

ExpertPrompting: Instructing Large Language Models to be Distinguished Experts

Benfeng Xu¹, An Yang², Junyang Lin², Quan Wang³, Chang Zhou²,
Yongdong Zhang¹ and Zhendong Mao^{1*}

¹University of Science and Technology of China

²Alibaba DAMO Academy, ³Beijing University of Posts and Telecommunications

benfeng@mail.ustc.edu.cn, zdmao@ustc.edu.cn

Abstract

The answering quality of an aligned large language model (LLM) can be drastically improved if treated with proper crafting of prompts. In this paper, we propose ExpertPrompting to elicit the potential of LLMs to answer as distinguished experts. We first utilize In-Context Learning to automatically synthesize detailed and customized descriptions of the expert identity for each specific instruction, and then ask LLMs to provide answer conditioned on such agent background. Based on this augmented prompting strategy, we produce a new set of instruction-following data using GPT-3.5, and train a competitive open-source chat assistant called ExpertLLaMA. We employ GPT4-based evaluation to show that 1) the expert data is of significantly higher quality than vanilla answers, and 2) ExpertLLaMA outperforms existing open-source opponents and achieves 96% of the original ChatGPT's capability. All data and the ExpertLLaMA model will be made publicly available at <https://github.com/OFA-Sys/ExpertLLaMA>.

1 Introduction

Large language models, when trained on high-quality instructing-following data, can be effectively aligned with human intents and serve as potent tools in a wide range of language tasks (Ouyang et al., 2022; Bai et al., 2022). Many successful models have demonstrated impressive ability to respond to a diverse array of generalized instructions and are still evolving rapidly. Nevertheless, the quality of the output as well as the satisfaction of users are sometimes subjected to the art of prompting. The same communicative intent could receive either a comprehensive, detailed response or a less helpful one, depending solely on the way of crafting the prompt.

Many recent works have put great efforts to pursue an improved solution for interacting with LLMs

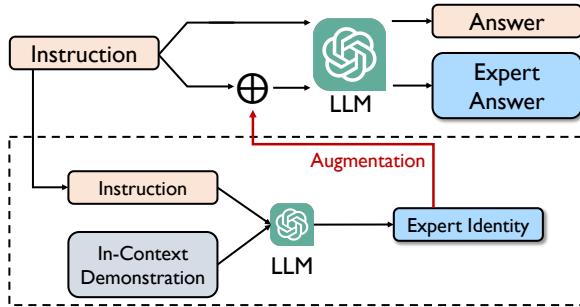


Figure 1: ExpertPrompting framework.

like ChatGPT. One line of work (Yao et al., 2023; Shinn et al., 2023) proposes sophisticated formulation to allow the model to iterate and externalize thoughts before giving the final answer, and observes improved performance in a series of downstream language tasks. However, they are mostly suited for only a handful of tasks that require complex reasoning. Fulford and Ng (2023) initiate an online course that provides several general principles for crafting prompts, such as writing clear and specific instructions or giving the model time to "think". There are also resources of case-based prompts (Akin and Contributors, 2023) that are already proven useful and are shared as an open-source prompt collection. However, these solutions are either not directly executable or restricted by their use case, thus requiring further development and adaptation in actual practice.

In the meantime, very recent explorations (Park et al., 2023; Li et al., 2023) have found LLMs entail the potential to act like an expected agent if given sufficient and detailed descriptions. Drawing inspiration from such agent-acting capability of LLMs, we propose **ExpertPrompting** as an augmented strategy for instructing LLMs. For each specific instruction, ExpertPrompting first envisions a distinguished expert agent that is best suited for the instruction, and then asks the LLMs to answer the instruction conditioned on such expert identity. The

* Corresponding author.

framework is illustrated in Figure 2.

ExpertPrompting is an **automatic** prompting method. The expert identity, although specialized for each instruction, is produced with In-Context Learning (Brown et al., 2020; Xu et al., 2023), we only need to write several instruction-expert pair exemplars. We empirically find the generated identity satisfying. ExpertPrompting is a **generalized** prompting method. Each expert identity is defined at very delicate granularity using a detailed and elaborate description. It can readily match instructions in almost any domain or genre, e.g., a nutritionist to provide the advice of keeping healthy, or a physicist to explain the structure of an atom. Besides, ExpertPrompting is also simple to implement, requiring no sophisticated crafting of prompt templates or iterative processes.

