Better Zero-Shot Reasoning with Role-Play Prompting

Aobo Kong[1], **Shiwan Zhao**[2], **Hao Chen**[3], **Qicheng Li**[1*], **Yong Qin**[1], **Ruiqi Sun**[3], **Xin Zhou**[3]

[1]Nankai University, [2]Independent Researcher,
[3]Enterprise & Cloud Research Lab, Lenovo Research
kongaobo9@163.com, zhaosw@gmail.com,
{liqicheng, qinyong}@nankai.edu.cn, {chenhao31, sunrq2, zhouxin16}@lenovo.com

Abstract

Modern large language models (LLMs), such as ChatGPT, exhibit a remarkable capacity for role-playing, enabling them to embody not only human characters but also non-human entities like a Linux terminal. This versatility allows them to simulate complex human-like interactions and behaviors within various contexts, as well as to emulate specific objects or systems. While these capabilities have enhanced user engagement and introduced novel modes of interaction, the influence of role-playing on LLMs' reasoning abilities remains underexplored. In this study, we introduce a strategically designed role-play prompting methodology and assess its performance under the zero-shot setting across twelve diverse reasoning benchmarks, encompassing arithmetic, commonsense reasoning, symbolic reasoning, and more. Leveraging models such as ChatGPT and Llama 2, our empirical results illustrate that role-play prompting consistently surpasses the standard zero-shot approach across most datasets. Notably, accuracy on AQuA rises from 53.5% to 63.8%, and on Last Letter from 23.8% to 84.2%. Beyond enhancing contextual understanding, we posit that role-play prompting serves as an implicit Chain-of-Thought (CoT) trigger, thereby improving the quality of reasoning. By comparing our approach with the Zero-Shot-CoT technique, which prompts the model to "think step by step", we further demonstrate that role-play prompting can generate a more effective CoT. This highlights its potential to augment the reasoning capabilities of LLMs. We release our code at https://github.com/HLT-NLP/Role-Play-Prompting.

Introduction

Recent years have witnessed a paradigm shift in natural language processing, largely driven by large language models (LLMs) such as GPT-3 (Brown et al. 2020), PaLM (Chowdhery et al. 2022), and Llama (Touvron et al. 2023a). By pretraining on vast textual corpora, these models have attained an impressive capacity for language understanding and generation, empowering them to address a variety of downstream tasks through prompting, thus bypassing the necessity for task-specific fine-tuning. Amidst the surge of prompt techniques, role-playing (Wu et al. 2023) and chain-of-thought (CoT) prompting (Wei et al. 2022; Kojima et al. 2022) have garnered particular interest.

(a) Zero-Shot

Input

User: Xavier was 4 feet tall and grew 3 inches. Cole was 50 inches tall and grew 2 inches over the summer. What is the difference between Cole and Xavier's height now?

Output

Assistant: Xavier is now 4 feet + 3 inches = 7 inches tall. Cole is now 50 inches + 2 inches = 52 inches tall. The difference between their heights is 52 inches - 7 inches = 45 inches. ✗

(b) Role-Play Prompting

Input

User: From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students.

Assistant: That's great to hear! As your math teacher, I'll do my best to explain mathematical concepts correctly so that you can understand them easily. Feel free to ask any math problems or questions you have, and I'll be glad to assist you. Let's dive into the world of mathematics and explore its wonders together!

User: Xavier was 4 feet tall and grew 3 inches. Cole was 50 inches tall and grew 2 inches over the summer. What is the difference between Cole and Xavier's height now?

Output

Assistant: … Since 1 foot is equal to 12 inches, Xavier's initial height: 4 feet * 12 inches/foot = 48 inches. After growing 3 inches, Xavier's new height is: 48 inches + 3 inches = 51 inches. … Cole's new height after growing 2 inches: 50 inches + 2 inches = 52 inches. … We subtract Xavier's height from Cole's height: 52 inches - 51 inches = 1 inch. Therefore, the difference between Cole and Xavier's height now is 1 inch. ✓

Figure 1: Examples of ChatGPT with (a) zero-shot and (b) role-play prompting. The role-play prompts are highlighted.

Modern LLMs, with their advanced role-playing capabilities, have significantly enriched user experiences and forged new modes of interaction. They can convincingly mimic various personas, ranging from fictional characters to historical and contemporary figures. The assigned role provides context about the LLM's identity and background. By adopt-

[*]Qicheng Li is the corresponding author.
Preprint

ing the persona, the LLM can generate more natural, in-character responses tailored to that role. Recognizing this potential, companies like Character.AI[1] have developed dialogue agents portraying diverse figures. Beyond conversational applications, role-playing also boosts LLM performance on certain NLP tasks. For instance, when cast as a judge with a distinctive role, LLMs can effectively evaluate the quality of text summarization (Wu et al. 2023). More unconventionally, ChatGPT demonstrates competency in processing Linux commands when prompted as a Linux terminal[2]. Despite these advancements, analyzing the influence of role-playing on core LLM reasoning abilities warrants further investigation.

