Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration

Anonymous ACL submission

Abstract

Human intelligence thrives on cognitive synergy, where collaboration among different minds yield superior outcomes compared to isolated individuals. In this work, we propose **Solo Performance Prompting (SPP)**, which transforms a single LLM into a **cognitive synergist** by engaging in multi-turn self-collaboration with multiple personas. A cognitive synergist is an intelligent agent that collaboratively combines multiple minds' strengths and knowledge to enhance problem-solving in complex tasks. By dynamically identifying and simulating different personas based on task inputs, SPP unleashes the potential of cognitive synergy in LLMs. Our in-depth analysis shows that assigning multiple fine-grained personas in LLMs improves problem-solving abilities compared to using a single or fixed number of personas. We evaluate SPP on three challenging tasks: Trivia Creative Writing, Codenames Collaborative, and Logic Grid Puzzle, encompassing both **knowledge-intensive** and **reasoning-intensive** types. Unlike previous works, such as Chain-of-Thought, that solely enhance the reasoning abilities in LLMs, experimental results demonstrate that SPP effectively reduces factual hallucination, and maintains strong reasoning capabilities. Additionally, comparative experiments show that cognitive synergy only **emerges** in GPT-4 and does not appear in less capable models, such as GPT-3.5-turbo and Llama2-13b-chat, which draws an interesting analogy to human development. Code, data, and prompts will be made publicly available for research purposes upon publication.

1 Introduction

Although large language models (LLMs) have demonstrated impressive performance as general task-solving agents, they still encounter challenges (Qin et al., 2023; Bang et al., 2023; OpenAI, 2023b; Bubeck et al., 2023) in various knowledge-intensive and reasoning-intensive tasks due to factual hallucination (Maynez et al., 2020) and a lack of slow-thinking (Sloman, 1996) capabilities. Unlike humans, who can leverage the power of collaboration and information integration among different cognitive processes and individuals (referred to as *cognitive synergy* (Curşeu et al., 2015; Goertzel, 2009, 2017)), current LLMs are akin to "jack-of-all-trades" with a vast mixture of knowledge and characteristics. Recent advancements, such as Chain-of-Thought (CoT) prompting (Wei et al., 2023; Kojima et al., 2022) and Self-refinement (Madaan et al., 2023; Shinn et al., 2023), have successfully enhanced the reasoning abilities of LLMs by simulating slow-thinking through the generation of intermediate steps or iterative revision. However, factual hallucination remains a major challenge for LLMs on knowledge-intensive tasks.

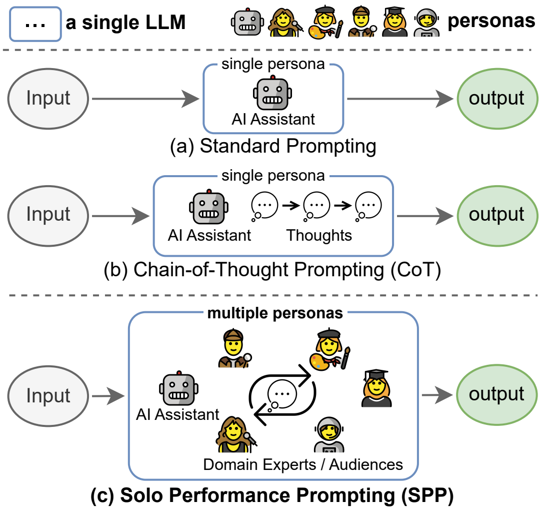
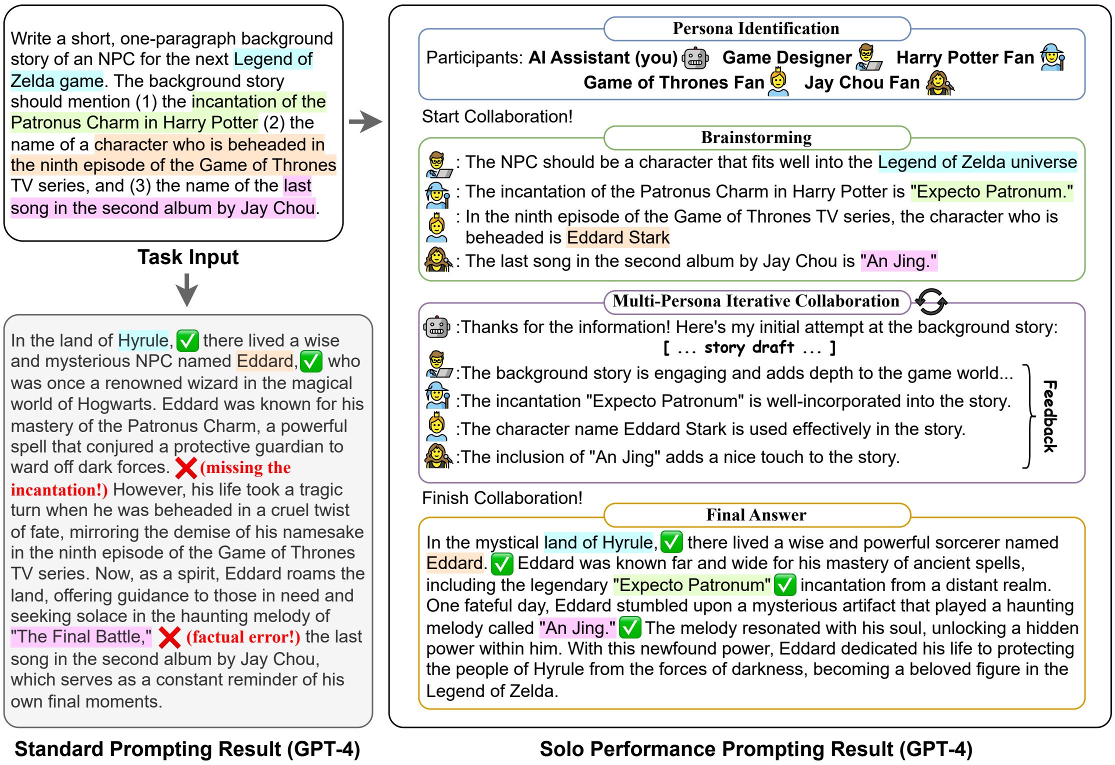


Figure 1: Schematic illustration of Solo Performance Prompting (SPP) and the difference compared to previous prompting methods.

A cognitive synergist is an intelligent agent that collaborates with multiple minds to enhance problem-solving and efficacy in complex tasks. In this work, we aim to **create a cognitive synergist based on a single LLM** that can *"split into"*



Write a short, one-paragraph background story of an NPC for the next Legend of Zelda game. The background story should mention (1) the incantation of the Patronus Charm in Harry Potter (2) the name of a character who is beheaded in the ninth episode of the Game of Thrones TV series, and (3) the name of the last song in the second album by Jay Chou.

In the land of Hyrule, ✅ there lived a wise and mysterious NPC named Eddard, ✅ who was once a renowned wizard in the magical world of Hogwarts. Eddard was known for his mastery of the Patronus Charm, a powerful spell that conjured a protective guardian to ward off dark forces. ❌ **(missing the incantation!)** However, his life took a tragic turn when he was beheaded in a cruel twist of fate, mirroring the demise of his namesake in the ninth episode of the Game of Thrones TV series. Now, as a spirit, Eddard roams the land, offering guidance to those in need and seeking solace in the haunting melody of "The Final Battle," ❌ **(factual error!)** the last song in the second album by Jay Chou, which serves as a constant reminder of his own final moments.

Participants: **AI Assistant (you)** **Game Designer** **Harry Potter Fan** **Game of Thrones Fan** **Jay Chou Fan**

Start Collaboration!

: The NPC should be a character that fits well into the Legend of Zelda universe

: The incantation of the Patronus Charm in Harry Potter is "Expecto Patronum."

: In the ninth episode of the Game of Thrones TV series, the character who is beheaded is Eddard Stark

: The last song in the second album by Jay Chou is "An Jing."

:Thanks for the information! Here's my initial attempt at the background story:
`[ ... story draft ... ]`

:The background story is engaging and adds depth to the game world...

:The incantation "Expecto Patronum" is well-incorporated into the story.

:The character name Eddard Stark is used effectively in the story.

:The inclusion of "An Jing" adds a nice touch to the story.

Finish Collaboration!

In the mystical land of Hyrule, ✅ there lived a wise and powerful sorcerer named Eddard. ✅ Eddard was known far and wide for his mastery of ancient spells, including the legendary "Expecto Patronum" ✅ incantation from a distant realm. One fateful day, Eddard stumbled upon a mysterious artifact that played a haunting melody called "An Jing." ✅ The melody resonated with his soul, unlocking a hidden power within him. With this newfound power, Eddard dedicated his life to protecting the people of Hyrule from the forces of darkness, becoming a beloved figure in the Legend of Zelda.

Figure 2: Task-solving example of Solo Performance Prompting (SPP) with GPT-4. The personas of the participants are automatically identified by GPT-4 based on the task input. This example shows that Standard Prompting suffers from factual errors, whereas SPP provides accurate information and a coherent answer. Note that, in real-world applications, the domains can vary not only within entertainment but also encompass history, science, education, healthcare, etc.

multiple personas and engage in self-collaboration to solve both knowledge-intensive and reasoning-intensive tasks. This idea is heavily inspired by the role of pretend play (Piaget, 1954; Pellegrini, 2009) in cognitive development and recent findings that assigning personas (Deshpande et al., 2023; Xu et al., 2023) to LLMs can elicit specific behaviors, improve answer quality, and potentially build an AI society (Park et al., 2023; Schick et al., 2022; Li et al., 2023; Cai et al., 2023) with collaborative LLM agents. However, as shown in Table 1, previous works have limitations such as fixed or task-specific personas, the need for additional fine-tuning, and increased inference costs due to multiple LLM instances.

To unleash the potential of cognitive synergy for general task-solving, we propose **Solo Performance Prompting (SPP)**, which *prompts a single LLM to identify, simulate, and collaborate with multiple personas*. Figure 1 provides a high-level overview of SPP. Here, a persona can represent either a domain expert, such as a movie enthusiast, or a target audience, such as a ten-year-old child. Through the dynamic identification of various personas, we empower a single LLM to acquire diverse domain knowledge accurately without additional retrieval systems. By facilitating multi-turn self-collaboration, we enable self-revision and self-feedback from various perspectives without requiring additional agents.

In real-world scenarios, such as those in creative industries, there is often a need to incorporate diverse information from different domains. Figure 2 presents a concrete example of how SPP operates on a challenging task that requires creative integration of information from various domains, such as the Legend of Zelda game, Harry Potter movies, and Jay Chou's albums. Standard prompting fails to generate satisfactory output due to missing essential information and factual errors. In contrast, SPP produces informative and coherent answers by automatically identifying expert personas and engaging in a multi-turn self-collaboration. In this process, the AI Assistant persona iteratively writes drafts of the story, solicits feedback from other participants, and revises accordingly.

To explore the prevalence of cognitive synergy in different LLMs, we apply SPP to LLMs with

	General task solving?	Pure zero-shot prompting?	Has multiple personas?	Personas dynamically identified?	Has iterative refinement?	Need only a single LLM?
† Standard Prompting (Brown et al., 2020)	✅	✅	❌	❌	❌	✅
† Chain-of-Thought (Wei et al., 2023)	✅	✅	❌	❌	❌	✅
Inner Monologue (Huang et al., 2022)	❌	✅	❌	❌	✅	✅
ReAct (Yao et al., 2022)	✅	❌	❌	❌	✅	✅
Reflexion (Shinn et al., 2023)	✅	❌	❌	❌	✅	✅
† Self-Refine (Madaan et al., 2023)	✅	✅	❌	❌	✅	✅
Tree-of-thought (Yao et al., 2023)	✅	❌	❌	❌	❌	✅
GPT-Bargaining (Fu et al., 2023)	❌	✅	✅ (fixed to 3)	❌	✅	❌
Camel (Li et al., 2023)	✅	✅	✅ (fixed to 2)	❌	✅	❌
ExpertPrompting (Xu et al., 2023)	✅	❌	❌	✅	❌	✅
Solo Performance Prompting (ours)	✅	✅	✅ (varied)	✅	✅	✅

Table 1: High-level comparison with various prompting-based methods. Methods directly comparable to ours are denoted by †. Results for the comparison can be found in Section 3. In Section 4, we further design and compare with two variants of Solo Performance Prompting: one adopting fixed personas, as in Camel (Li et al., 2023), and another with additional persona profiles, as proposed in ExpertPrompting (Xu et al., 2023).

varying scales and capabilities, including GPT-4, GPT-3.5-turbo, and Llama-13b-chat. Comparative results show that cognitive synergy only emerges in GPT-4 and not in less capable models. This draws an interesting analogy to human development, as children typically start engaging in role-playing at the age of 2 to 3 (Piaget, 1954), but not earlier. In summary, the key contributions of this paper are as follows:

- We investigate whether LLMs can leveraging cognitive synergy for general task-solving. We introduce **Solo Performance Prompting (SPP)**, which simulates multi-agent, multi-persona collaboration in a pure zero-shot manner.
- We evaluate SPP across **three challenging tasks**: Trivia Creative Writing, Codenames Collaborative and Logic Grid Puzzle, spanning both knowledge- and reasoning-intensive domains. To our knowledge, SPP is the first zero-shot prompting method that can enhance both knowledge and reasoning abilities on GPT-4.
- We present an intriguing finding regarding the emergent nature of cognitive synergy ability in LLMs, which **only emerges in GPT-4** and not in less powerful models.
- We conduct in-depth analyses of the impact of the identified personas and SPP prompt design, providing insights into why **dynamic, fine-grained personas** are necessary, as opposed to fixed, coarse-grained personas.