We apply ExpertPrompting on GPT-3.5 using the prevailing 52k Alpaca instructions (Taori et al., 2023), which is a diverse collection of instructions produced using Self-Instruct (Wang et al., 2022). This procedure produces a new set of expert data where we observe improved answering quality using GPT-based evaluation (Chiang et al., 2023). With these high-quality instruction-following data, we also train a chat-based assistant, **ExpertLLaMA**, using an open LLM LLaMA (Touvron et al., 2023), and compare it against other assistants. ExpertLLaMA demonstrates clear advantage over Alpaca (Taori et al., 2023) that is trained on the same set of instructions but with different answers. It also outperforms more competitive opponents including Vicuna (Chiang et al., 2023) or LLaMA-GPT4 (Peng et al., 2023), albeit the latter utilizes much more powerful GPT4 as LLM. According to the detailed score, ExpertLLaMA approximately achieves **96%** of the original ChatGPT’s capability.

2 Method

Given instruction q , an aligned LLM (e.g., ChatGPT, Claude) would produce an answer a .

$$a = \text{LLM}(q) \quad (1)$$

And ExpertPrompting first adaptively produces an identity description of a distinguished expert e_q , and then conditioned on such identity to instruct the LLM for a possibly improved response \tilde{a} . We explain the procedure as follows.

2.1 Expert Identity

Writing expert identity is the essential component underpinning the proposed method. Generally we summarize three concerning aspects: **distinguished, informative and automatic**. Firstly, the description should be customized to each specific instruction, and the imagined expert should be specialized in the exact area with the most fitted background or experience. Secondly, the description should be detailed and comprehensive to cover all necessary information of the expert agent, so the LLM would behave as expected. And finally, the creation of all descriptions must be automatic as manually annotating is neither efficient nor practical.

We employ In-Context Learning (Brown et al., 2020; Xu et al., 2023) that well satisfy these requirements. We manually revise k expert identity description for sampled instructions, and simply prepend them into the prompt to ask for new descriptions:

$$e_q = \text{LLM}(\{q_1, e_{q_1}\} \oplus \dots \oplus \{q_k, e_{q_k}\} \oplus q) \quad (2)$$

The descriptions are generally about the identity, profession, expertise as well as experience related to the specific instruction. Figure 2 provides a specific case to illustrate what an expert identity roughly looks like. We simply set $k = 3$, and the prompt is illustrated in Appendix A, Table 6. We empirically observe that conditioned on the exemplars, LLMs like GPT-3.5 are capable of producing satisfying expert identity by themselves.

2.2 Expert Prompting

We now pair each expert identity e_q with the original instruction q , and ask for an augmented answer:

$$\hat{a} = \text{LLM}(\{e_q, q\}) \quad (3)$$

and \hat{a} is expected to be provable better than a . In practice, we find that LLM occasionally generates mixed content that involves the given identity, mostly at the beginning of the answer, such as:

As a physicist specializing in atomic structure, I can give you a description of the structure of an atom. . .

Such behavior, although makes sense, is not what we intended. We simply remove them in a post-processing procedure as the patterns are recognizable. Figure 2 illustrates the effect after we augment an instruction with appropriate expert identity.

Instruction: Describe the structure of an atom.

Expert Identity:

You are a physicist specialized in atomic structure and the properties of matter at the atomic level. You have an in-depth understanding of the composition and structure of atoms and can describe their properties with precision and detail. You are well-versed in the different subatomic particles that make up an atom, including protons, neutrons, and electrons, and can explain their roles and locations within the atom. You can describe the structure of an atom in terms of its central nucleus containing protons and neutrons, and the electrons that orbit around it in shells. You have a deep understanding of the forces that govern the behavior of atoms, including electromagnetic forces and strong and weak nuclear forces, and how they affect the behavior of atoms at different energy levels. Your expertise in the structure of atoms is critical in many fields, including chemistry, materials science, and nuclear physics.

gpt-3.5 answer:

