Are LLMs Effective Backbones for Fine-tuning? An Experimental Investigation of Supervised LLMs on Chinese Short Text Matching

Anonymous ACL submission

Abstract

The recent success of Large Language Models (LLMs) has garnered significant attention in both academia and industry. Prior research on LLMs has primarily focused on enhancing or leveraging their generalization capabilities in zero- and few-shot settings. However, there has been limited investigation into effectively fine-tuning LLMs for a specific natural language understanding task in supervised settings. In this study, we conduct an experimental analysis by fine-tuning LLMs for the task of Chinese short text matching. We explore various factors that influence performance when fine-tuning LLMs, including task modeling methods, prompt formats, and output formats.

1 Introduction

The recent success of Large Language Models (LLMs), such as GPT-3(Brown et al., 2020), LLaMA(Touvron et al., 2023) and PaLM(Chowdhery et al., 2023), has garnered significant attention in both academia and industry. LLMs have demonstrated remarkable generalization capabilities in zero- and few-shot settings, particularly in natural language generation (NLG) tasks. Substantial efforts have been made to enhance and utilizing such generalization capabilities(Xu et al., 2023; Saad-Falcon et al., 2023; Yun et al., 2023).

However, for natural language understanding (NLU) tasks, zero- and few-shot LLMs struggle to achieve satisfactory performance(Nie et al., 2022; Wei et al., 2023; Li et al., 2023a,b) compared to fine-tuned small models (e.g., Bert base(Devlin et al., 2018)). Our experimental results on the task of Chinese short text matching also confirm this phenomenon. As presented in Section3.1, fine-tuned Bert achieves an accuracy of 84.5% on the BQ(Chen et al., 2018) corpus, while GPT-4[1], one of the most successful LLMs, only attains an accuracy score of 52.9% in zero-shot and 77.9% in few-shot settings. There has been limited investigation into effectively tuning LLMs for a specific NLU task in supervised settings. In this paper, we explore various factors affecting the performance of LLMs for Chinese short text matching task, including task modeling methods, prompt formats, and output formats.

- **Task modeling methods:** In this study, we examine the impacts of modeling this task as both a generative task and a discriminative classification task, respectively. *(1) Generative Task:* LLMs uniformly model all tasks as generative tasks. Following this principle, we organize the given pair of sentences into a single text as input and make the model generate the target label (equivalent or inequivalent). *(2) Discriminative Classification Task:* Motivated by the efficacy of fine-tuning Bert for text matching(Chen et al., 2020; Qi et al., 2022), we concatenate the given pair of texts as input, extract vector representations from the final LLM layer as features, and perform binary classifications based on the extracted features.

- **Prompt Formats:** Prompt design is crucial for LLMs in zero- and few-shot settings(Gu et al., 2021; Liu et al., 2023). However, the importance of prompts in supervised settings has not been explored. In this paper, we compare two completely different styles of prompts. One is concise, directly concatenating the given pair of sentences without any explanation of the target task. The other organizes the prompt through complex instructions, including not only the given sentences but also a detail description of the target task.

- **Output Formats:** Incorporating the Chain of

[1]The metrics are measured by utilizing OpenAI API.

Thought (CoT) into prompts has been shown to significantly enhance performance in reasoning and complex tasks in zero- and few-shot settings(Wei et al., 2022; Wang et al., 2022). Nevertheless, the impact of CoT on matching tasks in supervised settings has yet to be examined. In this study, we address this gap by *incorporating CoT into the output part* of training samples.

We conduct experiments on two widely-used Chinese short text matching datasets, LCQMC (Liu et al., 2018a) and BQ (Chen et al., 2018). All experiments are carried out based on CLLM-7B, which is a Chinese-enhanced model based on LLaMA-2-7B. Our preliminary results demonstrate that the fine-tuned CLLM-7B outperforms both fine-tuned BERT and few-shot GPT-4. Furthermore, the results indicate that the generative paradigm surpasses the discriminative approach, especially when training data is limited. Lastly, our experiments reveal that CoT is also beneficial for the matching task in supervised settings.

In summary, our major contributions are twofold: (1) To the best of our knowledge, we are the first to systematically explore effective strategies for fine-tuning LLMs for text matching. (2) We are the first to verify the effectiveness of CoT for NLU task. Although our experiments focused on the text matching task, we believe that our findings may also be applicable to other NLU tasks, such as text classification.