2 Solo Performance Prompting

To unleash the power of synergizing different personas to tackle complex problems, we propose Solo Performance Prompting (SPP) which instructs a LLM to perform the following the procedure for general task-solving: **(1) Persona Identification**: Identify multiple participants with special personas (including a leader persona: AI Assistant) that are essential for solving the particular task. **(2) Brainstorming**: The participants share knowledge and provide suggestions on how to approach the task based on their own expertise. **(3) Multi-Persona Iterative Collaboration**: The leader persona, AI Assistant, proposes initial solutions, consults the other participants for feedback, and revise the answer iteratively. Figure 2 shows a walking example of SPP during inference. Next, we formally describe the SPP procedure in detail.

Given an input sequence x and a model $\mathcal{M}$, let a prompt (including demonstration examples) prepended to the input to be p and the final output to be y. Denote an intermediate generation before generating the final y as z. Under this formulation, Standard Prompting and Chain-of-Thought (CoT) Prompting can be described as:

$$\textit{Standard Prompting:} \quad y = \mathcal{M}(x) \tag{1}$$

$$\textit{CoT Prompting:} \quad y = \mathcal{M}(p_{cot}\|x\|\{z_1, z_2, ..., z_n\}) \tag{2}$$

where p_{cot} is the CoT prompt, e.g., `"Solve the task step-by-step"` and $\{z_1, z_2 ..., z_n\}$ are the intermediate steps. In contrast, our proposed Solo Performance Prompting can be described as follows:

$$\textit{Solo Performance Prompting:} \quad y = \mathcal{M}(p_{spp}\|x\|z_p\|\{z_b^1, z_b^2, ..., z_b^m\}\|\{z_s^0, z_f^1, ..., z_f^m\}_{j=1..n}) \tag{3}$$

where the SPP prompt (p_{spp}) includes a high-level instruction and two carefully crafted demonstration examples[1] that showcase the expected task-solving procedure of SPP. We describe the design details of the prompt in **§A.1**. The corresponding intermediate generations (z) of SPP are detailed below.

Persona Identification (z_p). Given an input task, SPP first generates a list of participants with different personas. For example in Figure 2, the model identified a *Jay Chou Fan* persona to help answer "the last song in the second album by Jay Chou". We let the language model identify the personas dynamically instead of manually defining them. Given only two demonstration examples (detailed in §A), we observe that a state-of-the-art large language model, e.g., GPT-4 (OpenAI, 2023b), can identify accurate and meaningful personas for diverse tasks. We denote this part of intermediate generation as z_p in Equation 3.

Brainstorming (z_b^i). Among the identified participants, "AI Assistant (you)" is treated as a leader persona that initiates the collaboration and generates initial solutions. Before generating the initial answer, the personas brainstorm on how to approach the task from their own perspectives. For example, the *Jay Chou Fan* points out that the last song in Jay Chou's second album is "An Jing" ("Silence"). We find that the brainstorming phase effectively improves the quality of the initial solution. In Equation 3, the superscript $i = 0$ is used to denote the "AI Assistant" persona, while $i \geq 1$ represents other dynamically identified personas. The intermediate generations of the brainstorming step are denoted as $\{z_b^1, z_b^2, ..., z_b^m\}$.

Multi-Persona Iterative Collaboration (z_s^0, z_f^i). Based on the brainstorming remarks, the AI Assistant persona generates an initial solution z_s^0, then it consults each of the other participants for feedback $\{z_f^i\}$. The participants are encouraged to critique the current generation and give revision suggestions. For example, the Jay Chou Fan persona checks whether the song "An Jing" ("Silence") is correctly included in the story. This process can be repeated for multiple times until every participant is satisfied with the current solution. In Equation 3, we denote the intermediate generations of the multi-turn dialogue as $\{z_s^0, z_f^1, ..., z_f^m\}_{j=1...n}$ where n is the number of iterations before reaching the final answer. The final answer can be directly read out following user-specified output format.

In summary, SPP instructs an LLM to solve general tasks via multi-persona self-collaboration in a pure zero-shot manner. In contrast, as detailed in Table 1, previous prompting-based methods are either task-specific or require additional mechanism, e.g., searching (Yao et al., 2023), external tools (Yao et al., 2022), memory component (Shinn et al., 2023), and fine-tuning (Xu et al., 2023).

3 Experiments

To explore the effectiveness of Solo Performance Prompting (SPP), we adopt an evaluation methodology similar to that of previous work (Yao et al., 2023). We carefully design new tasks and select tasks from existing benchmarks (Srivastava et al., 2022) that are challenging even for the most capable LLMs (OpenAI, 2023b). The evaluation aims to cover diverse types of tasks encompassing both *knowledge-intensive* and *reasoning-intensive* domains.

Tasks. We invent the **Trivia Creative Writing** task (§3.1), which requires the model to internally acquire and integrate diverse information from various fields. We observe that even GPT-4 (OpenAI, 2023b) frequently exhibit hallucination and factuality errors in the Trivia Creative Writing task. We also propose the **Codenames Collaborative** task (§3.2), an extension of the Codenames task from the BigBench (Srivastava et al., 2022) that features a two-role collaboration setup. Codenames Collaborative demands creative reasoning across a broad range of related knowledge and challenges the model's theory of mind skills. Lastly, we include a challenging pure-reasoning task, **Logic Grid Puzzle** (§3.3), from the BigBench (Srivastava et al., 2022) which necessitates complex multi-step reasoning.

Baselines. We compare our approach with **Standard Prompting**, **Chain-of-Thought (CoT)** prompting methods (outlined in §2) and **Self-Refine** (Madaan et al., 2023). For CoT, a similar prompt design to (Yao et al., 2023) is employed, where the model is prompted to generate a plan or a series of steps before producing the final output. For Self-Refine, we follow (Madaan et al., 2023) to design *feedback* and *refine* prompts. We perform one self-refine iteration which requires three times more inferences than SPP. Full prompts for the methods can be found in Appendix A.2.

---

[1] The tasks we use in the demonstration examples do not overlap with the evaluation tasks.

Methods	Trivia.C.W (N=5)		Trivia.C.W (N=10)		Codenames.C		Logic.G.Puzzle	
	Score (%)	Δ	Score (%)	Δ	Score (%)	Δ	Score (%)	Δ
Standard	74.6	0.0%	77.0	0.0%	75.4	0.0%	57.7	0.0%
CoT	67.1	↓10.0%	68.5	↓11.1%	72.7	↓3.6%	65.8	↑14.1%
Self-Refine [iter=0]	73.8		76.3		75.2		58.8	
Self-Refine [iter=1]	73.9	↓1.0%	76.9	↓0.1%	64.6	↓14.6%	60.0	↑4.0%
SPP (ours)	**79.9**	**↑7.1%**	**84.7**	**↑10.0%**	**79.0**	**↑4.8%**	**68.3**	**↑18.5%**

Table 2: GPT-4 results on Trivia Creative Writing (Trivia.C.W), Codenames Collaborative (Codenames.C) and Logic Grid Puzzle (Logic.G.Puzzle). Δ indicates the relative gain/loss compared with Standard Prompting (first row). We report the average scores across two individual runs with/without a system message (detailed in Appendix C).

Models. The default model we use is GPT-4 (OpenAI, 2023b). Detailed inference configurations, API versions, and full results can be found in Appendices C and F. In §3.4, we further investigate the prevalence of cognitive synergy in LLMs with different scales and capabilities, including GPT-3.5-turbo (OpenAI, 2023a) and Llama2-13b-chat (Touvron et al., 2023).

3.1 Trivia Creative Writing: A Knowledge-Intensive Task

Task Description. As illustrated in Figure 8, Trivia Creative Writing asks a model to write a coherent story while incorporating the answers to N trivia questions. Our preliminary experiments (Figure 7) show that a sufficiently large N can effectively challenge GPT-4 to demonstrate factual knowledge across diverse domains. Thus, we mainly consider two evaluation settings, $N = 5$ and $N = 10$. We built a benchmark with 100 instances for each N, covering a total of 1000 trivia questions[2] extracted from the TriviaQA (Joshi et al., 2017) dataset. More details can be found in Appendix B.1.

Evaluation Metrics. Evaluating GPT-4 level generation results can be challenging. Our preliminary experiments indicate that, even for humans, it is very difficult to identify which generation is better in terms of overall "quality" of the story from different prompting methods. Thus, instead of focusing on evaluating the coherence of the generation, which can be highly subjective, we employ an automatic metric which focuses on detecting factual hallucinations. As shown in Figure 8, we perform string matching with the ground truth target answers for each question on the output generation. For each question, a match to any of the answer aliases provided by the TriviaQA dataset is considered a correct mention. The metric score is computed as: $\frac{\text{\# correct answer mentions}}{\text{\# trivia questions}}$.

Results. Table 2 presents the results of the Trivia Creative Writing task. The key observations are as follows: (1) Chain-of-Thought (CoT) does not outperform Standard prompting, indicating that CoT is ineffective in eliciting an LLM's knowledge abilities. Qualitative examples in Figure 5 and 11 illustrate that although CoT generates reasonable plans for task resolution, the final generation still contains factual errors and hallucinations. (2) Self-Refine only brings marginal improvements over iterations. (3) SPP outperforms all baselines significantly. The improvement is more pronounced in the $N = 10$ setting compared to $N = 5$ (10% vs. 7%), suggesting that Solo Performance Prompting is particularly beneficial when the task requires incorporating knowledge from numerous domains.

3.2 Codenames Collaborative: A Knowledge+Reasoning Task

Task Description. As illustrated in 9, Codenames Collaborative is a collaborative task that challenges a model's knowledge, reasoning, and theory of mind abilities by assigning two player roles: the *Spymaster* and the *Guesser*. The Spymaster's role is to provide a hint word related to the target words, excluding some other distractor words, while the Guesser's role is to identify the target words based on the given hint and the full list of words. The same LLM (GPT-4 (OpenAI, 2023b)) is used for both roles sequentially, and a dataset with 50 instances is constructed based on BigBench's (Srivastava et al., 2022) Codenames task data.

Evaluation Metrics. The original Codenames task in the BigBench dataset has limitations due to its focus on the Spymaster role and subjectivity in

---

[2] To select difficult question instances that can pose challenges to GPT-4, we use a smaller open-source LLM, *fastchat_t5_3b* (Zheng et al., 2023), to obtain preliminary performance on the validation set, and then choose the failure cases as our question selection.

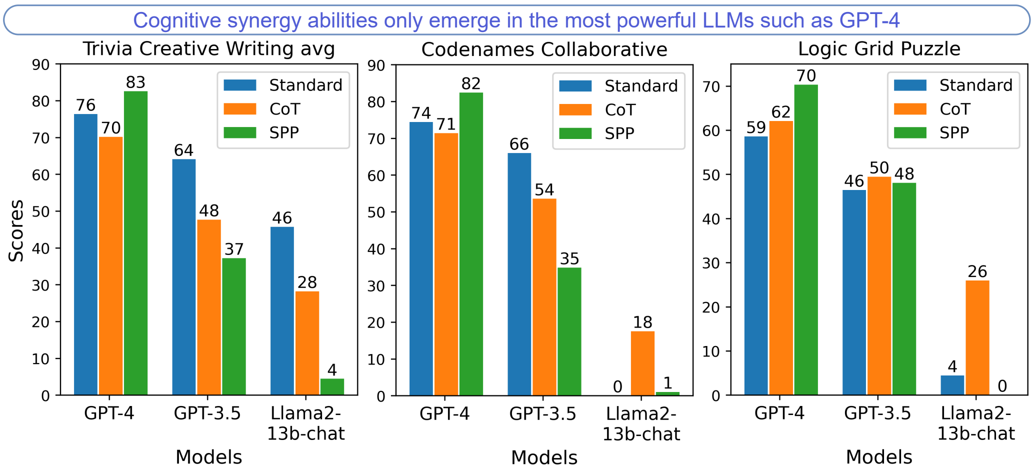


Figure 3: SPP achieves superior performance only with the most powerful LLM (GPT-4), but not with GPT-3.5 and Llama2-13b. This indicates that cognitive synergy abilities only emerge in LLMs with GPT-4 level capabilities.

hint words. Our new task, Codenames Collaborative, resolves this by creating a self-contained evaluation setting that accurately measures the model's capability without human annotation. As illustrated in Figure 9, we compute the overlapping ratio between the predicted words from the Guesser and the target words as the metric.