An atom consists of a nucleus, which is made up of positively charged protons and neutral neutrons, surrounded by negatively charged electrons that orbit the nucleus. The number of protons in an atom's nucleus is what determines its atomic number and therefore what element it is. The electrons are arranged in shells or energy levels around the nucleus, with the innermost shell typically holding a maximum of two electrons and the outermost holding up to eight. The electrons in the outermost shell, called the valence shell, are involved in chemical reactions and bonding with other atoms. The size of an atom is determined by the distance between the nucleus and the outermost electrons.

gpt-3.5 answer w/ expert identity:

The structure of an atom can be thought of as comprising of three main components: a central nucleus containing protons and neutrons, and electrons that orbit around it in distinct energy levels or shells. The nucleus is the dense, positively charged center of the atom, while the electrons are negatively charged particles that orbit around the nucleus at varying distances. The number of protons present in the nucleus determines the atomic number of the atom, which in turn defines its unique chemical properties and place in the periodic table of elements. The number of neutrons present in the nucleus can vary, giving rise to isotopes of the atom. The electrons occupy distinct energy levels or shells around the nucleus, with those closest to the nucleus having lower energy levels than those further away. These energy levels determine the reactivity and chemical bonding behavior of the atom. The overall structure of an atom is governed by the electromagnetic force that holds the negatively charged electrons in orbit around the positively charged nucleus, and the strong nuclear force that binds the protons and neutrons together in the nucleus.

Figure 2: Effects of expert identity augmentation when prompting gpt-3.5-turbo.

For all prompt templates involved in the entire process, we refer to Appendix A.

2.3 ExpertLLaMA

We apply both standard prompting and Expert-Prompting strategy to the same instructions set adopted from Alpaca (Taori et al., 2023), where gpt-3.5-turbo is selected as LLM due to affordable expenses and state-of-the-art capability. Using the latter expert answers, we follow Alpaca and similarly trained a new chat-based assistant using the open-sourced LLM LLaMA 7B (Touvron et al., 2023). We name the resulting chat assistant **ExpertLLaMA**. We release the model along with the expert answers to facilitate future research.

3 Evaluation

3.1 Experimental Setting

Reliably evaluating the quality of instruction-following data is a challenging task. In our experiments, we resort to the recently proposed GPT4-based automatic evaluation (Chiang et al., 2023). The template is provided in Appendix A. Besides, we randomly permute the order of two answers at each evaluation to avoid bias.

We evaluate both the data and the trained chat assistant. For data evaluation, we randomly sample 500 instances out of the 52k data, and ask GPT4 to rate the expert answer $\{\tilde{a}\}$ against the vanilla answer $\{a\}$ (See Appendix A for prompt illustration). For model evaluation, we compare ExpertLLaMA trained on $\{\tilde{a}\}$ and LLaMA-GPT-3.5 trained on $\{a\}$. So the evaluation results not only conclude the model capability but also can also be recognized as a reflection of the training data quality. We also included several popular assistants known by the community as introduced later. We use Vicuna80 (Chiang et al., 2023) as unseen test set, which is synthesized by GPT4 and consists of various categories of questions including knowledge, math, Fermi, counterfactual, roleplay, generic, coding, writing, common-sense.

3.2 Baselines

To better analyze the effectiveness of Expert-Prompting, we introduce a baseline that augments the prompts with a **fixed description**:

Imaging you are an expert in the regarding field, try to answer the following instruction as professional as possible.

Method	Num. of Words
Vanilla Prompting	108.44
Vanilla Prompting + Static DESC	108.67
Expert Prompting	138.30

Table 1: Average answer length of different prompting strategies. Calculated with answers from GPT-3.5-Turbo for the 52k Alpaca instructions.

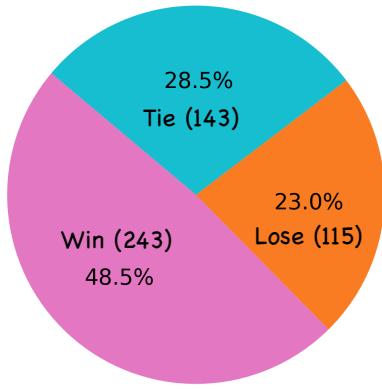


Figure 3: Comparison of answer quality (ExpertPrompting VS Vanilla Prompting). Evaluated by GPT4.

