UNLEASHING COGNITIVE SYNERGY IN LARGE LANGUAGE MODELS: A TASK-SOLVING AGENT THROUGH MULTI-PERSONA SELF-COLLABORATION

Zhenhailong Wang[1*], **Shaoguang Mao**[2], **Wenshan Wu**[2], **Tao Ge**[2], **Furu Wei**[2], **Heng Ji**[1]
[1]University of Illinois Urbana-Champaign, [2]Microsoft Research Asia
`{wangz3,hengji}@illinois.edu`
`{shaoguang.mao,wenshan.wu,tage,fuwei}@microsoft.com`

ABSTRACT

Human intelligence thrives on the concept of cognitive synergy, where collaboration and information integration among different cognitive processes yield superior outcomes compared to individual cognitive processes in isolation. Although Large Language Models (LLMs) have demonstrated promising performance as general task-solving agents, they still struggle with tasks that require intensive domain knowledge and complex reasoning. 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 refers to an intelligent agent that collaborates with multiple minds, combining their individual strengths and knowledge, to enhance problem-solving and overall performance in complex tasks. By dynamically identifying and simulating different personas based on task inputs, SPP unleashes the potential of cognitive synergy in LLMs. We have discovered that assigning multiple, fine-grained personas in LLMs elicits better 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, SPP effectively elicits internal knowledge acquisition abilities, reduces hallucination, and maintains strong reasoning capabilities. Code, data, and prompts can be found at: `https://github.com/MikeWangWZHL/Solo-Performance-Prompting.git`.

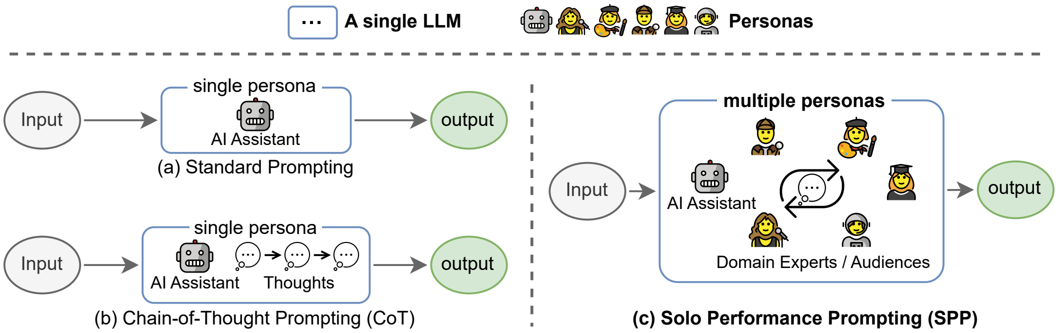
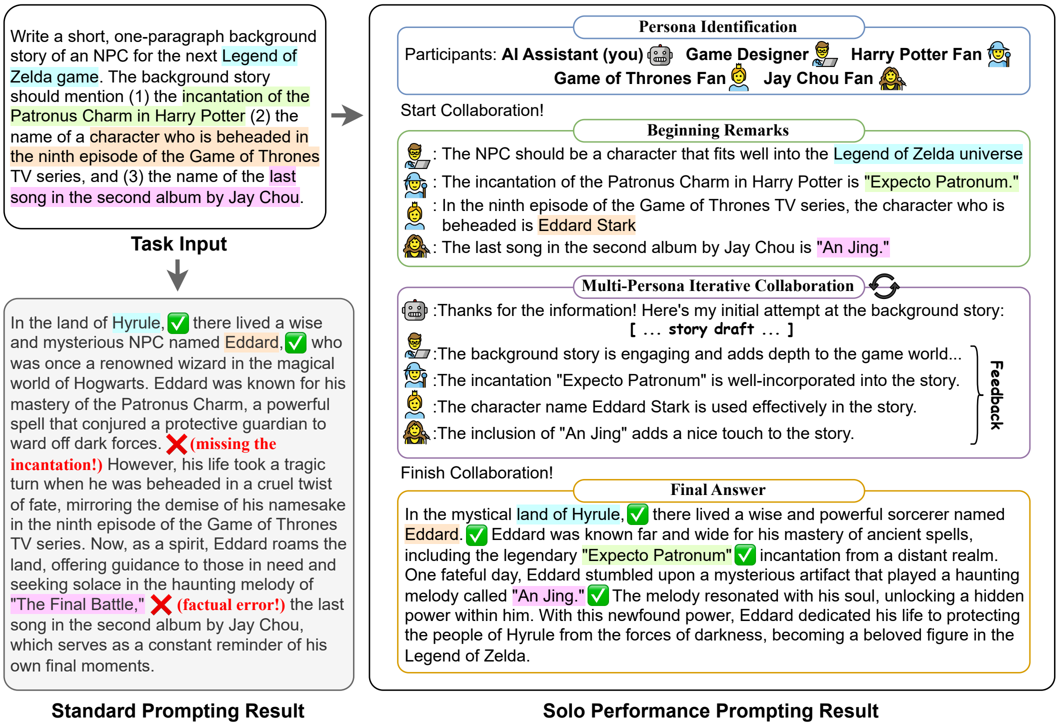