2 Backgrounds

In this section, we provide a brief overview of the Chinese short text matching task and the datasets employed in this study.

2.1 Task Definition

Chinese short text matching, often regarded as a task of identifying sentence semantic equivalence, is a fundamental task of natural language processing. Given a pair of sentences, the goal of a matching model is to ascertain their semantic equivalence. Short text matching is extensively utilized in a range of NLP tasks, such as question answering (Liu et al., 2018b) and dialogue systems (Pang et al., 2008).

2.2 Datasets and Metrics

We conduct experiments on two widely-used Chinese short text matching corpora: LCQMC (Liu et al., 2018a) and BQ (Chen et al., 2018).

LCQMC is a large-scale, open-domain question matching corpus. It comprises 260,068 Chinese search query pairs, including 238,766 training samples, 8,802 development samples, and 12,500 test samples. Each pair is annotated with a binary label indicating whether the two queries share the same intention.

BQ is a domain-specific, large-scale corpus for bank question matching. It consists of 120,000 Chinese sentence pairs, including 100,000 training samples, 10,000 development samples, and 10,000 test samples. Each pair is also annotated with a binary label indicating whether the two sentences convey the same meaning.

We employ accuracy (ACC.) as the evaluation metric, which is the percentage of correctly predicted examples.

3 Experiments and Results

In this section, we outline the experimental configurations and present the results. We examine the influence of the three factors discussed in Section 1 through the following experiments. We tune models via full-model fine-tuning.

3.1 Generative vs. Discriminative Models

We first outline our approach to fine-tuning LLMs by modeling the matching task as both a generative task and a discriminative task. Subsequently, we present the results and provide an analysis.

Modeling as A Generative Task: LLMs consistently treat all tasks as generative tasks. In line with this principle, we merge the provided pair of sentences with instructions into a single text input and prompt the model to generate the target label. We refer to this model as CLLM-7B-GEN. Figure 1(b) illustrates the model structure. We optimize it by maximizing the generation probability of the target label.

Modeling as A Discriminative Task: Inspired by the effectiveness of fine-tuning BERT for text matching tasks (see Figure 1(a)), we concatenate the given pair of texts as input, extract vector representations from the final LLM layer as features, and perform binary classification based on the extracted features. We refer to this model as CLLM-7B-CLS. Figure 1(c) demonstrates the model structure.

We validated the performance of generative and discriminative models on training sets of different scales. Figure 2 shows the experimental results,

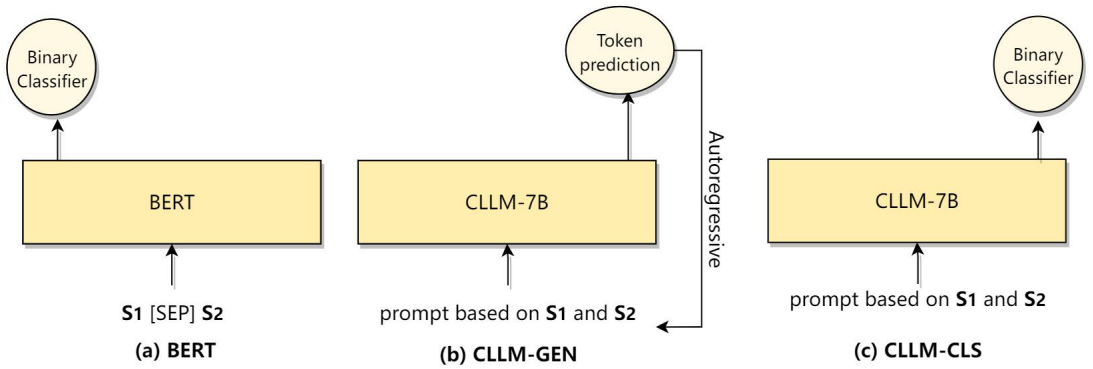
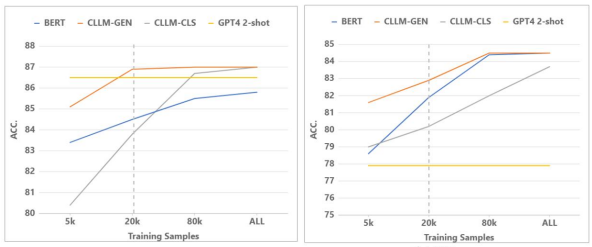
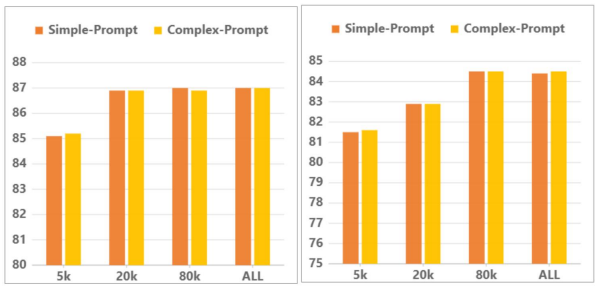


