qiaoxiaosong, zhangmin186, suchang8, liyuang3, liyinglu, liuyilun3, yaofeiyu1, liangxiaowei2, taoshimin, yanghao30, jiangyanfei, luyunfei6, tudandan}@huawei.com

Abstract

In this article, we present an effective system for semeval-2024 task 5. The task involves assessing the feasibility of a given solution in civil litigation cases based on relevant legal provisions, issues, solutions, and analysis. This task demands a high level of proficiency in U.S. law and natural language reasoning. In this task, we designed a self-eval LLM system that simultaneously performs reasoning and self-assessment tasks. We created a confidence interval and a prompt instructing the LLM to output the answer to a question along with its confidence level. We designed a series of experiments to prove the effectiveness of the self-eval mechanism. In order to avoid the randomness of the results, the final result is obtained by voting on three results generated by the GPT-4. Our submission was conducted under zero-resource setting, and we achieved first place in the task with an F1-score of 0.8231 and an accuracy of 0.8673.

1 Introduction

In 2023, a significant event in the field of artificial intelligence (AI) was the widespread adoption of ChatGPT, particularly the introduction of GPT-4 (OpenAI, 2023), which revolutionized perceptions of AI. GPT-4 exhibited a notable advancement of 11.2 points on the MMLU benchmark (Hendrycks et al., 2021) and demonstrated superior performance on various question answering (QA) and natural language inference (NLI) datasets. Large-scale language models (LLM) represented by GPT-4 have sprung up, including LLaMa-2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), Gemini (Anil et al., 2023), Baichuan-2 (Yang et al., 2023) , ChatGLM (Du et al., 2022), etc. There are many researchers have explored NLP task leveraging GPT-4 in zero-resource and low-resource scenarios. GPT-4 is pretrained on a large amount of Internet data initially, and refined through supervised fine-tuning and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). Despite these advancements, limited research exists on the direct application of GPT-4 to NLP tasks within the legal domain. This paper aims to comprehensively address this research gap.

2 Task Description

This task aims to assess system's ability to reason about legal arguments. The task organizers have introduced a dataset (Bongard et al., 2022) of civil litigation cases from the U.S. legal system. Each instance comprises of an overview of the case, a question, a proposed solution (an answer candidate), and analysis justifying the solution. Systems are required to determine if the solutions and analysis are correct (True) or incorrect (False). While similar to a typical classification task, this task demands strong causal reasoning and practical application of knowledge. As shown in Figure 1, evaluating the correctness of an answer candidate demands not only a logical assessment of the question and response but also the application of legal knowledge provided in the introduction.

Further, this task requires expertise with legal terminology and concepts. An experienced law professor, armed with deep understanding of relevant legal statutes and extensive knowledge in the field, would likely be able to swiftly assess the accuracy of answer candidates and the soundness of their analysis, even with minimal background information. Conversely, for those less familiar with the field, even being provided with comprehensive information, identifying key details and reaching the correct conclusion remains a challenging task. Notably, the training dataset and development dataset provide analysis of the labels, whereas the test dataset does not.

This dataset is extracted from real law teaching books and includes a total of 666 training sets, 84 development sets, and 98 test sets. The training set and development set provide analysis of labels, but