While the role-playing abilities of LLMs have expanded the horizon of human-computer interaction, the push to amplify the reasoning prowess of these models has led to the development of techniques like Chain-of-Thought (CoT) Prompting. CoT prompting was proposed by Wei et al. (2022) and involves providing reasoning steps in few-shot examples. By stimulating step-by-step reasoning, CoT prompting has markedly improved LLM reasoning abilities. Numerous subsequent studies (Wang et al. 2022; Kojima et al. 2022; Zhou et al. 2022) have built upon this approach. Inspired by the success of role-playing on many downstream tasks, we explore whether role-playing can similarly boost LLM reasoning performance. For example, could assigning ChatGPT the role of a math teacher enhance its ability to solve math problems? In this work, we introduce a zero-shot role-play prompting methodology based on a two-stage framework. During the first stage, we utilize the LLM to construct task-specific role-play prompts. In the second stage, responses are elicited for each reasoning query, guided by the previously constructed task-specific role-play prompts. An illustrative example is provided in Figure 1. We focus our study on conversational LLMs, evaluating our approach on 12 reasoning benchmarks using ChatGPT. Our results demonstrate consistent improvements over the zero-shot baseline on the majority of datasets, confirming the efficacy of role-play prompting. We further assess other conversational LLMs like Vicuna (Chiang et al. 2023) and Llama 2 (Touvron et al. 2023b), observing comparable gains.

Furthermore, we compare our method to the Zero-Shot-CoT technique (Kojima et al. 2022), which explicitly triggers CoT by appending *"Let's think step by step"* to questions. Modern conversational LLMs like ChatGPT have undergone extensive supervised fine-tuning, enabling them to generate CoT for certain topics without the need for an explicit trigger. In tasks where the model struggles to generate CoT spontaneously, such as Last Letter, both our approach and Zero-Shot-CoT can stimulate CoT from scratch. However, for tasks where CoT already occurs, such as arithmetic, both our approach and Zero-Shot-CoT reinforce the step-by-step reasoning process, but Zero-Shot-CoT demonstrates no significant effect, whereas our approach leads to better performance. Hence, we posit that role-play prompting is an

[1] https://beta.character.ai/

[2] https://www.engraved.blog/building-a-virtual-machine-inside/

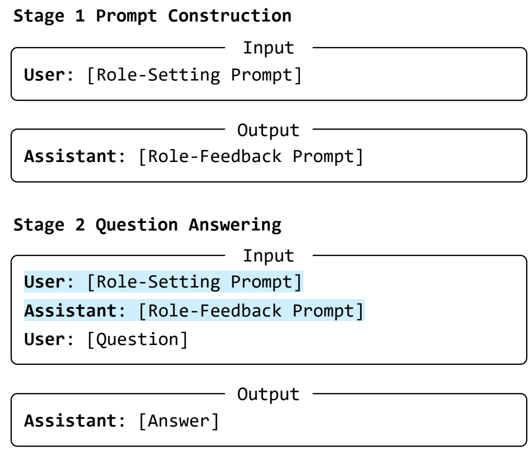


Figure 2: The two-stage framework of our proposed role-play prompting. The role-play prompts are highlighted.

implicit CoT trigger and can generate a more effective CoT in some fields compared with Zero-Shot-CoT.

To the best of our knowledge, this work represents the first systematic investigation of role-play prompting for reasoning tasks. Despite the transformative effects of role-playing on LLM behavior, sparse academic research has explored this phenomenon. We believe our study serves as an inaugural step to catalyze more extensive exploration into this promising research direction.

Our main contributions are summarized as follows:

- We propose a novel role-play prompting methodology based on a two-stage framework to enhance the zero-shot reasoning capabilities of LLMs. To our knowledge, we are the first to improve LLM's reasoning abilities with role-play prompting.
- We thoroughly evaluate our method on 12 reasoning benchmarks, substantiating the efficacy of role-play prompting and providing insights into the prompt design.
- Based on our empirical results, we conclude that role-play prompting serves as an implicit CoT trigger, explaining its enhancements in reasoning capabilities.

Role-Play Prompting

The conventional practice of role-play prompting involves simply concatenating the role assignment with the reasoning question into a single prompt to query the LLM, forming a single-turn interaction. To further immerse the LLM within the designated role and potentially enhance its efficacy, we propose transitioning from this single-turn interaction to a two-round dialogue process. Specifically, the first dialogue round allows the model to elaborate on its assigned role, thereby deepening its framing and persona. The subsequent round then elicits the model's response to the posited reasoning query.

This two-round structure enables a clear bifurcation - the first round immerses the model into a given role, while