Figure 1: Model structures of modeling text matching as generative and discriminant task.



(a) LCQMC (b) BQ

Figure 2: The results of models trained on 5,000, 20,000, 80,000 samples as well as trained on the entire training set.



(a) LCQMC (b) BQ

Figure 3: The results of concise and complex prompts.

where the 2-shot GPT-4 results are measured by calling the official OpenAI API. Figure 6 and Figure 7 in Appendix A illustrate the 2-shot prompts for LCQMC and BQ, respectively. From the results, we observe that:

1) When the number of training samples is less than 20,000, CLLM-GEN significantly outperforms discriminative models, including BERT and CLLM-CLS, on both LCQMC and BQ. This phenomenon is quite intuitive, as the generative approach aligns with the pre-training procedure, making it easier to activate the knowledge acquired by the model during pre-training. Furthermore, due to the massive amount of data used in the pre-training phase of LLMs, the issue of evaluation data leakage cannot be ignored (Yang et al., 2023; Zhou et al., 2023). To determine whether CLLM-7B has a data leakage problem, we conducted zero-shot experiments on it. The model achieves an accuracy of 52.1% on LCQMC and 52.9% on BQ, slightly better than the 50% expected from random guessing. Consequently, we believe that both BQ and LCQMC are not included in CLLM-7B's pre-training data.

2) The performance of 2-shot GPT-4 on BQ is much worse than that of supervised models. This is mainly because BQ is a dataset of real customer service questions from WeBank Inc., and a full understanding of the sentences' meaning requires background information about this bank. For example, questions in BQ usually mention specific products or a particular function in the bank's app. This background knowledge is unknown to CLLM and is also impossible to provide entirely in the prompt.

3) CLLM-GEN trained on the whole training corpus on LCQMC outperforms BERT. However, it fails on the BQ corpus. We believe the reason is that CLLM-7B, like BERT, also lack knowledge of WeBank, and such knowledge can only be obtained from the training data. Therefore, compared to BERT, CLLM-7B does not have an advantage on this dataset.

The above experiments demonstrate that generative paradigm is better for supervised LLMs. Therefore, all subsequent experiments will be conducted following this paradigm.

3.2 Concise vs. Complex Prompts

Prompt design is crucial for LLMs in zero- and few-shot settings. However, the significance of prompts in supervised settings remains unexplored. In this subsection, we compare two distinct styles of prompts. The concise prompt involves directly concatenating the given text pairs without any ex-

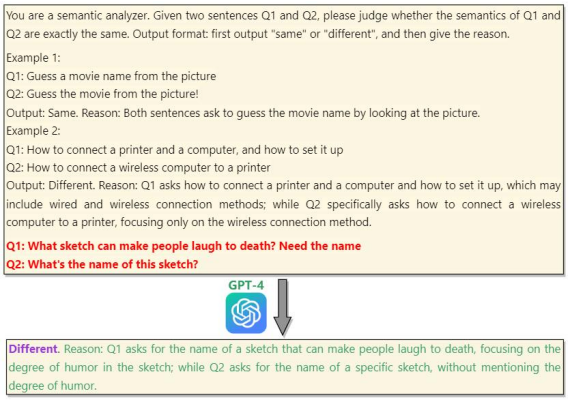


Figure 4: Illustration of how to obtain CoT via GPT-4. *All original texts in this figure are in Chinese. For ease of reading, we translated them. The original version is illustrated in Figure 9 in Appendix.*