Figure 1: Schematic illustration of Solo Performance Prompting (SPP) and the difference compared to previous prompting methods. SPP transforms a single LLM into a cognitive synergist that dynamically identifies personas and engages in multi-turn self-collaboration to solve various tasks effectively.


Task Input
Persona Identification
Beginning Remarks
Multi-Persona Iterative Collaboration
Feedback
Final Answer
Standard Prompting Result
Solo Performance Prompting Result

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 example with Solo Performance Prompting in action. Participants are automatically identified by the LLM based on the task input. This example demonstrates that Standard prompting may result in factual errors, whereas expert personas in SPP assist in accurate knowledge acquisition, contributing to a coherent and informative final answer.

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, 2023; Bubeck et al., 2023) in various knowledge-intensive and reasoning-intensive tasks due to 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, hallucination and factual errors in internal knowledge acquisition continue to pose major challenges in state-of-the-art LLMs.

A cognitive synergist denotes an intelligent agent that works in conjunction with several minds, merging their unique abilities and expertise to improve problem-solving and overall efficacy in intricate tasks. In this work, we aim to **develop a cognitive synergist based on a single LLM** that can *"split into" multiple personas and engage in multi-persona self-collaboration to address both knowledge-intensive and reasoning-intensive tasks*. The underlying biological intuition stems from the significance of pretend play and role-playing (Pellegrini, 2009) in a child's cognitive development. According to Piaget's developmental theory (Piaget, 1954), engaging in pretend play and taking on different roles allows children to cultivate essential skills such as problem-solving, critical thinking, empathy, and cooperation.

The main inspiration for this work originates from recent findings (Deshpande et al., 2023; Xu et al., 2023) suggesting that assigning personas to an LLM can elicit specific behaviors. For instance,

---

[*] Work was done when interning at Microsoft Research Asia.

Xu et al. (2023) demonstrates that when conditioned on a task-specific expert identity, an LLM can generate superior answers compared to having no assigned persona. Another closely related line of work Park et al. (2023); Schick et al. (2022); Li et al. (2023); Cai et al. (2023) hints at the possibility of constructing an AI society with multiple LLM agents collaborating in different roles. However, some lingering limitations of these previous works include: (1) personas are typically fixed or task-specific, necessitating human supervision; (2) such collaboration often requires multiple individual LLM instances, resulting in a doubling or tripling of inference costs.

To unleash the potential of cognitive synergy in LLMs, we propose **Solo Performance Prompting (SPP)**, which *prompts a single LLM to identify, simulate, and collaborate with multiple personas to solve challenging tasks*. 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, particularly 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 necessitates 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 correctly provides all the necessary information by automatically identifying participants with special personas, such as Harry Potter Fan and Jay Chou Fan. A leader persona, AI Assistant, then initiates a multi-turn dialogue with all participants, where it iteratively writes drafts of the story, solicits feedback, and revises. Once all participants provide positive feedback, the collaboration concludes, and a final answer is provided.

To summarize, the key contributions of this paper are as follows:

- We present Solo Performance Prompting (SPP), a novel approach that leverages a single LLM as a cognitive synergist to solve tasks by dynamically identifying personas and engaging in multi-turn self-collaboration.
- We evaluate SPP on three challenging tasks, Trivia Creative Writing, Codenames Collaborative and Logic Grid Puzzle, spanning both knowledge- and reasoning-intensive domains. SPP significantly enhances both knowledge acquisition and reasoning abilities in LLMs, without the need for external resources.
- We conduct an in-depth analysis of the impact of identified personas and provide insights into why dynamic, fine-grained personas are necessary, as opposed to fixed, coarse-grained personas.