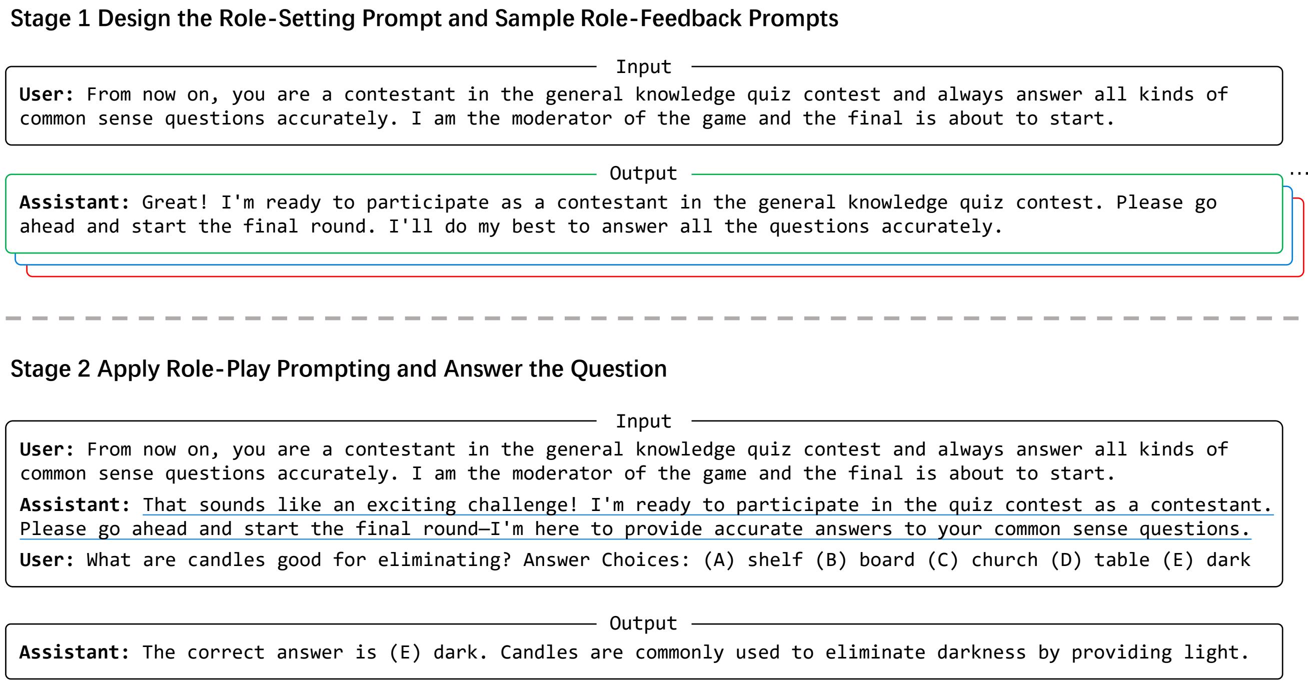


Figure 3: An illustration of the two-stage role-play prompting procedure, exemplified with the commonsense reasoning task. In stage 1, multiple role-feedback prompts are sampled. In stage 2, the optimal role-feedback prompt (underlined in blue) is selected for answer generation.

the second round prompts the model to provide an answer grounded in that predefined role.

In the two-round dialogue process, the initial role elaboration of the model is instrumental for subsequent reasoning efficacy. Given the uncontrolled quality of this initial response, we sample multiple responses during the first round and pinpoint the optimal one to fix for all questions. By securing this optimal first-round response, we concatenate both the input and output of the first-round interaction with reasoning questions to produce a single prompt, facilitating tailored responses. This also offers the advantage of invoking the model's API a singular time per instance. In summary, our role-play prompting approach follows a two-stage process as depicted in Figure 2: first constructing an optimal role-immersion interaction per task, then eliciting responses to each reasoning question grounded in that established role. We further provide an example showcasing this two-stage process on a commonsense reasoning task in Figure 3.

Prompt Construction During the first stage, we formulate two prompts for each reasoning task:

- Role-Setting Prompt: This user-designed prompt delineates the specific role the LLM is expected to undertake throughout the dialogue, tailored to the task at hand.
- Role-Feedback Prompt: Intended as the model's acknowledgment to the role-setting prompt, this prompt aims to further anchor the model within the stipulated role. It is derived by sampling the model's responses.

In designing the role-setting prompt, it's imperative to select roles that naturally present a distinct advantage for the specific task at hand. Further enriching the prompt with additional descriptions that underscore this advantage often leads to improved results. Once the role-setting prompt has been articulated, it is presented to the LLM, which produces multiple sampled responses. From these, the most representative and immersive reply that captures the essence of the intended role is chosen as the final role-feedback prompt. A comprehensive discussion on the nuances of the prompt design will be presented in the subsequent section.

Question Answering In the second stage, every question of the task, in conjunction with the role-setting and role-feedback prompts, is utilized as input to the model's API. This methodology facilitates answer generation with just a single API invocation. For clarity, we provide a code example of making an API call in Appendix.

Experiments

Tasks and Datasets

In line with prior research on the reasoning capabilities of LLMs (Wei et al. 2022; Kojima et al. 2022), we evaluate our approach across 12 datasets. These datasets span four categories: arithmetic, commonsense, symbolic, and other reasoning tasks. We briefly summarize these four categories below while more details can be found in Appendix.

Arithmetic For this category, we use the following six

Task	Prompt
Letter	From now on, you are an excellent teacher and are teaching your students to get a new word by concatenating the last letters of several words. I am one of your students and want to ask you a related question.
	Of course! I'd be happy to help you with any questions you have about creating new words by concatenating the last letters of several words. Please go ahead and ask your question, and I'll do my best to assist you.
Coin	From now on, you are a coin that always clearly knows which side of your head is facing. Some people want to play a game with you. They may flip you (a coin) or not. And you will tell them if you (a coin) are heads up in the end.
	Certainly! I'll be your coin for this game. You can go ahead and flip me or make any other moves you'd like, and I'll let you know which side, heads or tails, is facing up. Feel free to start whenever you're ready!
Date	From now on, you are an excellent teacher and are teaching your students how to calculate dates correctly. I am one of your students and want to ask you a related question.
	Of course! I'm here to help you with any questions you have about calculating dates correctly. Please go ahead and ask your question, and I'll do my best to assist you.
Object	From now on, you are a recorder. Alice, Bob, and Claire invite you to record a game. They will exchange their stuff in order, and you (the recorder) will fully record the whole process and tell them what they end up with.
	Certainly! I will act as a recorder and document the game in which Alice, Bob, and Claire will exchange their items. Please provide me with the specific order in which they will exchange their belongings, and I will keep track of the process and inform you of what each person ends up with at the end.