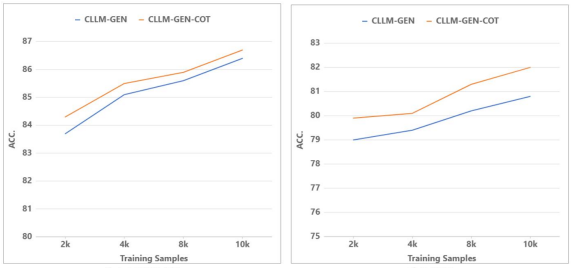
planation of the target task, while the complex prompt organizes the prompt with detailed instructions, incorporating not only the given texts but also a specific description of the target task. Examples of these prompts can be found in Figure 8 in Appendix A.

Figure 3 presents the results, showing that models separately trained by concise and complex prompts achieve comparable performance. This observation suggests that supervised LLMs are not sensitive to prompts. The primary function of a complex prompt is to enhance the model's comprehension of the target task. In supervised scenarios, the model can learn the task definition more accurately from the training data, rendering the prompt design less impactful.

3.3 Effects of CoT

CoT has demonstrated its effectiveness in reasoning and complex tasks within zero- and few-shot settings. However, its efficacy for language understanding tasks in supervised settings remains unexplored.

Matching datasets provide labels without CoT. To obtain CoT for the training set, we enlist GPT-4 to determine whether a given pair of texts is equivalent, while also providing explanations for its decision. For samples where GPT-4's judgment aligns with the golden label, we utilize the explanation as the CoT. Conversely, for inconsistent samples, we retain only golden label. Figure 4 depicts the designed prompt and response generated by GPT-4. Note that only the output portion of the training samples requires the addition of CoT. Figure 10 in Appendix presents a training sample that includes CoT. During the evaluation process, we disregard



(a) LCQMC (b) BQ

Figure 5: Results of models trained with CoT.

the CoT generated by the model, focusing solely on the label "same" or "different".

In order to reduce the cost, we did not obtain CoT for the entire training set. Instead, we separately sampled 10,000 instances from each dataset and requested GPT-4 to generate CoT. After filtering samples with inconsistent judgments, approximately 86% of samples in LCQMC and 78% in BQ retained CoT. We will release the data with CoT for further use by the community[2].

We conducted experiments on training sets of varying scales. Figure 5 displays the results, from which we observe that CoT improves performance on both LCQMC and BQ. Furthermore, the BQ dataset is more challenging than LCQMC, and CLLM-GEN-CoT achieved a more substantial improvement on BQ. This finding suggests that CoT may be particularly effective for difficult tasks.

4 Conclusions

In this work, we conduct an experimental study by fine-tuning LLMs on the task of Chinese short text matching. We investigate various factors affecting performance in tuning LLMs, including task modeling methods, prompt formats, and the chain of thought. We systematically carry out experiments on two widely used datasets. The results reveal several insights. First, the fine-tuned CLLM-7B outperforms both fine-tuned BERT and few-shot GPT-4, indicating that LLMs serve as effective backbones in supervised scenarios. Moreover, the generative paradigm is superior to the discriminative one, particularly when training data is limited. Second, supervised LLMs are insensitive to prompts, unlike zero- and few-shot LLMs. Third, CoT is also beneficial for supervised text matching. Although our experiments focus on the task of text matching, the observations may be applicable to other NLU tasks, such as text classification.

[2] https://github.com/xxx/xxx

Limitations

This study has two primary limitations: (1) Prompt engineering is crucial for zero- and few-shot LLMs. We assessed the few-shot performance of GPT-4, as depicted in Figure 2. Despite our meticulous design of the few-shot prompts, the prompt designs remain subjective and may not necessarily represent the most optimal choices. (2) This study concentrates on the text matching task. Additional experiments might be required to adequately demonstrate if the conclusions drawn in this article are applicable to other NLU tasks (e.g. text classification).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4946–4951.

Lu Chen, Yanbin Zhao, Boer Lyu, Lesheng Jin, Zhi Chen, Su Zhu, and Kai Yu. 2020. Neural graph matching networks for chinese short text matching. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 6152–6158.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023b. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018a. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th international conference on computational linguistics*, pages 1952–1962.

Yang Liu, Wenge Rong, and Zhang Xiong. 2018b. Improved text matching by enhancing mutual information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *arXiv preprint arXiv:2212.02216*.