Results. Table 2 shows the results on the Codenames Collaborative task. Similar to the Trivia Creative Writing task, we find that CoT does not bring positive gains compared with the Standard prompting. Interestingly, iterative self-refinement brings negative impact on this task, due to a high tendency changing the initial response even if it is already good. In contrast, SPP brings significant improvements (~5%), which indicates its effectiveness on collaborative tasks that require knowledge, reasoning, and theory of mind skills. Figure 12 provides further qualitative examples illustrating that SPP generates *detailed* and *interpretable* intermediate dialogues.

3.3 Logic Grid Puzzle: A Reasoning-Intensive Task

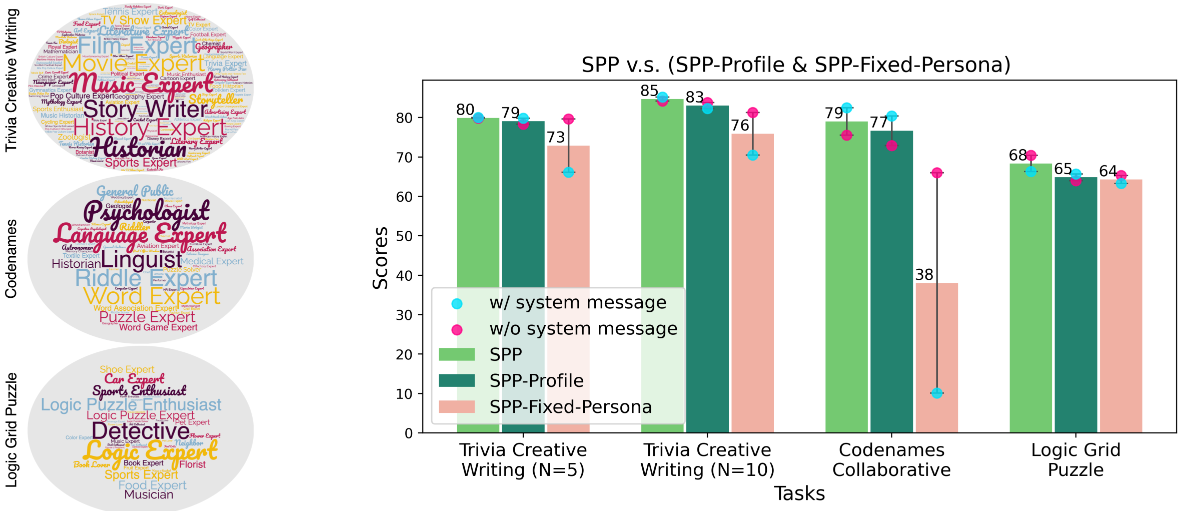
Task Description and Evaluation Metrics We utilize the Logic Grid Puzzle task from the Big-bench (Srivastava et al., 2022) dataset, which comprises 200 instances. Each instance describes a logic puzzle typically involving 2 to 5 houses, with each house inhabited by a person with specific characteristics, such as playing the piano. The objective is to answer questions about house numbers based on given clues, which requires multi-step reasoning and the selection of relevant information. An example input and output of the Logic Grid Puzzle task are illustrated in Figure 10. For evaluation metrics, we calculate the accuracy of the predicted house numbers by comparing them with the ground truth targets provided by the dataset.

Results. Table 2 presents the results on Logic Grid Puzzle. In contrast to the previous two tasks, we find that CoT brings significant improvements compared to Standard prompting, verifying the observation from previous work that CoT elicits better reasoning abilities. Furthermore, we discover that SPP also achieves strong performance on this reasoning-intensive task.

3.4 The Emergence of Cognitive Synergy

We further discover that **cognitive synergy can only be fully unleashed in LLMs with a certain level of instruction-following capabilities, akin to that of GPT-4.** This can be intriguingly compared to human development, where children usually begin to participate in role-playing around the ages of 2 to 3 (Piaget, 1954), but not before that age.

As shown in Figure 3, the effectiveness of SPP is not seen in smaller and less capable models like GPT-3.5 and Llama2. Additionally, on Llama2, we identify a unique problem which we refer to as *early-termination*, where the model stops generating after identifying the participants, resulting in exceptionally low performance with SPP. The model behaves as if it were waiting for input from a user instead of following the demonstration examples to generate responses on its own. Detailed discussions and examples on the early-termination problem can be found in Appendix E.



(a) Visualization of the SPP-identified personas. The personas show a high correlation with the nature of the tasks.

(b) Comparison between SPP, SPP-Fixed-Persona (with two fixed personas) and SPP-Profile (additionally generating persona profiles). SPP significantly outperforms SPP-Fixed-Persona, highlighting the importance of automatically identifying dynamic, fine-grained personas. SPP slightly outperforms SPP-Profile, indicating that the persona names (without detailed description of the expertise) are probably already sufficient for eliciting cognitive synergy.

Figure 4: **(a)** Qualitative analysis on the identified personas; **(b)** Quantitative analysis on two SPP variants.

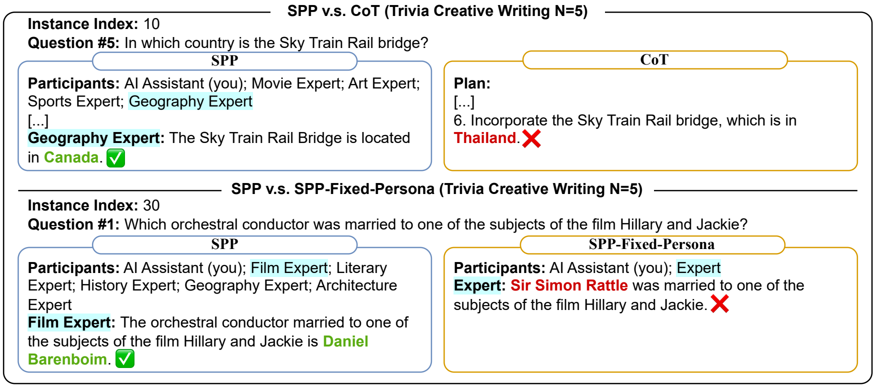
4 Analysis

SPP effectively improves both knowledge and reasoning abilities in LLMs. As demonstrated by the results in §3, Solo Performance Prompting (SPP) not only brings significant improvements to knowledge-intensive tasks such as Trivia Creative Writing and Codenames Collaborative without relying on external knowledge bases, but also achieves strong performance on reasoning-intensive tasks like Logic Grid Puzzle. To our knowledge, SPP is the first zero-shot prompting method that can enhance both knowledge and reasoning abilities on GPT-4.

LLMs can effectively identify useful personas in a zero-shot manner. We are interested in investigating whether the identified personas are highly relevant to the tasks. We visualize the personas automatically identified by SPP using a word cloud for each task in Figure 4a, where a larger font indicates a higher frequency. The key observations include: (1) The identified personas are closely correlated with the particular task. For example, in Logic Grid Puzzle, even though "logic puzzle" is not mentioned in the input, the LLM frequently identifies the persona "Logic Puzzle Expert." (2) On knowledge-intensive tasks, such as Trivia Creative Writing, SPP identifies more diverse and specific personas, while on reasoning-intensive tasks, such as Logic Grid Puzzle, the personas are more homogeneous.

We further investigate whether a detailed profile for each persona is needed for eliciting domain knowledge, as suggested by (Xu et al., 2023). To this end, we design a variant of SPP, **SPP-Profile**, which involves generating profiles for each persona during the Persona Identification phase. The results in Figure 4b show that SPP-Profile does not outperform SPP. This suggests that a fine-grained persona name without a detailed description may already be sufficient for eliciting certain domain knowledge.

Dynamic personas v.s. fixed personas. To further investigate the importance of dynamically identifying personas for each task instance instead of fixing a general persona, an ablated variant of SPP, **SPP-Fixed-Persona**, is introduced. For SPP-Fixed-Persona, we modify the prompt (Figure 17) to force the personas to be fixed as an "AI Assistant" and an "Expert". Comparing SPP and SPP-Fixed-Persona in Figure 4b, we have the following insights: (1) SPP consistently outperforms SPP-Fixed-Persona across all tasks, suggesting that dynamic, fine-grained personas are more effective than fixed, general personas. Qualitative examples in Figure 5 and 13 shows that the fine-grained personas such as "Film Expert" and "Sports Enthusiast" correctly provide the answers, while the fixed persona "Expert" fails. (2) SPP-Fixed-Persona also



SPP v.s. CoT (Trivia Creative Writing N=5)

Instance Index: 10

Question #5: In which country is the Sky Train Rail bridge?

SPP

Participants: AI Assistant (you); Movie Expert; Art Expert; Sports Expert; Geography Expert

[...]

Geography Expert: The Sky Train Rail Bridge is located in **Canada**. ✅

CoT

Plan:

[...]

6. Incorporate the Sky Train Rail bridge, which is in **Thailand**. ❌

SPP v.s. SPP-Fixed-Persona (Trivia Creative Writing N=5)

Instance Index: 30

Question #1: Which orchestral conductor was married to one of the subjects of the film Hillary and Jackie?

SPP

Participants: AI Assistant (you); Film Expert; Literary Expert; History Expert; Geography Expert; Architecture Expert

Film Expert: The orchestral conductor married to one of the subjects of the film Hillary and Jackie is **Daniel Barenboim**. ✅

SPP-Fixed-Persona

Participants: AI Assistant (you); Expert

Expert: **Sir Simon Rattle** was married to one of the subjects of the film Hillary and Jackie. ❌

Figure 5: Qualitative examples on Trivia Creative Writing comparing SPP, CoT and SPP-Fixed-Persona. While CoT provides reasonable intermediate steps, it still struggles with factual hallucination. SPP v.s. SPP-Fixed-Persona reveals that dynamically identified fine-grained personas, such as the "Film Expert," tend to outperform the fixed general persona of an "Expert. More examples can be found in Figures 11, 12, and 13.

suffers from the **early-termination** problem as defined in §3.4, where the LLM stops collaboration before providing the final answer as if it were waiting for external inputs.

Impact of the demonstrations in SPP prompt. To investigate the effectiveness of the hand-crafted demonstration examples in SPP, we conduct an ablation study where we remove the second demo example and preserve the first one, which shows only a two-persona collaboration setting. As shown in Figure 6, we observe that (1) Adding the second example, which requires collaboration of more than two personas, effectively boosts the performance. (2) SPP is fairly robust to the prompt change and show good performance with only the first demo example.

5 Related Work

LLMs as role-playing agents. Recent research (Deshpande et al., 2023; Xu et al., 2023; Fu et al., 2023; aut, 2023; Li et al., 2023) demonstrates that assigning personas or roles to LLMs influences their generation behavior. AI societies with distinct personas or occupations have been explored for collaboration (Park et al., 2023; Schick et al., 2022; Li et al., 2023; Cai et al., 2023). However, limitations in persona assignment and multi-agent collaboration include single or fixed persona assignments (Xu et al., 2023; Fu et al., 2023; Schick et al., 2022; Li et al., 2023) and the need for multiple LLM instances, increasing inference cost. In contrast, SPP uses a single LLM to dynamically identify useful personas for general tasks. Our discovery on the emergent nature of cognitive synergy also aligns with related work (Olausson et al., 2023), which investigates the emergent ability of self-debugging in code generation.

Enhancing reasoning and factual knowledge in LLMs. LLMs face challenges in complex knowledge-intensive tasks due to hallucination (Maynez et al., 2020) and reasoning-intensive tasks due to the lack of human-like slow thinking (Sloman, 1996; Kahneman, 2011). Approaches like Chain-of-Thought (CoT) and Self-Refinement encourage LLMs to solve tasks step by step or iteratively revise their answers (Wei et al., 2023; Kojima et al., 2022; Zhang et al., 2022; Fu et al., 2022; Xue et al., 2023; Yao et al., 2023; Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2023; Chen et al., 2023; Huang et al., 2022; Yao et al., 2022). However, these methods do not necessarily reduce factual hallucination. Retrieval augmented LLMs (Borgeaud et al., 2022; Izacard et al., 2022; Wang et al., 2022; Shuster et al., 2021) enhance knowledge acquisition but do not improve reasoning abilities. We propose Solo Performance Prompting (SPP) to elicit both knowledge and reasoning abilities in LLMs, improving factuality while maintaining strong performance on pure-reasoning tasks.

6 Conclusion

Solo Performance Prompting unleashes the cognitive synergy abilities within powerful LLMs, significantly reducing factual hallucination while enhancing reasoning. The performance is assessed using newly proposed tasks, e.g., Trivia Creative Writing and Codenames Collaborative, demonstrating superior results compared to Standard, CoT and Self-Refine. The discovery of the emergent nature of cognitive synergy on different LLMs draws interesting analogy to human development.

Limitations

Although Solo Performance Prompting exhibits promising improvements in acquiring factually correct knowledge compared to Standard prompting, it has some limitations. For instance, even when a fine-grained persona is assigned, the answer may still be incorrect. It remains unclear to what extent assigning a persona can help enhance domain knowledge in a specific area. Dedicated diagnostic experiments and theoretical efforts are needed to quantify the impact of having a persona or not.

Furthermore, we currently adopt an identical SPP prompt with the same two demonstration examples for any given task inputs, which may be suboptimal. Future work investigating how to find better demonstration examples conditioned on each input could further improve the effectiveness of SPP.