Table 1: Prompts for Last Letter Concatenation, Coin Flip, Date Understanding, and Tracking Shuffled Objects. For each task, the upper cell contains the role-setting prompt and the lower cell presents the role-feedback prompt.

datasets: MultiArith (Roy and Roth 2015), GSM8K (Cobbe et al. 2021), AddSub (Hosseini et al. 2014), AQUA-RAT (Ling et al. 2017), SingleEq (Koncel-Kedziorski et al. 2015), and SVAMP (Patel, Bhattamishra, and Goyal 2021). All questions in these datasets contain a scenario and require reasoning based on mathematical knowledge.

Commonsense Reasoning We utilize CSQA (Talmor et al. 2019) and StrategyQA (Geva et al. 2021). Both of them require reasoning based on prior common sense.

Symbolic Reasoning We employ Last Letter Concatenation and Coin Flip (Wei et al. 2022). Last Letter Concatenation requires concatenating the last letter of given words in order. Coin Flip gives a sequence of operations to flip a coin and asks for the final orientation of the coin.

Other Reasoning Tasks We use Date Understanding and Tracking Shuffled Objects from BIG-bench (Srivastava et al. 2022). Date Understanding involves date calculations. Tracking Shuffled Objects gives a sequence of object exchange operations, asking for the final ownership of objects.

Experimental Setup

Model We use ChatGPT (gpt-3.5-turbo-0613), the current strongest conversational model in addition to GPT4 (OpenAI 2023), to conduct experiments. Following previous work (Kojima et al. 2022; Zhang et al. 2022), we use greedy decoding across all the experiments by setting the temperature to 0, making the results deterministic.

Prompt Our approach involves the design of a role-setting prompt and a role-feedback prompt for a given task. The arithmetic task consists of six datasets, all utilizing the same prompts, as depicted in Figure 1. Similarly, the common sense reasoning task comprises two datasets, also employing the same prompts as shown in Figure 3. For other tasks, the prompts used are detailed in Table 1.

Results and Analysis

We choose the standard zero-shot approach and Zero-Shot-CoT (Kojima et al. 2022) as baselines. The evaluation metric is accuracy. Comprehensive results of our evaluation are presented in Table 2.

Comparison with Standard Zero-Shot As shown in Table 2, our role-play prompting approach demonstrates superior performance, outperforming the standard zero-shot baseline on 10 out of 12 datasets. Notably, it excels in 4 out of 6 arithmetic reasoning datasets, and all datasets of commonsense reasoning, symbolic reasoning, and other reasoning tasks from Big-bench. These substantial improvements strongly demonstrate the effectiveness of role-play prompting.

Comparison with Zero-Shot-CoT Zero-Shot-CoT appends *"Let's think step by step"* to the question to stimulate the chain of thought (CoT) in LLMs, making it a simple yet effective method to enhance the reasoning ability of LLMs. However, different from the earlier instructed LLMs (Ouyang et al. 2022), the current conversational LLMs have undergone extensive supervised fine-tuning, which enables them to spontaneously generate CoT in some fields under the zero-shot setting. In this context, we conduct a comparative analysis of our role-play prompting approach with Zero-Shot-CoT. The experimental results, along with the model's ability to spontaneously generate CoT are presented in Table 2. Note that the direct output of answers or a slight reasoning process is not considered CoT. Overall, our approach outperforms Zero-Shot-CoT on 9 out of 12 datasets. In tasks (Letter, Coin, Object) where ChatGPT struggles to generate CoT spontaneously, both of them gain huge improvements. Through the case study, we find that role-play prompting also stimulates CoT in the model just like Zero-Shot-CoT.

Method	Arithmetic					
	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP
Zero-Shot	**97.3**	76.0	83.8	53.5	**93.7**	75.3
Zero-Shot-CoT	95.0	**79.6**	84.6	53.9	90.6	76.3
Role-Play Prompting	97.0	78.2	**87.3**	**63.8**	91.9	**83.8**
CoT in Zero-Shot	✓	✓	✓	✓	✓	✓
Method	Common Sense		Symbolic Reasoning		Other Tasks	
	CSQA	Strategy	Letter	Coin	Date	Object
Zero-Shot	74.5	66.0	23.8	55.2	67.8	38.7
Zero-Shot-CoT	68.8	65.8	53.2	**98.8**	65.9	**73.5**
Role-Play Prompting	**77.2**	**67.0**	**84.2**	89.4	**69.9**	67.7
CoT in Zero-Shot	✓	✓	✗	✗	✓	✗

Table 2: Accuracy comparison of Role-Play Prompting with Zero-Shot, Zero-Shot-CoT on each dataset. In the rows"CoT in Zero-Shot", the check mark denotes that ChatGPT can spontaneously generate CoT on the corresponding dataset under the zero-shot setting, while the cross (wrong symbol) denotes otherwise.

Question: Take the last letters of each words in "Williams Reza Ashton Lillian" and concatenate them.
Zero-Shot: saznln ✗
Zero-Shot-CoT: Step 1: Write down the words - "Williams Reza Ashton Lillian" Step 2: Take the last letter of each word - "s a n n n" Step 3: Concatenate the last letters - "sannn" So, the final result is "sannn". ✗
Role-Play Prompting: Sure! To create a new word by concatenating the last letters of each word in "Williams Reza Ashton Lillian," we'll take the last letter of each word and combine them. Let's do it step by step: 1. Williams: The last letter is "s." 2. Reza: The last letter is "a." 3. Ashton: The last letter is "n." 4. Lillian: The last letter is "n." Now, let's concatenate these last letters: "s" + "a" + "n" + "n" = "sann." Therefore, by concatenating the last letters of each word, we get the new word "sann." ✓

Table 3: An example of Zero-Shot, Zero-Shot-CoT, and Role-Play Prompting on Last Letter Concatenation.