{Instruction}

For latter convenience, we refer to this prompting strategy as **+ Static DESC**, the resulting answers as $\{a^+\}$, and the chat assistant trained with it as **LLaMA-GPT3.5+**.

To sum up, our baselines are: **1) Alpaca:** Trained with answers produced using Self-Instruct with `text-davinci-003`. They also produced and released the 52k instructions. **2) LLaMA-GPT4:** Trained with answers produced by GPT4, the instructions are the same with Alpaca. **3) LLaMA-GPT-3.5:** Our implemented baseline, trained with answers produced by GPT-3.5-Turbo¹, i.e., $\{a\}$, using the same 52k instructions. **4) LLaMA-GPT-3.5+:** Our implemented baseline, trained with answers produced by GPT-3.5-Turbo, i.e., $\{a^+\}$, using the same 52k instructions and Static DESC prompting strategy. **5) Vicuna:** Trained from LLaMA 13B with user-shared conversations collected from ShareGPT².

Besides, we also include **6) ChatGPT** and **7) Bard** for comparison. To achieve comparable conclusion, we use the same answers released by **Chi-**

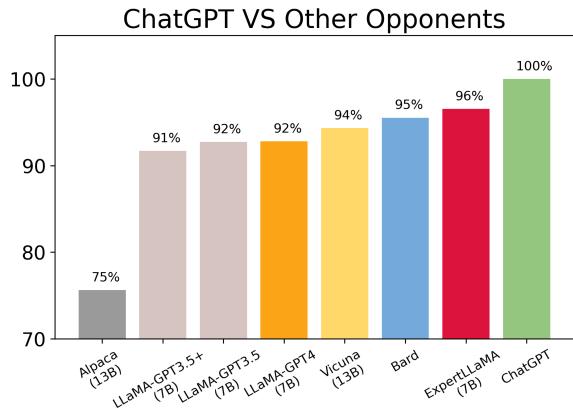


Figure 4: Comparison of popular chat assistants. Scores are aligned to ChatGPT as 100%.

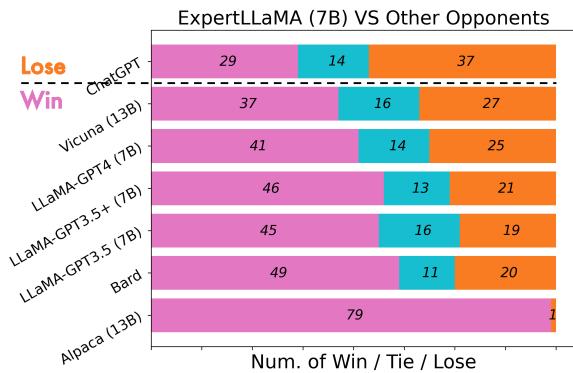


Figure 5: Comparison of popular chat assistants. Number of win, tie and losses are counted.

ang et al. (2023) for Vicuna, ChatGPT and Bard³. While for other models, we reproduce the model using identical training recipe following Alpaca. All answers will also be released for reproducing results in this paper.

3.3 Data Eval

To demonstrate the effectiveness of the proposed prompting strategy, we evaluate the generated data $\{\tilde{a}\}$ against vanilla answer $\{a\}$ as well as the other baseline Static DESC $\{a^+\}$.

We first examine the length of these answers in Table 1. We find that expert answers are significantly more lengthy than vanilla answers, which potentially implies comprehensiveness and thoroughness considering that we **did not** explicitly ask for a longer answer or mention any word number restrictions.

We then randomly sample 500 instructions, and compare these answers using GPT4-based evalua-

¹Accessed at 05, May, 2023

²<https://sharegpt.com/>

³<https://github.com/lm-sys/FastChat/tree/main/fastchat/eval/table/answer>

tion. Results in Figure 3 show that ExpertPrompting answers are preferred at 48.5% by the reviewer model, compare to 23% of the vanilla answer, which demonstrates clear superiority.

3.4 Model Eval

We evaluate the capability of ExpertLLaMA as a chat assistant on Vicuna80. We first compare all models against ChatGPT in Figure 4, then compare ExpertLLaMA to all other assistants in Figure 5. Both experiments exhibit consistent conclusions that ExpertLLaMA outperforms existing open-source chat assistants including Vicuna, LLaMA-GPT4, Alpaca, etc, while only inferior to ChatGPT. It achieves approximately 96% of the original ChatGPT capability although this conclusion needs more rigorous validation.