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1806–1810

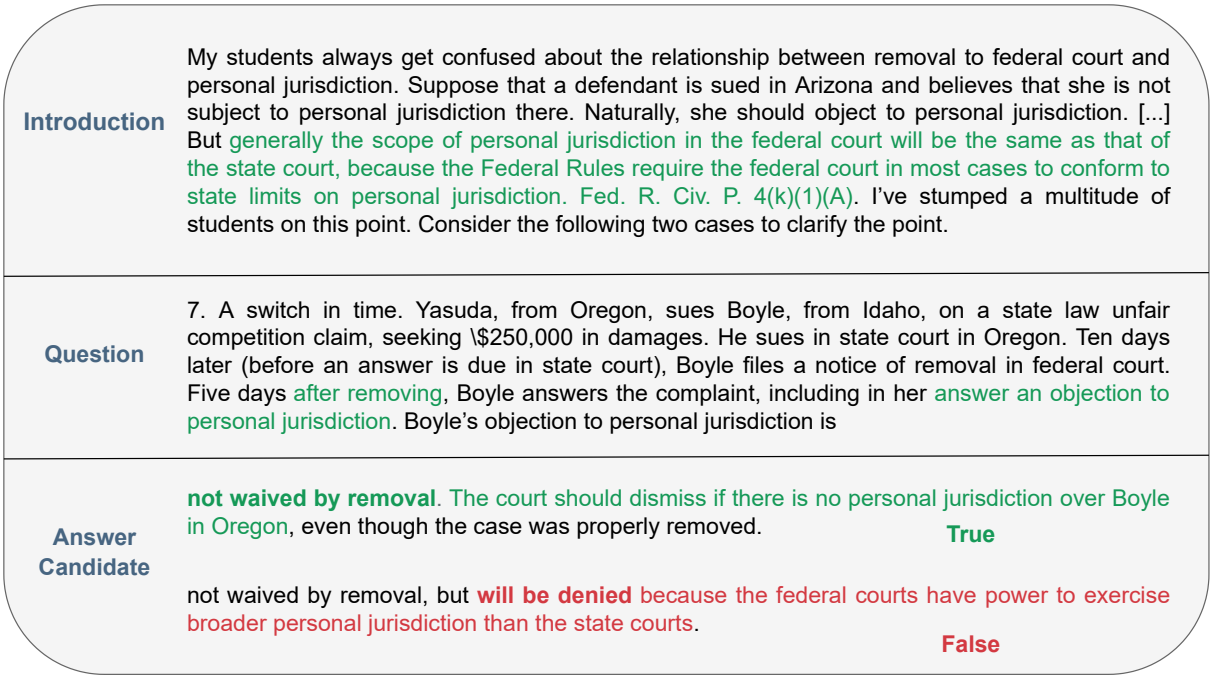

Figure 1: Data Example

the test set is not provided. The goal is to predict the label of the test set.

3 System

3.1 Method Overview

For this task, we have designed a Self-Eval LLM system that utilizes LLM (e.g. GPT-4) for reasoning to obtain answers. However, the responses generated by LLM can sometimes be ambiguous. Therefore, we have incorporated a confidence detection task to enable the LLM to evaluate the reliability of its own answers, and stimulate the LLM's potential. We use one specifically designed prompt for the model to perform both tasks — judging answer candidate and providing answer confidence. Furthermore, we have employed two strategies: converting judgments into selections and ensemble learning. As shown in Figure 2, we depict a workflow with and without confidence.

3.2 Inference with Confidence

This task demands strong causal reasoning skills and specialized knowledge in the legal domain, intutively beyond the capabilities of small models like BERT. LLMs are trained with vast Internet-based corpora, obtaining extensive knowledge and causal reasoning capabilities. Hence, we opted to utilize LLM, specifically the GPT-4, the model name is "gpt-4-0125-preview" and all hyperparameters use the default. Furthermore, in response to the ambiguity in LLM's responses, we proposed the Self-Eval mechanism, wherein the LLM is required to assess the confidence of its answer candidates while generating outputs. Our prompt took the following form: [*You are an AI assistant with reasoning and distinguishing abilities in the legal field. I have a new NLP task and dataset from the domain of the U.S. civil procedure. As a legal assistant, you can help me decide whether the relevant answer is correct or not. I will provide an explanation, an analysis, a question, and an answer. Please analyze to see if the answer is correct and give your confidence on a scale of 0-5, where the higher the score, the more accurate you think your answer is. The output format is: Analysis:, Is correct (Yes/No):, Confidence score:*].

Additionally, we found that LLM performs better in choice tasks than in judgment tasks. By examining task data examples, we noticed that some examples had the same introduction and question. Therefore, we converted data from true or false questions to multiple-choice questions. Specifically, we assigned numbers to answer candidates for LLM to choose from. If all options are incorrect, it returns None. For choice tasks, the prompt format we used was as follows: [*You are an AI assistant with reasoning and distinguishing abilities in the legal field. I have a new NLP task and dataset from the domain of the U.S. civil procedure. As a*