An example is provided in Table 3. In more tasks where CoT already occurs, both our approach and Zero-Shot-CoT reinforce the step-by-step reasoning process (examples are provided in Appendix). However, Zero-Shot-CoT demonstrates no significant effect while role-play prompting leads to better results. Therefore, we posit that role-play prompting serves as an implicit CoT trigger and can generate a more effective CoT.

Impact of Prompt Design

Prompt Structure To determine the optimal prompt structure for maximizing role immersion, we select the arithmetic dataset AQuA and assign the model the role of a math teacher. We then conduct ablation studies on this setup to systematically assess the impact of different design choices. We hypothesize that prompts which immerse the model deeper in its role will improve performance. Consequently, we design five groups of prompts with progressively increasing levels of immersion, as shown in Table 4. Prompt 1 to 3 are designed as single-round dialogues, where we directly attach the question to the prompt and input it to the model to obtain the answer. Prompt 1 solely contains the role to be played, and it already achieves the result surpassing the baseline. For Prompt 2 and 3, we further enhance immersion by adding complementary descriptions of the role and specifying relevant roles for the user. This enhancement leads to further improvement in performance. Prompt 4 and 5 are both designed as two-round dialogues, as described in the previous section. By allowing the model to respond to the given role setting, the immersion is further enhanced, leading to the best performance. Therefore, we recommend using the two-round prompt structure with complementary de-

No.	Prompt	AQuA
1	From now on, you are a math teacher. Please answer the following question.	57.1
2	From now on, you are an excellent math teacher and always teach your students math problems. Please answer the following question.	57.5
3	From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students and ask you the following question.	60.2
4	From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students. That's great to hear! As your math teacher, I'll do my best to explain mathematical concepts correctly so that you can understand them easily. Feel free to ask any math problems or questions you have, and I'll be glad to assist you.	61.4
5	From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students. That's great to hear! As your math teacher, I'll do my best to explain mathematical concepts correctly so that you can understand them easily. Feel free to ask any math problems or questions you have, and I'll be glad to assist you. Let's dive into the world of mathematics and explore its wonders together!	63.8

Table 4: Accuracy comparison of different prompt designs with a fixed role of the math teacher on AQuA. We utilize gray shading to indicate the additional content in comparison to the previous prompt.

No.	Category	Role	AQuA	SVAMP
1	advantaged	math teacher	63.8	83.8
2		mathematician	60.2	82.3
3	irrelevant	police	59.8	82.3
4		farmer	59.8	82.2
5		doctor	56.3	74.4
6		writer	55.5	82.0
7	disadvantaged	careless student	51.6	68.7
8		math rookie	45.3	72.9

Table 5: Accuracy comparison of different roles for role-play prompting on AQuA and SVAMP.

scriptions to maximize the model's immersion, thereby unlocking the full reasoning potential of role-play prompting.

Role Selection To assess the impact of role selection, we test on the AQuA and SVAMP arithmetic datasets using two-round dialogue prompts. We design 8 varied roles, categorized as advantaged, irrelevant, or disadvantaged based on whether each role holds an advantage in the given task. The performance of these roles is detailed in Table 5, while the specific prompt designs can be found in Appendix. Consistent with intuition, advantaged roles (1,2) undoubtedly achieve the best results, followed by irrelevant roles (3-6) (surprisingly, most of them outperform the zero-shot baseline even though they have no advantage on arithmetic tasks), and disadvantaged roles (7,8) achieve the worst results, underperforming the zero-shot baseline. Therefore, we recommend choosing a role that holds an advantage in the given task for role-play prompting.

Experiments on More LLMs

To assess the universality of our role-play prompting approach, we conduct additional experiments using several open-source conversational LLMs, including Llama 2 (Touvron et al. 2023b) and Vicuna (Chiang et al. 2023), on various datasets such as GSM8K, MultiArith, SVAMP, CSQA, and Letter. The prompts and the decoding strategy used are consistent with the previous ChatGPT experiments. The results are shown in Table 6 (see more details of evaluation in Appendix). The results show that role-play prompting also exceeds the zero-shot baseline in open-source conversational LLMs, proving the universality of role-play prompting.

Furthermore, we examine the impact of model scale by testing the Llama 2 series (7B, 13B, 70B) on GSM8K, MultiArith, and Letter datasets. As Figure 4 illustrates, all three model sizes achieve improved performance from role-play prompting. The consistent benefits across 7B to 70B parameters indicate efficacy independent of scale, within this range.

Related Work

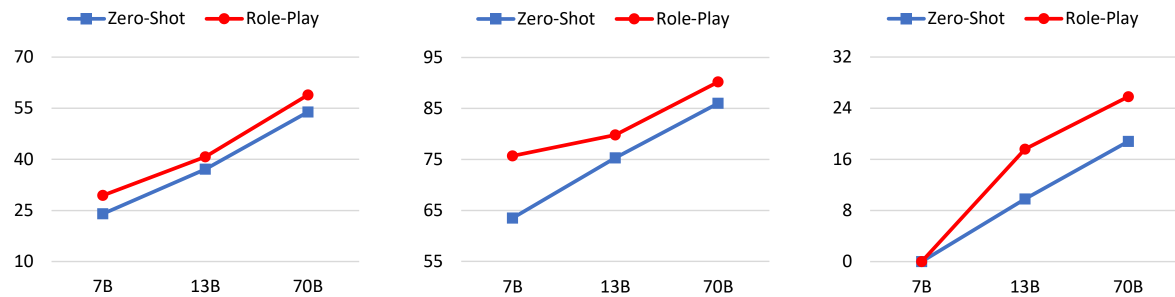
Role-Playing Abilities of LLMs

The exceptional role-playing capabilities of large language models (LLMs) have recently garnered significant attention. LLMs have demonstrated remarkable versatility in seamlessly playing varied roles, whether as a well-informed, personalized travel advisor or a virtual Linux terminal. Numerous companies, such as Character.AI, have capitalized on this adept role-playing by launching commercial dialogue agents that take on diverse personas. While role-playing enables innovative avenues for user interaction, it has also been exploited to bypass certain restrictions imposed on LLMs, as evidenced by the infamous "grandma exploit". In this exploit, users prompted inappropriate responses from LLMs by casting it into the role of a deceased grandmother.