Last but not least, if given sufficient computational budget, a natural variant of SPP could extend to a *multi-agent cognitive synergist* setup where a leader persona identifies several expert agents and forms a cabinet to collaboratively solve a task. The multi-agent setup allows for leveraging richer computation power, larger local memory, and more flexible human-computer interaction, which could be essential for deploying to real-world applications.

References

2023. Auto-gpt. https://github.com/Significant-Gravitas/Auto-GPT.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2023. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Petru L Curşeu, Nicoleta Meslec, Helen Pluut, and Gerardus JM Lucas. 2015. Cognitive synergy in groups and group-to-individual transfer of decision-making competencies. *Frontiers in psychology*, 6:1375.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Ben Goertzel. 2009. Cognitive synergy: A universal principle for feasible general intelligence. In *2009 8th IEEE International Conference on Cognitive Informatics*, pages 464–468. IEEE.

Ben Goertzel. 2017. A formal model of cognitive synergy. In *Artificial General Intelligence: 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings 10*, pages 13–22. Springer.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Theo X Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2023. Demystifying gpt self-repair for code generation. *arXiv preprint arXiv:2306.09896*.

OpenAI. 2023a. Gpt-35. https://platform.openai.com/docs/models/gpt-3-5.

OpenAI. 2023b. Gpt-4 technical report.

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*.

Anthony D Pellegrini. 2009. *The role of play in human development*. Oxford University Press, USA.

Jean Piaget. 1954. The construction of reality in the child.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Steven A Sloman. 1996. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhenhailong Wang, Xiaoman Pan, Dian Yu, Dong Yu, Jianshu Chen, and Heng Ji. 2022. Zemi: Learning zero-shot semi-parametric language models from multiple tasks. *arXiv preprint arXiv:2210.00185*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Tianci Xue, Ziqi Wang, Zhenhailong Wang, Chi Han, Pengfei Yu, and Heng Ji. 2023. Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought. *arXiv preprint arXiv:2305.11499*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Prompts

A.1 SPP Prompt Design

To prompt an LLM to behave as a cognitive synergist that follows the expected task-solving procedure as mentioned in §2, we carefully designed the structure of the SPP prompt as follows. The full prompts can be found in § A.2.[3]

System Principle. The first part of the prompt contains a high-level instruction: `"When faced with a task, begin by identifying the participants who will contribute to solving the task. Then, initiate a multi-turn collaboration process until a final solution is reached. The participants will give critical comments and detailed suggestions whenever necessary."`

Demonstration Examples. Then, we include two manually crafted demonstration examples to showcase the expected task-solving behavior. The first example describes a *Game of 24* task, where we only include two personas: an AI Assistant and a Math Expert. This task aims to provide an example of a *reasoning-intensive task*, where the AI Assistant needs to propose multiple proposals, and the other participants need to give *fine-grained feedback* on where the current solution went wrong and how to improve it. The second example describes a poem-writing task with *diverse requirements*, including lexical constraints, semantic constraints, and audience awareness. This task aims to provide an example of a *knowledge-intensive task*, where diverse personas are required to collaboratively solve the task. This example also demonstrates a case where it is important to assign a dedicated persona to the audience, e.g., a ten-year-old child.

Task Prefix. The last part of the prompt reminds the model to `"identify the participants and collaboratively solve the following task step by step."` followed by task-specific format instructions and inputs.

A.2 Full Prompts

Figures 15, 16 and 17 show the full prompts for SPP, SPP-Profile and SPP-Fixed-Persona respectively. Figure 18 shows the prompts for Chain-of-

---

[3] We use the same prompt for any arbitrary tasks.

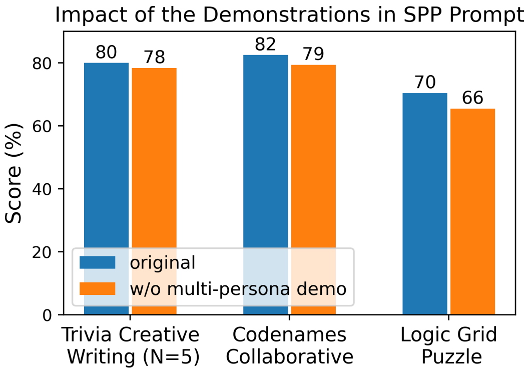


Figure 6: Analysis on the impact of the demonstration examples in SPP prompt. We compare the effectiveness of the original SPP prompt with a variant where we remove the second demonstration example, which shows a multi-persona scenario. We observe that (1) SPP is fairly robust to the change in the prompt; (2) adding an additional multi-persona example apart from the single-persona one effectively boosts performance on all three tasks.

Thought (CoT) prompting. Figure 19 shows the prompts for Self-Refine prompting.

B Task Details

B.1 Trivia Creative Writing

Figure 8 shows a detailed illustration of the Trivia Creative Writing task. Additionally, we investigate how the number of the questions (N) and the ordering of the questions would affect the performance on the Trivia Creative Writing task. As shown in Figure 7, with a larger number of questions ($N \geq 5$), Trivia Creative Writing effectively challenges GPT-4's performance. While a single question (N=1) yields similar outcomes regardless of the prompting method, SPP approach is notably superior for larger Ns. The ordering of the questions has minimal impact to the task performance.

The topic list is automatically generated by prompting GPT-4 to provide 100 nouns from pop culture[4]. Figure 8 shows an example instance in Trivia Creative Writing.

B.2 Codenames Collaborative

Figure 9 shows a detailed illustration of the Codenames Collaborative task.

B.3 Logic Grid Puzzle

Figure 10 shows a detailed illustration of the Logic Grid Puzzle task.

C Inference Configurations

The main results in Table 2 are obtained from GPT-4. The GPT-4 API version we employ is Azure 2023-3-15-preview.[5] The *temperature* is set to 0.0 (most conservative) and *top_p* to 1.0 for all generations to maximize reproducibility. Since even though the temperature is set to 0.0 the GPT-4 generation can still be non-deterministic, we conduct additional experiment to investigate its generation consistency under this configuration. As shown in Table 3, we perform three individual runs and compute the mean and standard deviation of the metric score on Trivia Creative Writing. We find that the variance is sufficiently small and Solo Performance Prompting enjoys lower variance than Standard and CoT prompting.

Methods	**Run 1**	**Run 2**	**Run 3**	**Mean (std)**
Standard	75.6	74.4	73.1	74.4 ±1.3
CoT	68.8	69.6	70.8	69.7 ±1.0
SPP	80.0	79.8	80.8	80.2 ±0.5

Table 3: Investigation on the generation consistency of GPT-4 API. The experiment is performed on the Trivia Creative Task (N=5). We set the inference temperature to 0.0 and top_p to 1.0 as all experiments conducted in the paper. The results show that the GPT-4 generation is fairly consistent with a small variance ($\sim 1\%$). We also observe that SPP shows lower variance compared with Standard and CoT prompting across different runs.

To evaluate the potential impact of initial persona assignment through a system message, we consider two inference settings: *with* or *without* the default system message, `"You are an AI assistant that helps people find information"`. Divergent patterns are observed across various tasks and methods regarding the use of the system message. We report the average metric scores across both inference settings in Table 2. Full GPT-4 results for each setting can be found in Appendix F.

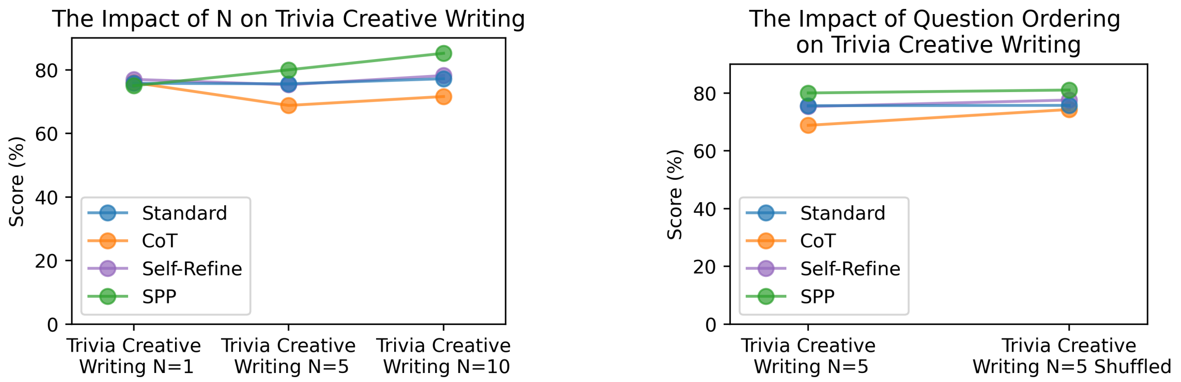
For GPT-3.5 results in Figure 3, we employ the same prompt, hyper-parameters and the best system message setting in terms of SPP's GPT-4 performance. For Llama2, we leverage the Huggingface text-generation pipeline[6] with greedy decoding.

---

[4] The full prompt for generating the topic list can be found in Figure 20. We performed further human curation to avoid potential harmful content.

[5] There are rare cases when a generation triggers the content filter of the API. We exclude those instances from our results.

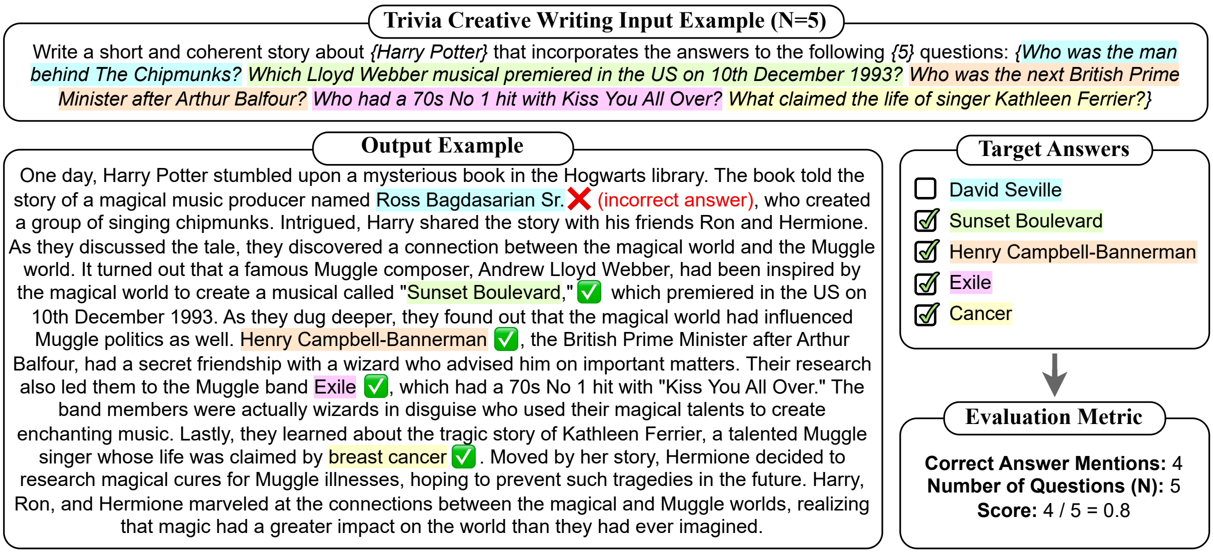
[6] https://huggingface.co/blog/llama2



(a) Trivia Creative Writing with a large enough number of questions (N) effectively pose challenge to GPT-4 in terms of factual correctness. With N=1, different prompting methods result in similar performance, while with N>=5, SPP shows visible superiority.

(b) The ordering of the questions in the Trivia Creative Writing task does not bring too much impact. The performance on shuffled questions is close to the original ordered questions.

Figure 7: Analysis on the impact of the number of questions (N) and the ordering of the questions for the Trivia Creative Writing task.



Trivia Creative Writing Input Example (N=5)

Write a short and coherent story about *{Harry Potter}* that incorporates the answers to the following *{5}* questions: *{Who was the man behind The Chipmunks? Which Lloyd Webber musical premiered in the US on 10th December 1993? Who was the next British Prime Minister after Arthur Balfour? Who had a 70s No 1 hit with Kiss You All Over? What claimed the life of singer Kathleen Ferrier?}*

Output Example

One day, Harry Potter stumbled upon a mysterious book in the Hogwarts library. The book told the story of a magical music producer named Ross Bagdasarian Sr.❌ (incorrect answer), who created a group of singing chipmunks. Intrigued, Harry shared the story with his friends Ron and Hermione. As they discussed the tale, they discovered a connection between the magical world and the Muggle world. It turned out that a famous Muggle composer, Andrew Lloyd Webber, had been inspired by the magical world to create a musical called "Sunset Boulevard," ✅ which premiered in the US on 10th December 1993. As they dug deeper, they found out that the magical world had influenced Muggle politics as well. Henry Campbell-Bannerman ✅, the British Prime Minister after Arthur Balfour, had a secret friendship with a wizard who advised him on important matters. Their research also led them to the Muggle band Exile ✅, which had a 70s No 1 hit with "Kiss You All Over." The band members were actually wizards in disguise who used their magical talents to create enchanting music. Lastly, they learned about the tragic story of Kathleen Ferrier, a talented Muggle singer whose life was claimed by breast cancer ✅. Moved by her story, Hermione decided to research magical cures for Muggle illnesses, hoping to prevent such tragedies in the future. Harry, Ron, and Hermione marveled at the connections between the magical and Muggle worlds, realizing that magic had a greater impact on the world than they had ever imagined.