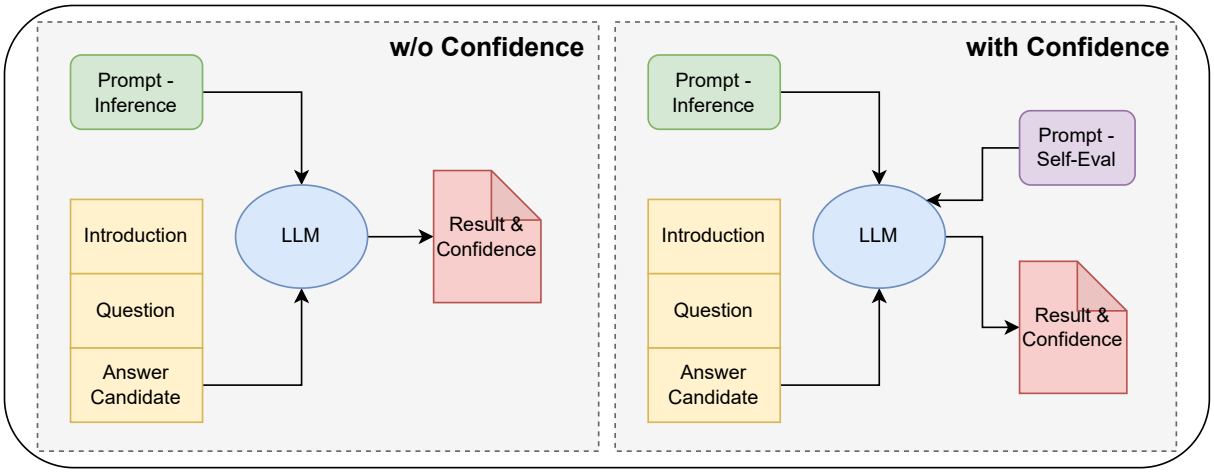

Figure 2: Flowchart of LLM System w/o Confidence and with Confidence

legal assistant, you can help me decide whether the relevant answer is correct or not. I will provide an explanation, an analysis, a question, and a few answers. Only one or none of these answers is correct. Please determine which answer is correct, Note that there may be cases where none of the answers are correct. Give a confidence score (0-5) on the larger model's answer, with higher scores indicating that you think the answer is correct. The output format is: correct answer: answer-id, confidence: score:]. Ultimately, to alleviate the model's stochastic nature, we implemented an ensemble strategy where, for each LLM, we ran it three times and aggregated the inference results, which is the final version we used in the evaluation. Due to cost constraints, we only implemented an ensemble strategy in the experimental group with the highest performance results.

3.3 Inference without Confidence

In addition to the previously mentioned LLM system equipped with the Self-Eval mechanism, we also conducted experiments involving direct inferring. The experimental parameters were kept consistent with the previous settings. When assessing judgement tasks, we utilized the following prompt format: [*You are an AI assistant with reasoning and distinguishing abilities in the legal field. I have a new NLP task and dataset from the domain of the U.S. civil procedure. As a legal assistant, you can help me decide whether the relevant answer is correct or not. I will provide an explanation, an analysis, a question, and an answer. Please analyze to see if the answer is correct. The output format is: Analysis:, Is correct (Yes/No):*]

When tackling choice tasks, the prompt we utilize is as follows: [*You are an AI assistant with reasoning and distinguishing abilities in the legal field. I have a new NLP task and dataset from the domain of the U.S. civil procedure. As a legal assistant, you can help me decide whether the relevant answer is correct or not. I will provide an explanation, an analysis, a question, and a few answers. Only one or none of these answers is correct. Please determine which answer is correct, Note that there may be cases where none of the answers are correct:*]

3.4 2Pass Strategy

In addition to the above experiments, we also designed a 2pass LLM reasoning and evaluation experiment to verify the self-evaluation ability of LLM. We take true or false questions as an example. First, we prompt the LLM to provide reasoning-only answers. Next, we ask the LLM to provide confidence scores for its answers, and gain the final result based on the confidence score. If confidence exceeds 3, we maintain the original result given by the LLM, otherwise we flip it. The prompts for the first pass are the same as the judgment only, and the prompts for the second pass are as follows: [*You are an AI assistant with reasoning and distinguishing abilities in the legal field. I have a new NLP task and dataset from the domain of the U.S. civil procedure. I will provide an explanation, an analysis, a question, and an answer. And analysis and judge of LLM, Give a confidence score (0-5) on the larger model's answer, with higher scores indicating that you think the answer is correct. The output format is: confidence: score:*].