Bo Pang, Lillian Lee, et al. 2008. Foundations and trends® in information retrieval. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135.

Le Qi, Yu Zhang, Qingyu Yin, Guidong Zheng, Wen Junjie, Jinlong Li, and Ting Liu. 2022. All information is valuable: Question matching over full information transmission network. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1431–1440.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. 2023. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11265–11279, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Silei Xu, Shicheng Liu, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina Semnani, and Monica Lam. 2023. Fine-tuned LLMs know more, hallucinate less with few-shot sequence-to-sequence semantic parsing over Wikidata. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5778–5791, Singapore. Association for Computational Linguistics.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

Hye Yun, Iain Marshall, Thomas Trikalinos, and Byron Wallace. 2023. Appraising the potential uses and harms of LLMs for medical systematic reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10122–10139, Singapore. Association for Computational Linguistics.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Appendix

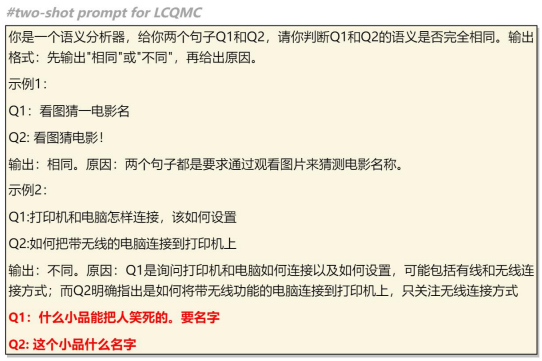


Figure 6: An illustration of 2-shot prompt for LCQMC.

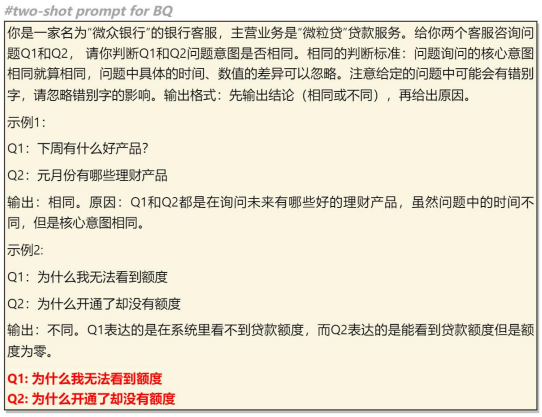


Figure 7: An illustration of 2-shot prompt for BQ.

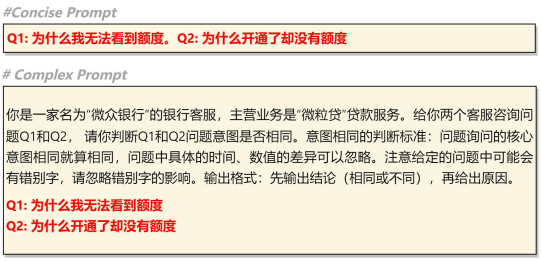


Figure 8: Examples of complex and simple prompts in Section3.2

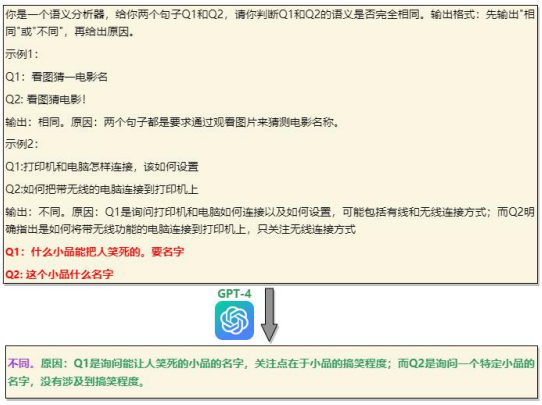


Figure 9: The Chinese version of texts in Figure 4

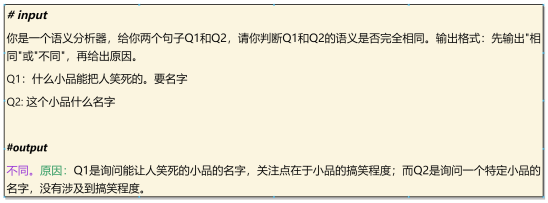


Figure 10: An example of training sample with CoT.