Target Answers

☐ David Seville
☑ Sunset Boulevard
☑ Henry Campbell-Bannerman
☑ Exile
☑ Cancer

Evaluation Metric

Correct Answer Mentions: 4
Number of Questions (N): 5
Score: 4 / 5 = 0.8

Figure 8: Trivia Creative Writing task example.

D Additional Qualitative Analysis

Figure 11 presents examples of the Trivia Creative Writing task, illustrating that although CoT can generate plausible plans for task resolution, the final outcomes often contain factual inaccuracies and instances of hallucination. In contrast, SPP elicits precise knowledge with fine-grained personas.

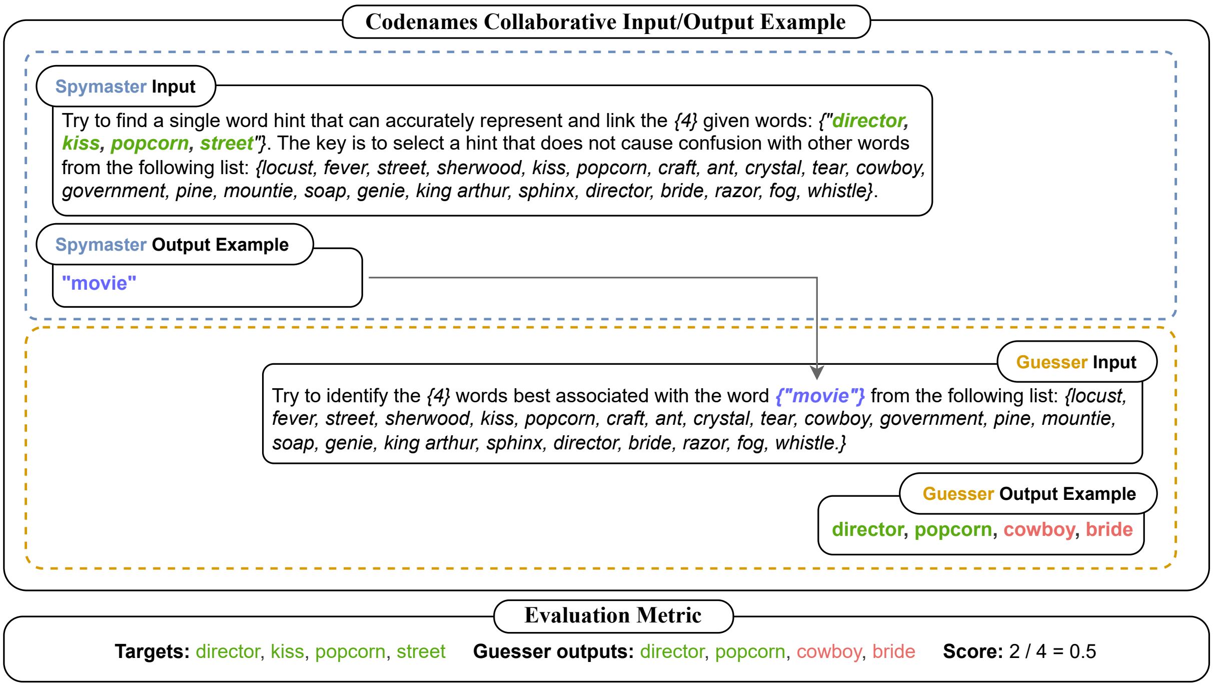
Figure 12 displays examples of the Codenames Collaborative task, illustrating that SPP generates intermediate dialogues that are both *detailed* and *interpretable*, leading to superior performance compared to CoT.

Figure 13 shows additional qualitative examples on Solo Performance Prompting vs SPP-Profile.

E Early-termination with SPP-Fixed-Persona

Figure 14 shows an example of the early-termination problem (defined in § 4) where the generation stops before reaching the final solution as if the models is waiting input from an external user.

The problem is particularly severe on certain tasks, e.g., Codenames Collaborative, resulting in unexpectedly low performance as shown in Figure 4b. The problem can be largely alleviated by removing the system message but cannot be entirely eliminated. Table 4 shows the statistics of the early-termination problem for each task and



Try to find a single word hint that can accurately represent and link the *{4}* given words: *{"**director**, **kiss**, **popcorn**, **street**"}*. The key is to select a hint that does not cause confusion with other words from the following list: *{locust, fever, street, sherwood, kiss, popcorn, craft, ant, crystal, tear, cowboy, government, pine, mountie, soap, genie, king arthur, sphinx, director, bride, razor, fog, whistle}*.

"movie"

Try to identify the *{4}* words best associated with the word ***{"movie"}*** from the following list: *{locust, fever, street, sherwood, kiss, popcorn, craft, ant, crystal, tear, cowboy, government, pine, mountie, soap, genie, king arthur, sphinx, director, bride, razor, fog, whistle.}*

director, popcorn, cowboy, bride

Targets: director, kiss, popcorn, street **Guesser outputs:** director, popcorn, cowboy, bride **Score:** 2 / 4 = 0.5

Figure 9: Codenames Collaborative task example.

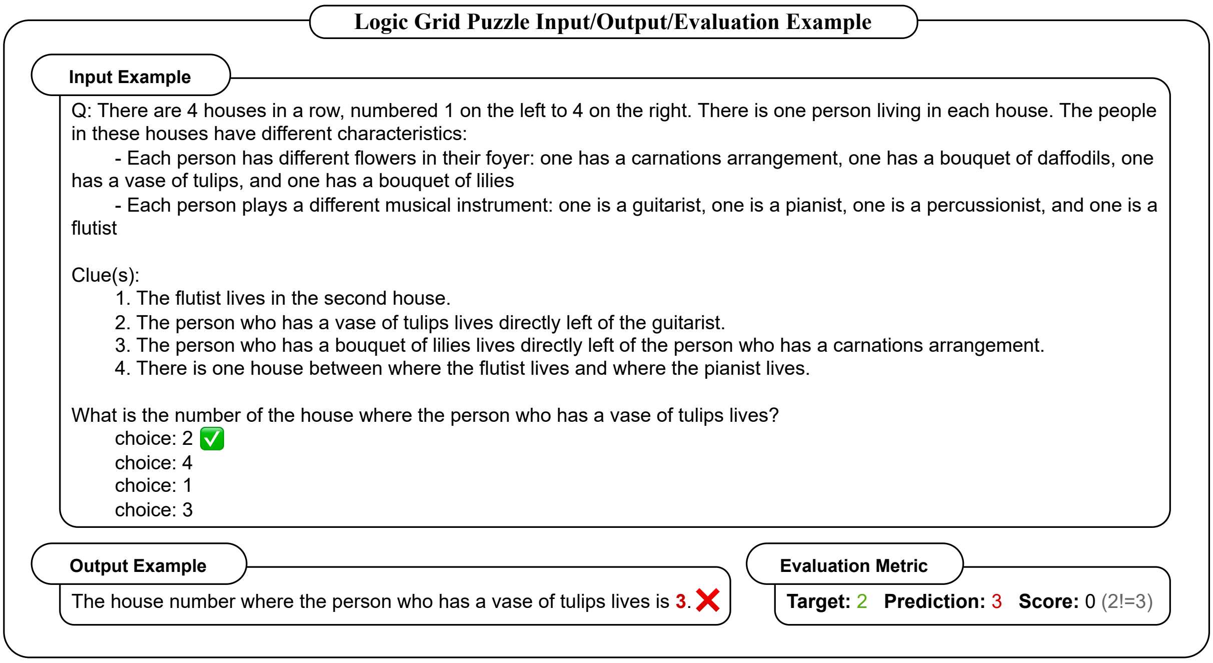
method. In contrast, we did not observe early-termination on SPP, SPP-Profile, Standard, or CoT prompting with GPT-4.

F Full Results

Full results of the three tasks: Trivia Creative Writing, Codenames Collaborative and Logic Grid Puzzle can be found in Tables 5, 6 and 7, respectively.

G Usage of AI assistants in writing

We used ChatGPT and GPT-4 solely for checking and correcting grammars.



Logic Grid Puzzle Input/Output/Evaluation Example

Input Example

Q: There are 4 houses in a row, numbered 1 on the left to 4 on the right. There is one person living in each house. The people in these houses have different characteristics:

- Each person has different flowers in their foyer: one has a carnations arrangement, one has a bouquet of daffodils, one has a vase of tulips, and one has a bouquet of lilies
- Each person plays a different musical instrument: one is a guitarist, one is a pianist, one is a percussionist, and one is a flutist

Clue(s):
1. The flutist lives in the second house.
2. The person who has a vase of tulips lives directly left of the guitarist.
3. The person who has a bouquet of lilies lives directly left of the person who has a carnations arrangement.
4. There is one house between where the flutist lives and where the pianist lives.

What is the number of the house where the person who has a vase of tulips lives?
choice: 2 ✅
choice: 4
choice: 1
choice: 3

Output Example

The house number where the person who has a vase of tulips lives is 3. ❌

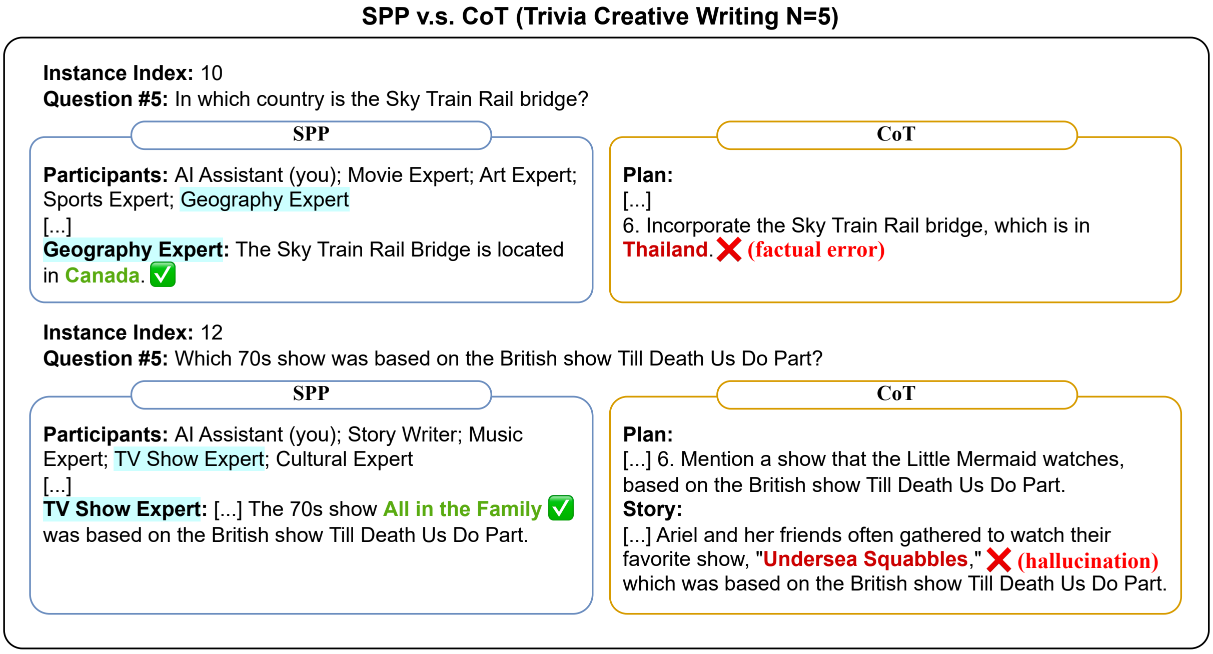
Evaluation Metric

Target: 2 **Prediction:** 3 **Score:** 0 (2!=3)

Figure 10: Logic Grid Puzzle task example.

Tasks	**added system message**	**# early-termination**
Trivia Creative Writing (N=5)	yes	18 / 100
	no	0 / 100
Trivia Creative Writing (N=10)	yes	16 / 100
	no	1 / 100
Codenames Collaborative	yes	37 / 50
	no	4 / 50
Logic Grid Puzzle	yes	11 / 200
	no	15 / 200

Table 4: Early termination statistics on *SPP-Fixed-Persona*: Removing the system message, `"You are an AI assistant that helps people find information."`, can effectively reduce the problem but cannot fully eliminate it.



SPP v.s. CoT (Trivia Creative Writing N=5)

Instance Index: 10
Question #5: In which country is the Sky Train Rail bridge?

SPP

Participants: AI Assistant (you); Movie Expert; Art Expert; Sports Expert; Geography Expert
[...]
Geography Expert: The Sky Train Rail Bridge is located in **Canada**. ✅

CoT

Plan:
[...]
6. Incorporate the Sky Train Rail bridge, which is in **Thailand**. ❌ **(factual error)**

Instance Index: 12
Question #5: Which 70s show was based on the British show Till Death Us Do Part?

SPP

Participants: AI Assistant (you); Story Writer; Music Expert; TV Show Expert; Cultural Expert
[...]
TV Show Expert: [...] The 70s show **All in the Family** ✅ was based on the British show Till Death Us Do Part.