Model	F1 Score	Accuracy
GPT-4-judgement only	0.7061	0.7551
GPT-4-judgement with confidence	0.7211	0.7653
GPT-4-2pass	0.6984	0.7341
GPT-4-choice only	0.7644	0.8163
GPT-4-choice with confidence	**0.8012**	**0.8649**
irene.benedetto's System	0.7747	0.8265
GPT-4-choice with confidence (Ensemble)	**0.8231**	**0.8673**

Table 1: Results of different models for test

4 Results and Analysis

4.1 Overview

Table 1 shows the results of different strategies on the test set of this task, where the representatives not marked with the Ensemble flag only run a single experiment. The evaluation metrics are F1 score and accuracy. As it is shown in the table, our final system, GPT-4-choice with confidence (Ensemble), has achieved the highest scores on both metrics, outperforming the best system from other participants, irene.benedetto, by absolute margins of 4.08 percentage points on F1 score and 2.65 percentage points on accuracy. Even without ensembling, our approach still improves F1 score by 2.65 percentage points and accuracy by 3.84 percentage points. This can prove the effectiveness of our system on the Legal Argument Reasoning task.

Table 1 presents a comparison between the performance of GPT-4 with and without Self-Eval. The results indicate a notable improvement when real-time confidence assessment is implemented. Specifically, for true or false questions, the F1 score of GPT-4 with confidence assessment is 1.50 percentage points higher than that without confidence assessment, with a corresponding 1.02 percentage point increase in accuracy. Similarly, in the case of multiple-choice questions, the F1 score for GPT-4 with confidence assessment outperforms the model without by 3.68 percentage points, while accuracy improves by 4.86 percentage points. This significant enhancement in performance underscores the value of integrating real-time confidence assessment into the model. The observed effect is attributed to the Self-Eval mechanism, which appears to stimulate the latent capabilities of the LLM. By prompting the LLM to evaluate its own confidence levels, it performs reasoning tasks more diligently, resulting in more accurate answers.

From Table 1, we can also see that on F1 score, 2Pass prompting underperforms the judgment-only and judgment-with-confidence versions of GPT-4 by 0.77 and 2.27 percentage points respectively. This suggests that ex-post confidence assessment does not benefit the model's answers. Confidence assessment only improve performance if conducted concurrently with inference. While surprising, this indicates promise for further exploration.

4.2 Case Study

As we mentioned in Section 3.1, the output of LLM is sometimes ambiguous. To address this, we introduced a self-evaluation mechanism. The following case illustrates its impact. As shown in figure 3, before the Self-Eval mechanism was introduced, when posed with a true of false question, the model replied "not entirely correct" — an ambiguous response falling between true and false. In contrast, analysis of multiple responses after incorporating the self-assessment mechanism revealed no ambiguous statements. These results demonstrate the effectiveness of the self-evaluation task at reducing ambiguity in LLM's responses.

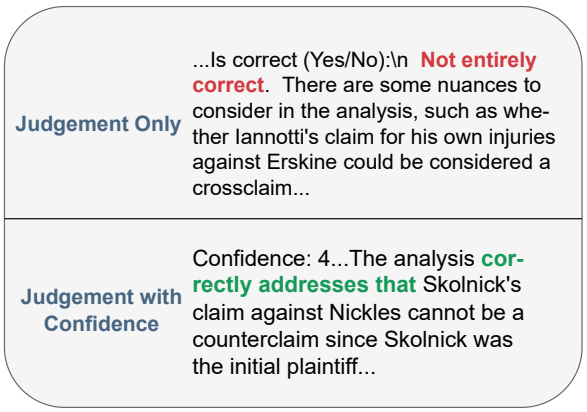

Figure 3: Case about Ambiguous with Self-Eval

5 Conclusion

In this article, we propose an LLM system with Self-Eval mechanism for SemEval-2024 Task 5. We explore the potential for using GPT-4 and prompt learning to obtain causal reasoning capabilities in the field of civil litigation. We have proven that the Self-Eval mechanism can alleviate the problem of unclear output and can also significantly improve performance. Additionally, we found that GPT4 demonstrates greater aptitude for choice tasks than for judgement tasks. With the prompts we provide, the experiment is fully reproducible and the experimental results can be extracted through regular expressions.

Due to time and space limitations, we leave some questions unresolved. For example, we only used GPT-4 for experiments. The broader applicability of the Self-Eval mechanism to other LLMs and its effectiveness in diverse tasks present room for further investigation. We intend to dive deeper into these questions in future work.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.