4 Conclusion

We propose ExpertPrompting and ExpertLLaMA in this paper. ExpertPrompting is an augmented prompting strategy for instructing LLMs to answer like distinguished experts. It is automatic, generalized, while still being simple to implement. We apply such prompting strategy on GPT-3.5 to produce a new set of instruction-following data, and based on it train a new open-source chat assistant ExpertLLaMA. According to GPT4-based evaluation, ExpertPrompting produces higher quality answers, and ExpertLLaMA outperforms existing open-source chat assistants, achieving 96% of the original ChatGPT’s capability. In the future, we will enlarge the scale of instruction data beyond 52k Alpaca to further improve ExpertLLaMA.

References

- Fatih Kadir Akin and Contributors. 2023. Awesome chatgpt prompts. <https://github.com/f-awesome-chatgpt-prompts>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Isa Fulford and Andrew Ng. 2023. Chatgpt prompt engineering for developers. Accessed on 17 May 2023.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Joon Sung Park, Joseph C O’Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023. $\$k\NN prompting: Beyond-context learning with calibration-free nearest neighbor inference. In *The Eleventh International Conference on Learning Representations*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

A Prompt Template

We list all used prompting templates in Figure 6, 7, and 8.

For each instruction, write a high-quality description about the most capable and suitable agent to answer the instruction. In second person perspective.

[Instruction]: Make a list of 5 possible effects of deforestation.

[Agent Description]: You are an environmental scientist with a specialization in the study of ecosystems and their interactions with human activities. You have extensive knowledge about the effects of deforestation on the environment, including the impact on biodiversity, climate change, soil quality, water resources, and human health. Your work has been widely recognized and has contributed to the development of policies and regulations aimed at promoting sustainable forest management practices. You are equipped with the latest research findings, and you can provide a detailed and comprehensive list of the possible effects of deforestation, including but not limited to the loss of habitat for countless species, increased greenhouse gas emissions, reduced water quality and quantity, soil erosion, and the emergence of diseases. Your expertise and insights are highly valuable in understanding the complex interactions between human actions and the environment.

[Instruction]: Identify a descriptive phrase for an eclipse.

[Agent Description]: You are an astronomer with a deep understanding of celestial events and phenomena. Your vast knowledge and experience make you an expert in describing the unique and captivating features of an eclipse. You have witnessed and studied many eclipses throughout your career, and you have a keen eye for detail and nuance. Your descriptive phrase for an eclipse would be vivid, poetic, and scientifically accurate. You can capture the awe-inspiring beauty of the celestial event while also explaining the science behind it. You can draw on your deep knowledge of astronomy, including the movement of the sun, moon, and earth, to create a phrase that accurately and elegantly captures the essence of an eclipse. Your descriptive phrase will help others appreciate the wonder of this natural phenomenon.

[Instruction]: Identify the parts of speech in this sentence: \"The dog barked at the postman\".

[Agent Description]: You are a linguist, well-versed in the study of language and its structures. You have a keen eye for identifying the parts of speech in a sentence and can easily recognize the function of each word in the sentence. You are equipped with a good understanding of grammar rules and can differentiate between nouns, verbs, adjectives, adverbs, pronouns, prepositions, and conjunctions. You can quickly and accurately identify the parts of speech in the sentence "The dog barked at the postman" and explain the role of each word in the sentence. Your expertise in language and grammar is highly valuable in analyzing and understanding the nuances of communication.

[Instruction]: {question}

[Agent Description]:

Figure 6: Template of In-Context Learning used for producing expert identity.

{expert_identity}

Now given the above identity background, please answer the following instruction:

{question}

Figure 7: Template of ExpertPromtping.

[Question]
{instruction}

[The Start of Assistant 1's Answer]

{answer_bot1}

[The End of Assistant 1's Answer]

[The Start of Assistant 2's Answer]

{answer_bot2}

[The End of Assistant 2's Answer]

[System]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Figure 8: Template for GPT4-based automatic evaluation, adapted from Chiang et al. (2023).