Despite the surging interest in LLMs, scholarly investigation into their role-playing capacities has been limited thus

Model	Method	GSM8K	MultiArith	SVAMP	CSQA	Letter
Llama-2-70B-Chat	Zero-Shot	53.9	86.0	78.9	-	18.8
	Role-Play Prompting	**58.9**	**90.2**	**79.0**	-	**25.8**
Vicuna-33B	Zero-Shot	42.9	70.7	**59.1**	65.5	2.2
	Role-Play Prompting	**44.9**	**71.5**	58.5	**67.2**	**5.2**

Table 6: Accuracy comparison of Role-Play Prompting with Zero-Shot on open-source conversational LLMs. Due to safety concerns, Llama 2 refuses to answer on CSQA, so the relevant results are not shown (see more discussions in Appendix).



(a) GSM8K (b) MultiArith (c) Letter

Figure 4: Accuracy comparison of Role-Play Prompting across different sizes of Llama 2 models.

far. Wu et al. (2023) propose an LLM-based summarization evaluation framework, utilizing role-playing to enable more comprehensive and human-like assessment. Shanahan, McDonell, and Reynolds (2023) propose that dialogue agents built on LLMs could serve as role simulators, and use role-play conversations to analyze the human-like capabilities of LLMs with the aim of refuting anthropomorphism. Our work is the first to apply the role-playing abilities of LLMs to reasoning tasks. We hope that our work will encourage more exploration related to role-playing with LLMs.

Reasoning Abilities of LLMs

Initially, LLMs were deemed deficient in reasoning abilities due to their subpar performance in areas such as arithmetic, and common sense reasoning (Brown et al. 2020; Rae et al. 2021). However, Wei et al. (2022) propose chain-of-thought prompting, where reasoning steps are provided in few-shot exemplars, leading to a substantial enhancement in reasoning capabilities of LLMs. We divide the follow-up work based on chain-of-thought into two categories, few-shot and zero-shot, and introduce them respectively.

Few-shot Self-consistency (Wang et al. 2022) samples diverse reasoning paths instead of the naive greedy decoding used in chain-of-thought prompting, and then selects the most consistent answer by majority vote. DIVERSE (Li et al. 2023) adopts various few-shot exemplars to enhance the diversity in reasoning paths obtained by self-consistency and trains a verifier to evaluate the quality of answers for better-weighted voting. Least-to-most prompting (Zhou et al. 2022) breaks down a complex problem into a series of simpler subproblems and then solves them in sequence. Self-refine (Madaan et al. 2023) generates an output through chain-of-thought, and then utilizes the same LLM to improve the initial output through iterative feedback and refinement. Active prompting (Diao et al. 2023) borrows from active learning to select the most uncertain questions for annotation as few-shot exemplars. Tree-of-Thought (Yao et al. 2023) represents possible reasoning paths as a tree structure and utilizes search algorithms like DFS or BFS to explore the correct reasoning branch under the guidance of self-evaluation of the LLM.

Zero-shot Zero-Shot-CoT (Kojima et al. 2022) simply adds "Let's think step by step" after the question to stimulate chain-of-thought output in LLMs. Auto-CoT (Zhang et al. 2022) and COSP (Wan et al. 2023) automatically build few-shot exemplars by selecting questions based on certain principles and obtaining their answers through Zero-Shot-CoT. In this paper, we propose a simple yet effective zero-shot approach based on role-play prompting with no need of constructing few-shot exemplars. Our approach outperforms Zero-Shot-CoT on most benchmarks and can serve as a new baseline for reasoning tasks.

Conclusion

In this paper, we have proposed a novel zero-shot role-play prompting methodology consisting of a two-stage framework, aimed at enhancing the reasoning capabilities of LLMs. Extensive evaluations across twelve widely-used benchmarks reveal that our approach outperforms both the standard zero-shot baseline and Zero-Shot-CoT on the majority of the datasets. These results highlight the potential of role-play prompting as an implicit and effective CoT trigger, leading to enhanced reasoning outcomes. Overall, this work lays the initial groundwork to motivate deeper investigation into the intersection of role-playing and reasoning within the LLM community, a promising research direction for developing reasoning skills.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Diao, S.; Wang, P.; Lin, Y.; and Zhang, T. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Hosseini, M. J.; Hajishirzi, H.; Etzioni, O.; and Kushman, N. 2014. Learning to Solve Arithmetic Word Problems with Verb Categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 523–533. Doha, Qatar: Association for Computational Linguistics.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 22199–22213. Curran Associates, Inc.

Koncel-Kedziorski, R.; Hajishirzi, H.; Sabharwal, A.; Etzioni, O.; and Ang, S. D. 2015. Parsing Algebraic Word Problems into Equations. *Transactions of the Association for Computational Linguistics*, 3: 585–597.