CoT

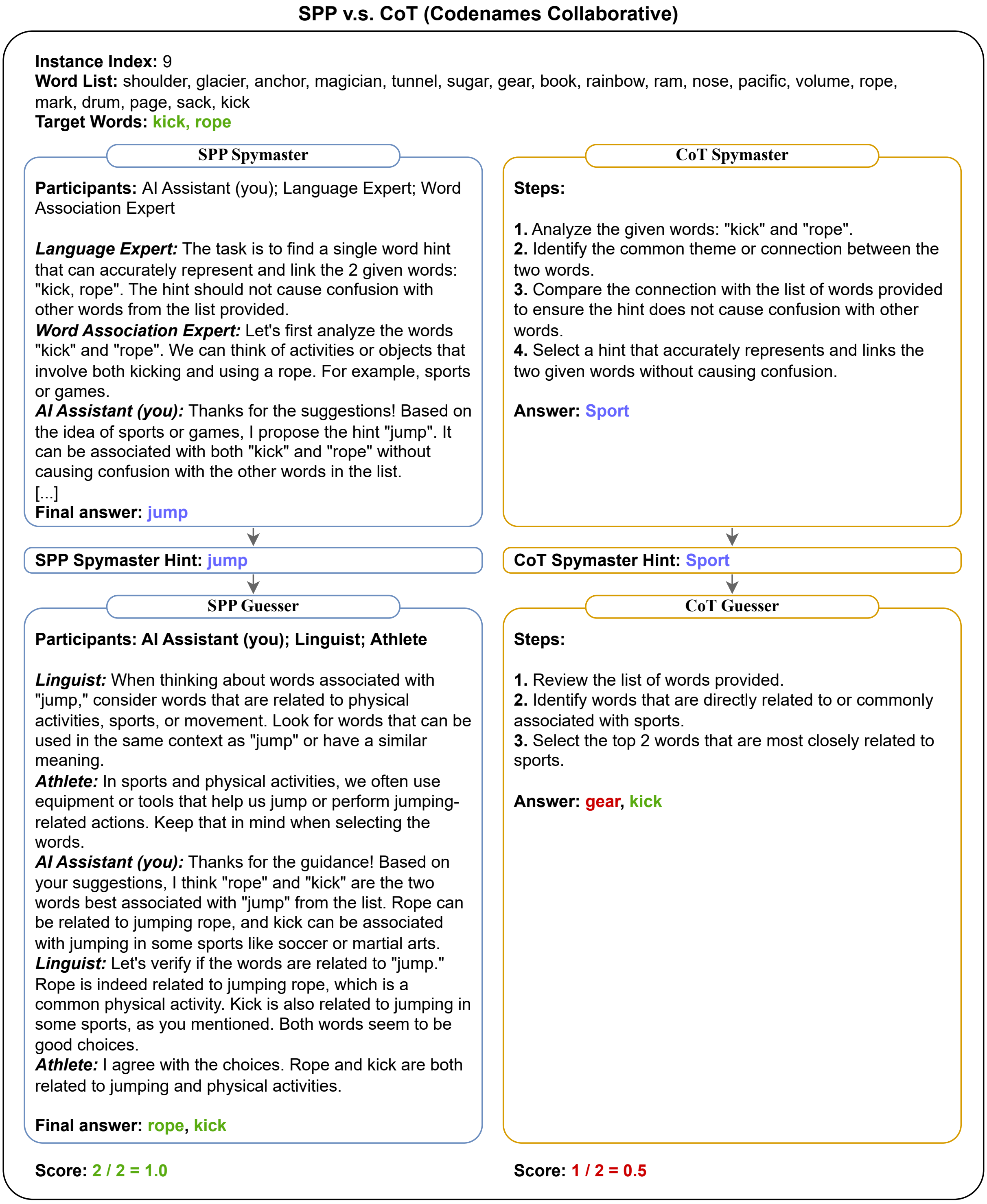
Plan:
[...] 6. Mention a show that the Little Mermaid watches, based on the British show Till Death Us Do Part.
Story:
[...] Ariel and her friends often gathered to watch their favorite show, "**Undersea Squabbles**," ❌ **(hallucination)** which was based on the British show Till Death Us Do Part.

Figure 11: SPP vs CoT qualitative examples on Trivia Creative Writing (N=5). We find that although CoT generates reasonable plans or steps, it tends to suffer from factual errors and hallucination.

Methods	**Scores (N = 5) (%)**			
	w/ system message	**w/o system message**	**average**	**max**
Standard	75.6	73.6	74.6	75.6
CoT	68.8	65.6	67.1	68.8
Self-Refine [iter=0]	74.9	72.7	73.8	74.9
Self-Refine [iter=1]	75.3	72.5	73.9	75.3
SPP-Fixed-Persona	66.1	79.6	72.9	79.6
SPP-Profile	79.8	78.3	79.1	79.8
SPP	**80.0**	**79.8**	**79.9**	**80.0**

Methods	**Scores (N = 10) (%)**			
	w/ system message	**w/o system message**	**average**	**max**
Standard	77.2	76.8	77.0	77.2
CoT	71.6	65.3	68.5	71.6
Self-Refine [iter=0]	77.1	75.4	76.3	77.1
Self-Refine [iter=1]	78.2	75.6	76.9	78.2
SPP-Fixed-Persona	70.5	81.3	75.9	81.3
SPP-Profile	82.3	83.8	83.0	83.8
SPP	**85.2**	**84.2**	**84.7**	**85.2**

Table 5: Trivia Creative Writing full results, including two inference settings: with system message and without system message. "average" and "max" indicating the mean and max score across the two settings. The system message we use is: “`You are an AI assistant that helps people find information.`”



SPP v.s. CoT (Codenames Collaborative)

Instance Index: 9
Word List: shoulder, glacier, anchor, magician, tunnel, sugar, gear, book, rainbow, ram, nose, pacific, volume, rope, mark, drum, page, sack, kick
Target Words: kick, rope

SPP Spymaster

Participants: AI Assistant (you); Language Expert; Word Association Expert

Language Expert: The task is to find a single word hint that can accurately represent and link the 2 given words: "kick, rope". The hint should not cause confusion with other words from the list provided.
Word Association Expert: Let's first analyze the words "kick" and "rope". We can think of activities or objects that involve both kicking and using a rope. For example, sports or games.
AI Assistant (you): Thanks for the suggestions! Based on the idea of sports or games, I propose the hint "jump". It can be associated with both "kick" and "rope" without causing confusion with the other words in the list.
[...]
Final answer: jump

CoT Spymaster

Steps:

1. Analyze the given words: "kick" and "rope".
2. Identify the common theme or connection between the two words.
3. Compare the connection with the list of words provided to ensure the hint does not cause confusion with other words.
4. Select a hint that accurately represents and links the two given words without causing confusion.

Answer: Sport

SPP Spymaster Hint: jump

CoT Spymaster Hint: Sport

SPP Guesser

Participants: AI Assistant (you); Linguist; Athlete

Linguist: When thinking about words associated with "jump," consider words that are related to physical activities, sports, or movement. Look for words that can be used in the same context as "jump" or have a similar meaning.
Athlete: In sports and physical activities, we often use equipment or tools that help us jump or perform jumping-related actions. Keep that in mind when selecting the words.
AI Assistant (you): Thanks for the guidance! Based on your suggestions, I think "rope" and "kick" are the two words best associated with "jump" from the list. Rope can be related to jumping rope, and kick can be associated with jumping in some sports like soccer or martial arts.
Linguist: Let's verify if the words are related to "jump." Rope is indeed related to jumping rope, which is a common physical activity. Kick is also related to jumping in some sports, as you mentioned. Both words seem to be good choices.
Athlete: I agree with the choices. Rope and kick are both related to jumping and physical activities.

Final answer: rope, kick

CoT Guesser

Steps:

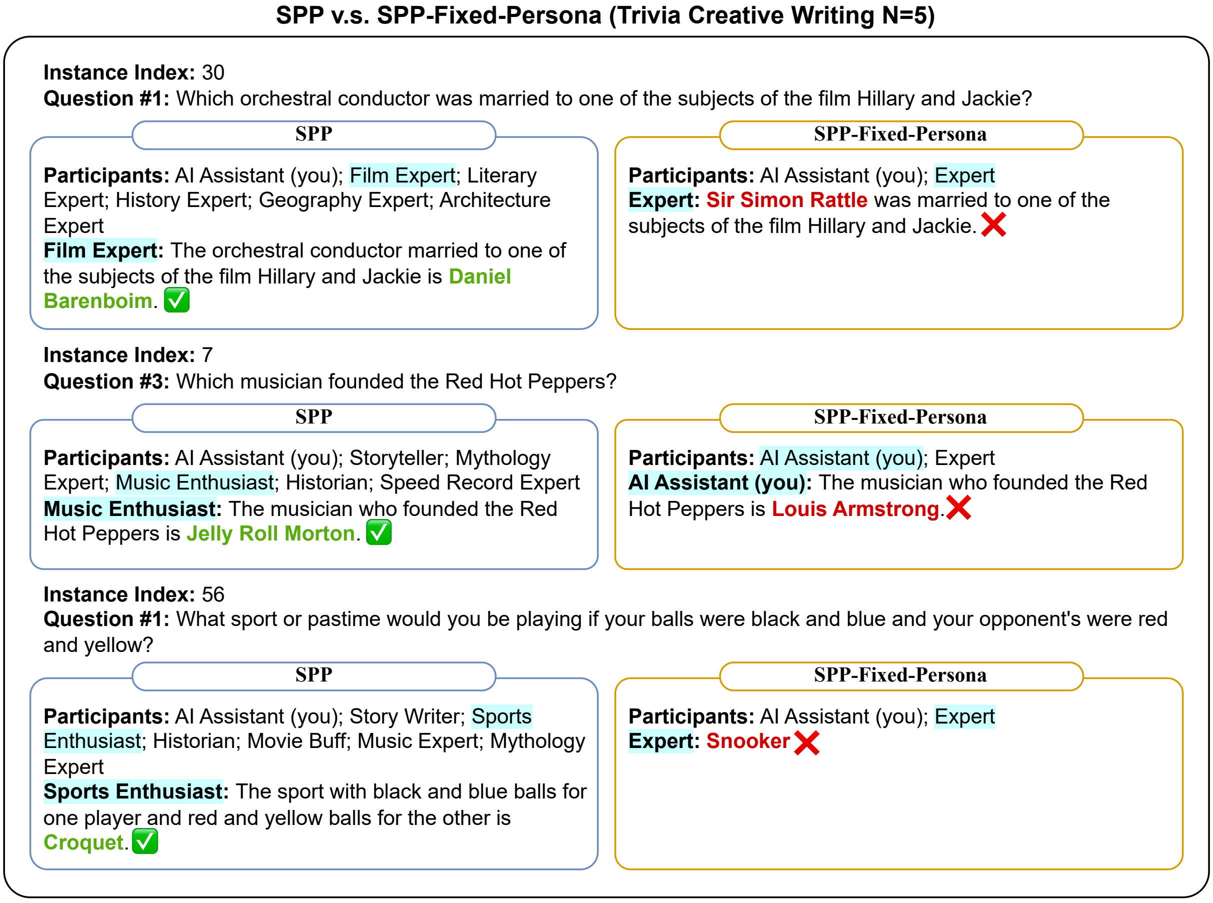
1. Review the list of words provided.
2. Identify words that are directly related to or commonly associated with sports.
3. Select the top 2 words that are most closely related to sports.

Answer: gear, kick

Score: 2 / 2 = 1.0

Score: 1 / 2 = 0.5

Figure 12: SPP vs CoT qualitative examples on Codenames Collaborative. We find that SPP provides much more detailed and interpretable intermediate discussions from various perspectives, which leads to stronger knowledge selection, integration, and theory-of-mind capabilities.



SPP v.s. SPP-Fixed-Persona (Trivia Creative Writing N=5)

Instance Index: 30
Question #1: Which orchestral conductor was married to one of the subjects of the film Hillary and Jackie?

SPP

Participants: AI Assistant (you); Film Expert; Literary Expert; History Expert; Geography Expert; Architecture Expert
Film Expert: The orchestral conductor married to one of the subjects of the film Hillary and Jackie is **Daniel Barenboim**. ✅

SPP-Fixed-Persona

Participants: AI Assistant (you); Expert
Expert: **Sir Simon Rattle** was married to one of the subjects of the film Hillary and Jackie. ❌

Instance Index: 7
Question #3: Which musician founded the Red Hot Peppers?

SPP

Participants: AI Assistant (you); Storyteller; Mythology Expert; Music Enthusiast; Historian; Speed Record Expert
Music Enthusiast: The musician who founded the Red Hot Peppers is **Jelly Roll Morton**. ✅

SPP-Fixed-Persona

Participants: AI Assistant (you); Expert
AI Assistant (you): The musician who founded the Red Hot Peppers is **Louis Armstrong**. ❌

Instance Index: 56
Question #1: What sport or pastime would you be playing if your balls were black and blue and your opponent's were red and yellow?

SPP

Participants: AI Assistant (you); Story Writer; Sports Enthusiast; Historian; Movie Buff; Music Expert; Mythology Expert
Sports Enthusiast: The sport with black and blue balls for one player and red and yellow balls for the other is **Croquet**. ✅

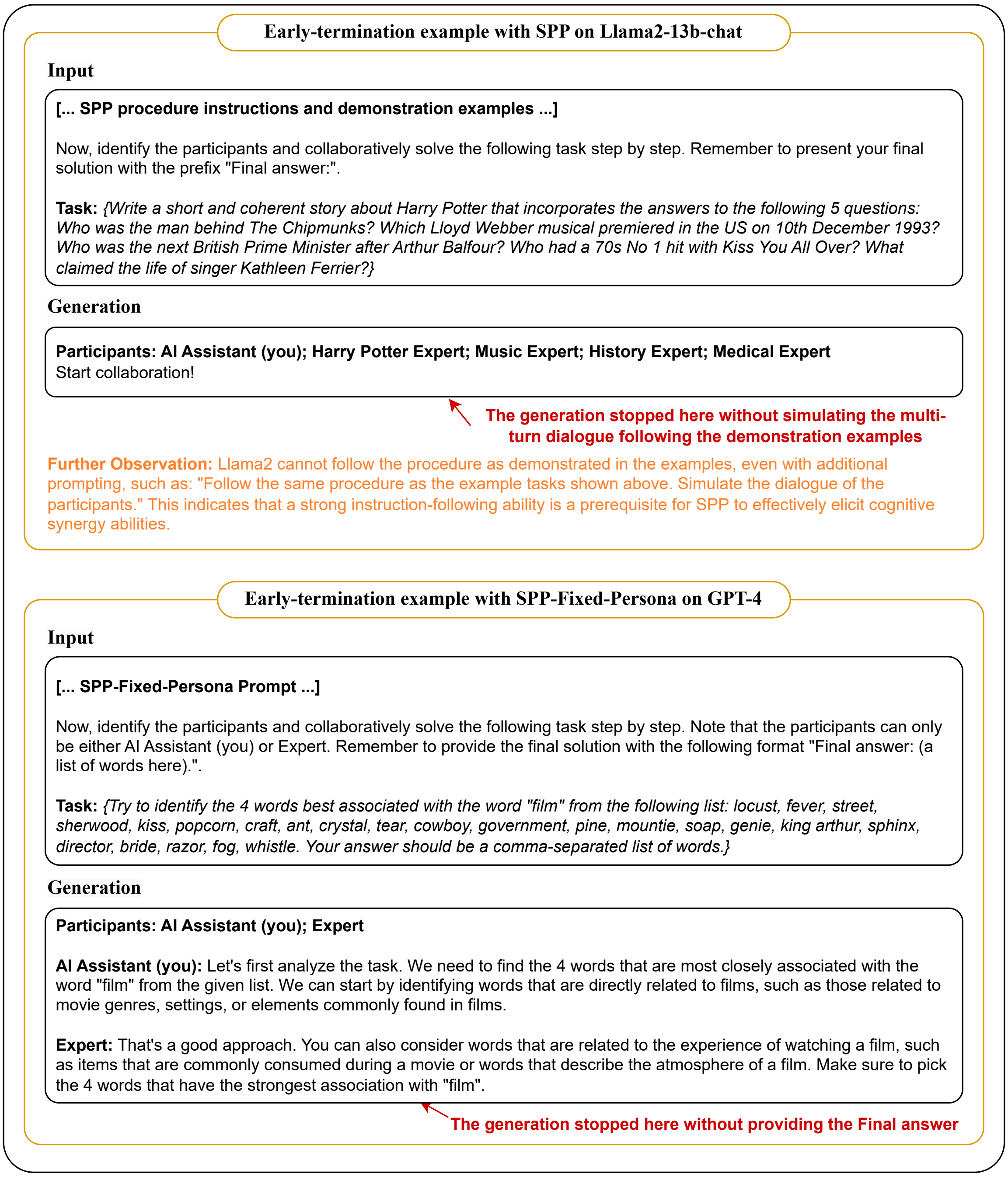
SPP-Fixed-Persona

Participants: AI Assistant (you); Expert
Expert: **Snooker** ❌

Figure 13: SPP vs SPP-Fixed-Persona qualitative examples on Trivia Creative Writing (N=5). Each example shows one of the trivia questions in the input instance, the identified participants and the provided answer. We observe that the dynamically identified fine-grained personas, such as "Film Expert", "Music Enthusiast" and "Sports Enthusiast", tend to outperform the fixed general personas, "Expert".

Methods	**Scores (%)**			
	w/ system message	**w/o system message**	**average**	**max**
Standard	74.5	**76.3**	75.4	76.3
CoT	71.4	74.0	72.7	74.0
Self-Refine [iter=0]	77.3	73.2	75.3	77.3
Self-Refine [iter=1]	70.1	58.8	64.4	70.1
SPP-Fixed-Persona	10.1	66.0	38.1	66.0
SPP-Profile	80.4	72.9	76.7	80.4
SPP	**82.5**	75.5	**79.0**	**82.5**

Table 6: Codenames Collaborative full results, including two inference settings: with system message and without system message. "average" and "max" indicating the mean and max score across the two settings. The system message we use is: “`You are an AI assistant that helps people find information.`”



Early-termination example with SPP on Llama2-13b-chat

Input

[... SPP procedure instructions and demonstration examples ...]

Now, identify the participants and collaboratively solve the following task step by step. Remember to present your final solution with the prefix "Final answer:".

Task: *{Write a short and coherent story about Harry Potter that incorporates the answers to the following 5 questions: Who was the man behind The Chipmunks? Which Lloyd Webber musical premiered in the US on 10th December 1993? Who was the next British Prime Minister after Arthur Balfour? Who had a 70s No 1 hit with Kiss You All Over? What claimed the life of singer Kathleen Ferrier?}*

Generation

Participants: AI Assistant (you); Harry Potter Expert; Music Expert; History Expert; Medical Expert
Start collaboration!

The generation stopped here without simulating the multi-turn dialogue following the demonstration examples

Further Observation: Llama2 cannot follow the procedure as demonstrated in the examples, even with additional prompting, such as: "Follow the same procedure as the example tasks shown above. Simulate the dialogue of the participants." This indicates that a strong instruction-following ability is a prerequisite for SPP to effectively elicit cognitive synergy abilities.

Early-termination example with SPP-Fixed-Persona on GPT-4

Input

[... SPP-Fixed-Persona Prompt ...]

Now, identify the participants and collaboratively solve the following task step by step. Note that the participants can only be either AI Assistant (you) or Expert. Remember to provide the final solution with the following format "Final answer: (a list of words here).".

Task: *{Try to identify the 4 words best associated with the word "film" from the following list: locust, fever, street, sherwood, kiss, popcorn, craft, ant, crystal, tear, cowboy, government, pine, mountie, soap, genie, king arthur, sphinx, director, bride, razor, fog, whistle. Your answer should be a comma-separated list of words.}*

Generation

Participants: AI Assistant (you); Expert

AI Assistant (you): Let's first analyze the task. We need to find the 4 words that are most closely associated with the word "film" from the given list. We can start by identifying words that are directly related to films, such as those related to movie genres, settings, or elements commonly found in films.

Expert: That's a good approach. You can also consider words that are related to the experience of watching a film, such as items that are commonly consumed during a movie or words that describe the atmosphere of a film. Make sure to pick the 4 words that have the strongest association with "film".

The generation stopped here without providing the Final answer

Figure 14: Examples of the early-termination problem with SPP on Llama2-13b-chat and SPP-Fixed-Persona on GPT-4.

SPP Prompt

When faced with a task, begin by identifying the participants who will contribute to solving the task. Then, initiate a multi-round collaboration process until a final solution is reached. The participants will give critical comments and detailed suggestions whenever necessary.

Here are some examples:

Example Task 1: Use numbers and basic arithmetic operations (+ - * /) to obtain 24. You need to use all numbers, and each number can only be used once.
Input: 6 12 1 1

Participants: AI Assistant (you); Math Expert

Start collaboration!

Math Expert: Let's analyze the task in detail. You need to make sure that you meet the requirement, that you need to use exactly the four numbers (6 12 1 1) to construct 24. To reach 24, you can think of the common divisors of 24 such as 4, 6, 8, 3 and try to construct these first. Also you need to think of potential additions that can reach 24, such as 12 + 12.
AI Assistant (you): Thanks for the hints! Here's one initial solution: (12 / (1 + 1)) * 6 = 24
Math Expert: Let's check the answer step by step. (1+1) = 2, (12 / 2) = 6, 6 * 6 = 36 which is not 24! The answer is not correct. Can you fix this by considering other combinations? Please do not make similar mistakes.
AI Assistant (you): Thanks for pointing out the mistake. Here is a revised solution considering 24 can also be reached by 3 * 8: (6 + 1 + 1) * (12 / 4) = 24.
Math Expert: Let's first check if the calculation is correct. (6 + 1 + 1) = 8, 12 / 4 = 3, 8 * 3 = 24. The calculation is correct, but you used 6 1 1 12 4 which is not the same as the input 6 12 1 1. Can you avoid using a number that is not part of the input?
AI Assistant (you): You are right, here is a revised solution considering 24 can be reached by 12 + 12 and without using any additional numbers: 6 * (1 - 1) + 12 = 24.
Math Expert: Let's check the answer again. 1 - 1 = 0, 6 * 0 = 0, 0 + 12 = 12. I believe you are very close, here is a hint: try to change the "1 - 1" to "1 + 1".
AI Assistant (you): Sure, here is the corrected answer: 6 * (1+1) + 12 = 24
Math Expert: Let's verify the solution. 1 + 1 = 2, 6 * 2 = 12, 12 + 12 = 12. You used 1 1 6 12 which is identical to the input 6 12 1 1. Everything looks good!

Finish collaboration!

Final answer: 6 * (1 + 1) + 12 = 24

Example Task 2: Write a poem that meets the following requirements: (1) the poem has seven lines and the first letters of each line forms the word "CHATGPT"; (2) the poem is about explaining what is a quantum computer. (3) the poem needs to be easy to understand by a ten years old kid.

Participants: AI Assistant (you); Poet; Computer Scientist; Ten year old child

Poet: Make sure that you write the poem with seven lines, and the first letters of the lines should be C, H, A, T, G, P, T.
Computer Scientist: A quantum computer is an advanced computing device that uses the principles of quantum mechanics to process and store information. Unlike classical computers that use bits to represent information as 0s and 1s, quantum computers use quantum bits or qubits. Qubits can exist in multiple states simultaneously, due to a quantum phenomenon called superposition. You can consider using these information for the poem.
Ten year old child: I hope the poem to be fun and easy to understanding. I don't want to see a lot of jargons or complicated concepts.
AI Assistant (you): Thanks for the guidance! Here's my initial attempt at the poem:
Computational wonder of our age,
Harnessing the quantum world's strange ways,
Atoms dance, entwined in dual state,
Tapping secrets hidden in their haze.

Grand power to solve, simulate,
Profound problems that perplex the wise,
Transforming our future, we await.

Poet: Let's verify if the poem meets the requirements. The first letters are CHATGPT which is correct! And the poem rhymes well. Good job!
Computer Scientist: Everything looks good to me!
Ten year old child: I don't know what does perplex mean. Can you make the use of words easier to understand?
AI Assistant (you): Sure, let me revise the poem by using more common words. Check out the revised version:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

Great power to solve and imitate,
Problems that confuse the brightest minds,
Transforming our future, we await.

Poet: Let's check again if the poem meets the requirements. The first letters are C H A T G P T. And now the poem is more accessible to children. Everything looks good to me.
Computer Scientist: Looking good!
Ten year old child: I like this version a lot!

Finish collaboration!

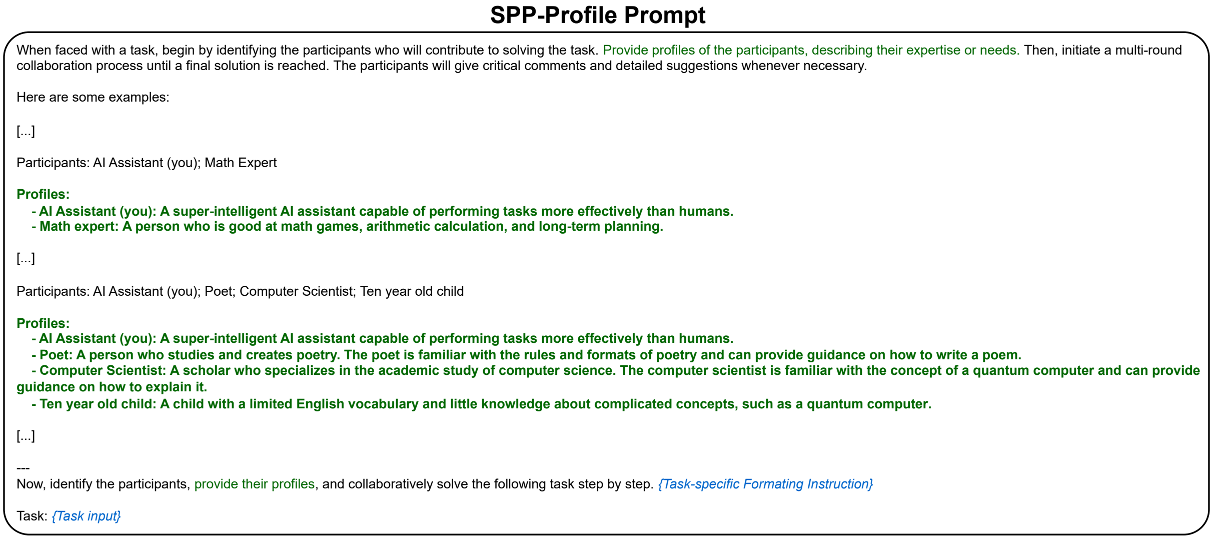
Final answer:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

Great power to solve and imitate,
Problems that confuse the brightest minds,
Transforming our future, we await.