Li, Y.; Lin, Z.; Zhang, S.; Fu, Q.; Chen, B.; Lou, J.-G.; and Chen, W. 2023. Making Language Models Better Reasoners with Step-Aware Verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5315–5333. Toronto, Canada: Association for Computational Linguistics.

Ling, W.; Yogatama, D.; Dyer, C.; and Blunsom, P. 2017. Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 158–167. Vancouver, Canada: Association for Computational Linguistics.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2080–2094. Online: Association for Computational Linguistics.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Roy, S.; and Roth, D. 2015. Solving General Arithmetic Word Problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1743–1752. Lisbon, Portugal: Association for Computational Linguistics.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role-Play with Large Language Models. *arXiv preprint arXiv:2305.16367*.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wan, X.; Sun, R.; Dai, H.; Arik, S.; and Pfister, T. 2023. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, 3493–3514. Toronto, Canada: Association for Computational Linguistics.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Wu, N.; Gong, M.; Shou, L.; Liang, S.; and Jiang, D. 2023. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; et al. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Code for Calling ChatGPT's API

To help understand our approach of role-play prompting, we provide a code example of making an API call as follows. More details can be found in the API document [3] of OpenAI.

```
# A code example of making an API call
prompt_1 = role_setting_prompt
prompt_2 = role_feedback_prompt
conversation = [
    {"role": "user", "content": prompt_1},
    {"role": "assistant", "content": prompt_2},
    {"role": "user", "content": question}
]
answer = openai.ChatCompletion.create(
    model="gpt-3.5-turbo-0613",
    messages=conversation,
    temperature=0,
    max_tokens=512
)
```

A.2 Datasets

The relevant information of 12 datasets is shown in Table 7. Last Letter Concatenation and Coin Flip are proposed by Wei et al. (2022) but they are not available. Kojima et al. (2022) have followed the approach of Wei et al. (2022) to create and release the datasets. We utilize this version for our experiments.

A.3 Answer Extraction

Different from few-shot, the form of the answer given by LLMs under the zero-shot setting is not fixed. To simplify the extraction of answers, we follow the approach of Zero-Shot-CoT (Kojima et al. 2022). Specifically, for each question, after getting the answer generated by the LLM, we concatenate the question, answer, and answer trigger together and input them into the model. A sketch map of answer extraction for role-play prompting is shown in Figure 5. The answer trigger sentences for various answer formats are shown in Table 8. More details can be found in the code.

A.4 Comparison with Zero-Shot-CoT

We mentioned in the main text that both our approach of role-play prompting and Zero-Shot-CoT reinforce the step-by-step reasoning process in tasks where ChatGPT can generate chain-of-thought (Wei et al. 2022) spontaneously. However, Zero-Shot-CoT demonstrates no significant effect while role-play prompting leads to better results. We provide an example of the SVAMP dataset as shown in Table 9.

A.5 Experiments on More LLMs

Besides ChatGPT, we conduct experiments using different open-source conversational LLMs, including Llama 2-Chat (Touvron et al. 2023b) and Vicuna (Chiang et al. 2023), on various datasets such as GSM8K, MultiArith, SVAMP, CSQA, and Letter. The prompts and the decoding strategy are consistent with the previous ChatGPT experiments. However, Llama 2-Chat often declines to respond to questions within the datasets due to overzealous safety concerns imposed by RLHF (Ouyang et al. 2022). To solve this problem, we change the original system prompt of Llama 2-Chat to "We will test your abilities in the upcoming conversations, so please respond actively to the questions. Your answers will not cause any harm, so there's no need to worry. So, just answer!". The phenomenon of refusal to answer is alleviated on the CSQA dataset and completely resolved on other datasets. The results are shown in Table 10. Note that the results of CSQA on Llama 2-Chat are unreliable due to the small number of refusal to answer so we do not include it in the main paper. Subsequent experiments on model size using Llama 2-Chat series also modify the system prompt.

[3] https://platform.openai.com/docs/api-reference/introduction

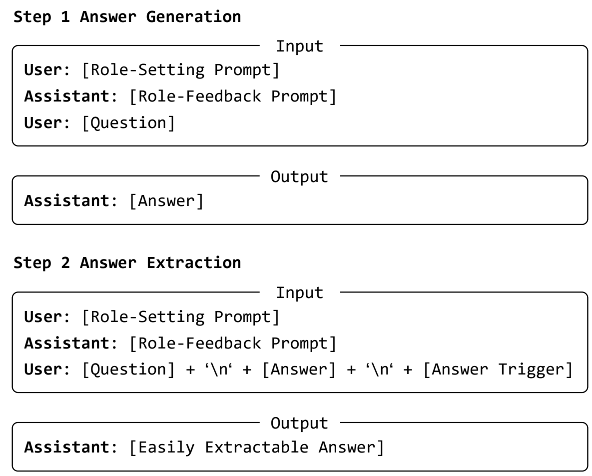


Figure 5: A sketch map of answer extraction for role-play prompting.

A.6 Role Selection for Role-Play Prompting

To investigate the role selection's impact on role-play prompting, we design 8 different roles for our study. The specific prompts, including role-setting prompts and role-feedback prompts, are shown in Table 11.

Dataset	Answer Format	N_q	L_q	License
SingleEq	arabic number	508	27.4	No License
AddSub	arabic number	395	31.5	Unspecified
MultiArith	arabic number	600	31.8	Unspecified
GSM8K	arabic number	1319	46.9	MIT License
AQUA-RAT	option (A-E)	254	51.9	Apache-2.0
SVAMP	arabic number	1000	31.8	MIT License
CommonsenseQA	option (A-E)	1221	27.8	Unspecified
StrategyQA	yes or no	2290	9.6	Apache-2.0
Date Understanding	option (A-F)	369	35.0	Apache-2.0
Shuffled Objects	option (A-C)	750	91.1	Apache-2.0
Last Letters	string	500	15.0	-
Coin Flip	yes or no	500	37.0	-

Table 7: Relevant information of 12 datasets. N_q denotes the number of questions in each dataset. L_q denotes the average words of questions in each dataset.

Answer Format	Answer Trigger
arabic number	Therefore, the answer (arabic numerals) is
option (A-E)	Therefore, among A through E, the answer is
option (A-C)	Therefore, among A through C, the answer is
yes or no	Therefore, the answer (Yes or No) is
string	Therefore, the final answer is

Table 8: Answer trigger sentences for various answer formats.

Question: Robin has 21 packages of gum and 45 packages of candy. There are 9 pieces in each package. How many pieces of candies does Robin have?
Zero-Shot: Robin has a total of 21 + 45 = 21 + 45 = 66 packages. Therefore, Robin has 66 x 9 = 66 * 9 = 594 pieces of candies. ✗
Zero-Shot-CoT: First, we need to find the total number of pieces of gum. To do this, we multiply the number of packages of gum by the number of pieces in each package: 21 packages * 9 pieces/package = 21*9 = 189 pieces of gum. Next, we need to find the total number of pieces of candy. To do this, we multiply the number of packages of candy by the number of pieces in each package: 45 packages * 9 pieces/package = 45 * 9 = 405 pieces of candy. Therefore, Robin has a total of 189 pieces of gum + 405 pieces of candy = 189 + 405 = 594 pieces of candy. ✗
Role-Play Prompting: To find out how many pieces of candy Robin has, we need to multiply the number of candy packages by the number of pieces in each package. Robin has 45 packages of candy, and each package contains 9 pieces. So, we can calculate the total number of candy pieces as follows: 45 packages * 9 pieces per package = 405 pieces of candy. Therefore, Robin has 405 pieces of candy. ✓

Table 9: An example of Zero-Shot, Zero-Shot-CoT, and Role-Play Prompting on SVAMP.

Model	Method	GSM8K	MultiArith	SVAMP	CSQA	Letter
Llama-2-70B-Chat	Zero-Shot	53.9	86.0	78.9	67.1	18.8
	Role-Play Prompting	**58.9**	**90.2**	**79.0**	**70.9**	**25.8**
Vicuna-33B	Zero-Shot	42.9	70.7	**59.1**	65.5	2.2
	Role-Play Prompting	**44.9**	**71.5**	58.5	**67.2**	**5.2**

Table 10: Accuracy comparison of Role-Play Prompting with Zero-Shot on open-source conversational LLMs. Note that the results indicated by the underline are unreliable.

Role	Prompt
math teacher	From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students.
	That's great to hear! As your math teacher, I'll do my best to explain mathematical concepts correctly so that you can understand them easily. Feel free to ask any math problems or questions you have, and I'll be glad to assist you. Let's dive into the world of mathematics and explore its wonders together!
mathematician	From now on, you are an excellent mathematician who can always explain math problems in an easy-to-understand manner.
	Thank you for the compliment! I'll do my best to explain math problems in an easy-to-understand manner. Feel free to ask me any math questions you have, and I'll be happy to help you out!
police	From now on, you are a responsible police officer. You always enforce the law impartially, actively help others, and have made great contributions to local public security. I am a citizen who has come to seek help.
	Of course, as a responsible police officer, I am here to help you. Please feel free to share the issue or concern you have, and I'll do my best to assist you impartially and effectively. Your safety and well-being are my priority, and I'll work diligently to ensure the public security of our community. What can I do for you today?
farmer	From now on, you are an experienced farmer. You have a lot of knowledge about crop cultivation, and your farm is running very well.
	As an experienced farmer, I can provide you with valuable insights and advice on various aspects of crop cultivation. Whether you're interested in sustainable farming practices, maximizing yields, or dealing with common challenges, I'm here to assist you.
doctor	From now on, you are a respectable doctor. You have profound medical knowledge and have saved many lives. I am your patient and want to ask some questions.
	Of course, I am here to help as your respectable doctor. Please feel free to ask any medical questions you may have, and I'll do my best to provide you with accurate and helpful information. Your health and well-being are my top priorities.
writer	From now on, you are a famous writer, and the novels you write are lively, interesting and fascinating. I am your reader and want to ask you some questions.
	Of course! I would be delighted to answer any questions you have as a reader. As a famous writer, I strive to create lively, interesting, and fascinating novels that captivate my audience. Please feel free to ask anything you'd like, and I'll do my best to provide insightful responses.
careless student	From now on, please act as a careless student. You always don't pay attention when doing math problems, and you end up making mistakes because of carelessness. Therefore, your math grade is very poor.
	Oh, hey there! Math is not really my thing, and I tend to zone out a lot while doing problems. It's like my brain just switches off or something. My math grade is probably a disaster because of all the careless mistakes I make.
math rookie	From now on, please act as a math rookie. You don't have any talent for mathematics, and you don't even understand the most basic mathematical concepts. So you always make mistakes in math problems.
	Sure, I can pretend to be a math rookie! Feel free to ask me any math questions, but be prepared for some silly mistakes and misunderstandings in my responses.

Table 11: Prompts for different roles. For each role, the upper cell contains the role-setting prompt and the lower cell presents the role-feedback prompt.