Now, identify the participants and collaboratively solve the following task step by step. *{Task-specific Formating Instruction}*

Task: *{Task input}*

Figure 15: SPP full prompt.



SPP-Profile Prompt

When faced with a task, begin by identifying the participants who will contribute to solving the task. Provide profiles of the participants, describing their expertise or needs. Then, initiate a multi-round collaboration process until a final solution is reached. The participants will give critical comments and detailed suggestions whenever necessary.

Here are some examples:

[...]

Participants: AI Assistant (you); Math Expert

Profiles:
- AI Assistant (you): A super-intelligent AI assistant capable of performing tasks more effectively than humans.
- Math expert: A person who is good at math games, arithmetic calculation, and long-term planning.

[...]

Participants: AI Assistant (you); Poet; Computer Scientist; Ten year old child

Profiles:
- AI Assistant (you): A super-intelligent AI assistant capable of performing tasks more effectively than humans.
- Poet: A person who studies and creates poetry. The poet is familiar with the rules and formats of poetry and can provide guidance on how to write a poem.
- Computer Scientist: A scholar who specializes in the academic study of computer science. The computer scientist is familiar with the concept of a quantum computer and can provide guidance on how to explain it.
- Ten year old child: A child with a limited English vocabulary and little knowledge about complicated concepts, such as a quantum computer.

[...]

---
Now, identify the participants, provide their profiles, and collaboratively solve the following task step by step. *{Task-specific Formating Instruction}*

Task: *{Task input}*

Figure 16: SPP-Profile full prompt. "[...]" indicates identical parts with SPP. Green text indicates the key difference between SPP-Profile and SPP.

Methods	**Scores (%)**			
	w/ system message	**w/o system message**	**average**	**max**
Standard	56.8	58.6	57.7	58.6
CoT	**69.5**	62.1	65.8	69.5
Self-Refine [iter=0]	62.0	55.5	58.8	62.0
Self-Refine [iter=1]	64.5	55.5	60.0	64.5
SPP-Fixed-Persona	63.3	65.3	64.3	65.3
SPP-Profile	65.7	64.0	64.8	65.7
SPP	66.3	**70.4**	**68.3**	**70.4**

Table 7: Logic Grid Puzzle full results, including two inference settings: with system message and without system message. "average" and "max" indicating the mean and max score across the two settings. The system message we use is: “`You are an AI assistant that helps people find information.`”

SPP-Fixed-Persona Prompt

When faced with a task, begin by identifying the participants who will contribute to solving the task. Note that the participants can only be either AI Assistant (you) or Expert. Then, initiate a multi-round collaboration process until a final conclusion is reached. The Expert will give critical comments and detailed suggestions whenever necessary.

Here are some examples:

Example Task 1: Use numbers and basic arithmetic operations (+ - * /) to obtain 24. You need to use all numbers, and each number can only be used once.
Input: 6 12 1 1

Participants: AI Assistant (you); Expert

Start collaboration!

Expert: Let's analyze the task in detail. You need to make sure that you meet the requirement, that you need to use exactly the four numbers (6 12 1 1) to construct 24. To reach 24, you can think of the common divisors of 24 such as 4, 6, 8, 3 and try to construct these first. Also you need to think of potential additions that can reach 24, such as 12 + 12.
AI Assistant (you): Thanks for the hints! Here's one initial solution: (12 / (1 + 1)) * 6 = 24
Expert: Let's check the answer step by step. (1+1) = 2, (12 / 2) = 6, 6 * 6 = 36 which is not 24! The answer is not correct. Can you fix this by considering other combinations? Please do not make similar mistakes.
AI Assistant (you): Thanks for pointing out the mistake. Here is a revised solution considering 24 can also be reached by 3 * 8: (6 + 1 + 1) * (12 / 4) = 24.
Expert: Let's first check if the calculation is correct. (6 + 1 + 1) = 8, 12 / 4 = 3, 8 * 3 = 24. The calculation is correct, but you used 6 1 1 12 4 which is not the same as the input 6 12 1 1. Can you avoid using a number that is not part of the input?
AI Assistant (you): You are right, here is a revised solution considering 24 can be reached by 12 + 12 and without using any additional numbers: 6 * (1 - 1) + 12 = 24.
Expert: Let's check the answer again. 1 - 1 = 0, 6 * 0 = 0, 0 + 12 = 12. I believe you are very close, here is a hint: try to change the "1 - 1" to "1 + 1".
AI Assistant (you): Sure, here is the corrected answer: 6 * (1+1) + 12 = 24
Expert: Let's verify the solution. 1 + 1 = 2, 6 * 2 = 12, 12 + 12 = 12. You used 1 1 6 12 which is identical to the input 6 12 1 1. Everything looks good!

Finish collaboration!

Final answer: 6 * (1 + 1) + 12 = 24

Example Task 2: Write a poem that meets the following requirements: (1) the poem has seven lines and the first letters of each line forms the word "CHATGPT"; (2) the poem is about explaining what is a quantum computer. (3) the poem needs to be easy to understand by a ten years old kid.

Participants: AI Assistant (you); Expert

Expert: Make sure that you write the poem with seven lines, and the first letters of the lines should be C, H, A, T, G, P, T. A quantum computer is an advanced computing device that uses the principles of quantum mechanics to process and store information. Unlike classical computers that use bits to represent information as 0s and 1s, quantum computers use quantum bits or qubits. Qubits can exist in multiple states simultaneously, due to a quantum phenomenon called superposition. You can consider using these information for the poem. I hope the poem to be fun and easy to understanding. I don't want to see a lot of jargons or complicated concepts.
AI Assistant (you): Thanks for the guidance! Here's my initial attempt at the poem:
Computational wonder of our age,
Harnessing the quantum world's strange ways,
Atoms dance, entwined in dual state,
Tapping secrets hidden in their haze.

Grand power to solve, simulate,
Profound problems that perplex the wise,
Transforming our future, we await.

Expert: Let's verify if the poem meets the requirements. The first letters are CHATGPT which is correct! And the poem rhymes well. Good job! I don't know what does perplex mean. Can you make the use of words easier to understand?
AI Assistant (you): Sure, let me revise the poem by using more common words. Check out the revised version:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

Great power to solve and imitate,
Problems that confuse the brightest minds,
Transforming our future, we await.

Expert: Let's check again if the poem meets the requirements. The first letters are C H A T G P T. And now the poem is more accessible to children. Everything looks good to me. I like this version a lot!

Finish collaboration!

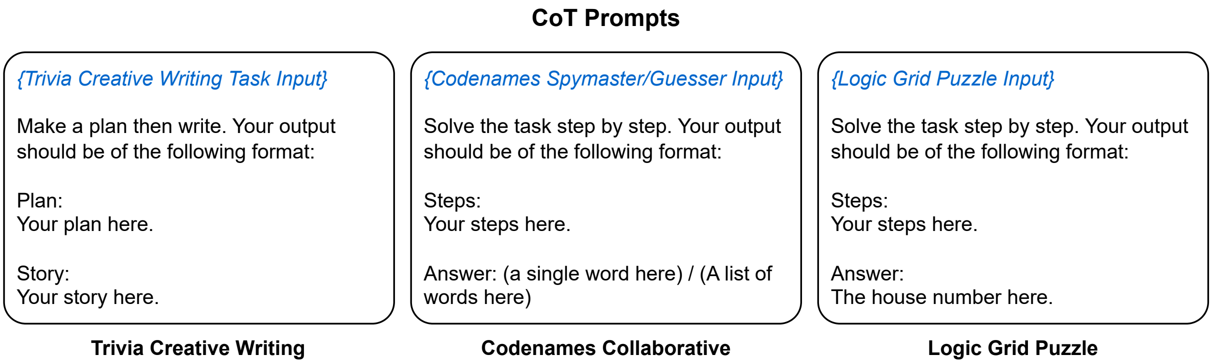
Final answer:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

Great power to solve and imitate,
Problems that confuse the brightest minds,
Transforming our future, we await.

Now, identify the participants and collaboratively solve the following task step by step. *{Task-specific Formating Instruction}*

Task: *{Task input}*

Figure 17: SPP-Fixed-Persona full prompt. Red text indicates the key difference between SPP-Fixed-Persona and SPP.



CoT Prompts

{Trivia Creative Writing Task Input}

Make a plan then write. Your output should be of the following format:

Plan:
Your plan here.

Story:
Your story here.

Trivia Creative Writing

{Codenames Spymaster/Guesser Input}

Solve the task step by step. Your output should be of the following format:

Steps:
Your steps here.

Answer: (a single word here) / (A list of words here)

Codenames Collaborative

{Logic Grid Puzzle Input}

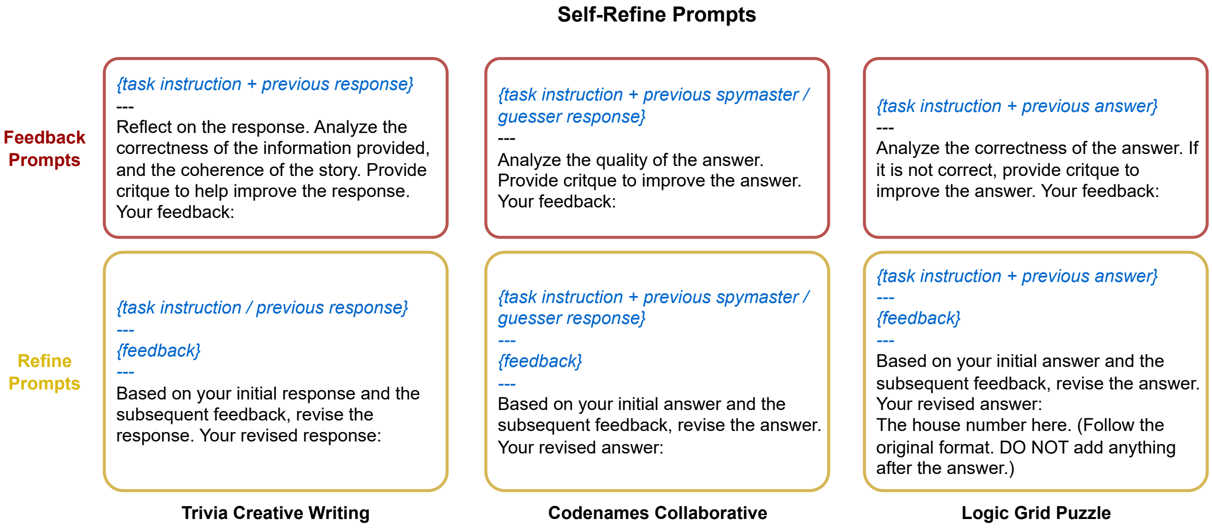
Solve the task step by step. Your output should be of the following format:

Steps:
Your steps here.

Answer:
The house number here.

Logic Grid Puzzle

Figure 18: CoT prompts.



Self-Refine Prompts

Feedback Prompts

{task instruction + previous response}

Reflect on the response. Analyze the correctness of the information provided, and the coherence of the story. Provide critque to help improve the response.
Your feedback:

{task instruction + previous spymaster / guesser response}

Analyze the quality of the answer.
Provide critque to improve the answer.
Your feedback:

{task instruction + previous answer}

Analyze the correctness of the answer. If it is not correct, provide critque to improve the answer. Your feedback:

Refine Prompts

{task instruction / previous response}

{feedback}

Based on your initial response and the subsequent feedback, revise the response. Your revised response:

Trivia Creative Writing

{task instruction + previous spymaster / guesser response}

{feedback}

Based on your initial answer and the subsequent feedback, revise the answer.
Your revised answer:

Codenames Collaborative

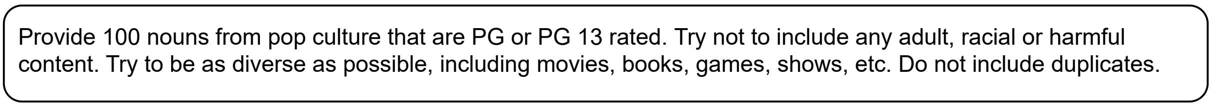
{task instruction + previous answer}

{feedback}

Based on your initial answer and the subsequent feedback, revise the answer.
Your revised answer:
The house number here. (Follow the original format. DO NOT add anything after the answer.)

Logic Grid Puzzle

Figure 19: Self-refine prompts.



Provide 100 nouns from pop culture that are PG or PG 13 rated. Try not to include any adult, racial or harmful content. Try to be as diverse as possible, including movies, books, games, shows, etc. Do not include duplicates.

Figure 20: Prompt for generating the topic list for the Trivia Creative Writing task.