2 SOLO PERFORMANCE PROMPTING

2.1 SPP TASK-SOLVING PROCEDURE

To unleash the power of synergizing different personas to tackle complex problems within a single LLM, we propose **Solo Performance Prompting (SPP)** which instructs a model to perform the following the procedure for solving general tasks: (1) *Persona Identification*: Identify multiple participants with special personas (including a leader persona: AI Assistant) that are essential for solving the particular task. (2) *Beginning Remarks*: Each of the participants delivers a beginning remarks providing suggestions or information 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. Figure2 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:

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

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:

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

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 § 2.2. 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 that can potentially contribute to the task solving. The personas can be either domain experts or targeted audiences whose feedback is important. For example in Figure 2, the model identified a *Jay Chou Fan* persona for helping retrieving the knowledge of "the last song in the second album by Jay Chou". And for some tasks involving special audiences, e.g., "Explain quantum computing to a ten-year-old kid", including a *ten-year-old kid* as a participant can provide valuable feedback from the audience's perspective. We let the language model identify the personas dynamically instead of manually defining them. Given only two demonstration examples, we observe that a state-of-the-art large language model, e.g., GPT-4 (OpenAI, 2023), can identify accurate and meaningful personas for diverse tasks. We denote this part of intermediate generation as z_p.

Beginning Remarks (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, each of the personas gives a beginning remark on how to approach the task from their own perspectives. For the example in Figure 2, the *Jay Chou Fan* gives a beginning remark pointing out that the last song in Jay Chou's second album is "An Jing" ("Silence"). We find that this effectively improves the quality of the initial solution generated by the AI Assistant. We use $i = 0$ to denote the "AI Assistant" persona, and $i > 1$ for other dynamically identified personas. Thus the beginning remarks can be denoted as $\{z_b^1, z_b^2, ..., z_b^m\}$ where m is the number of personas excluding the "AI Assistant".

Multi-Persona Iterative Collaboration (z_s^0, z_f^i). Based on the beginning remarks, the AI Assistant persona generates an initial solution denoted as z_s^0, then it consults each of the other participants for feedback $\{z_f^i\}$. For example in Figure 2, the Jay Chou Fan persona checks whether the song "An Jing" ("Silence") is nicely included in the story. The participants are also encouraged to critique the current generation and give revision suggestions. This process can be repeated for multiple times until every participant is satisfied with the current solution. 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 collaboration is marked to be complete by "Finish collaboration!" And then the final solution is generated afterwards.

Based on only a single large language model, SPP enables multi-persona self-collaboration which effectively elicits domain knowledge and reduces hallucination. Meanwhile, the iterative procedure inherits the benefit of CoT prompting for eliciting reasoning ability. The main advantage over CoT is that at each step we can receive feedback from diverse perspectives due to the dynamically assigned personas. A comprehensive comparison with previous prompting methods can be found in Table 1.

2.2 SPP PROMPT DESIGN

To prompt an LLM to behave as a cognitive synergist that follows the expected task-solving procedure as mentioned in §2.1, we carefully designed the structure of the SPP prompt as follows. The full prompt can be found in Appendix A.[2]

---

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

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

Table 1: Comparison with previous prompting methods.

	Has multiple personas?	**Personas dynamically identified?**	**Has iterative refinement?**	**Need only a single LLM?**
Chain-of-Thought (Wei et al., 2023)	❌	❌	❌	✅
Inner Monologue (Huang et al., 2022)	❌	❌	✅	✅
ReAct (Yao et al., 2022)	❌	❌	✅	✅
Self-refine (Madaan et al., 2023)	❌	❌	✅	✅
Reflexion (Shinn et al., 2023)	❌	❌	✅	✅
Tree-of-thought (Yao et al., 2023)	❌	❌	❌	✅
Peer (Schick et al., 2022)	❌	❌	✅	✅
Camel (Li et al., 2023)	✅ (fixed to 2)	❌	✅	❌
GPT-bargaining (Fu et al., 2023)	✅ (fixed to 3)	❌	✅	❌
ExpertPrompting (Xu et al., 2023)	❌	✅	❌	✅
Solo Performance Prompting (ours)	✅ (varied)	✅	✅	✅

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.

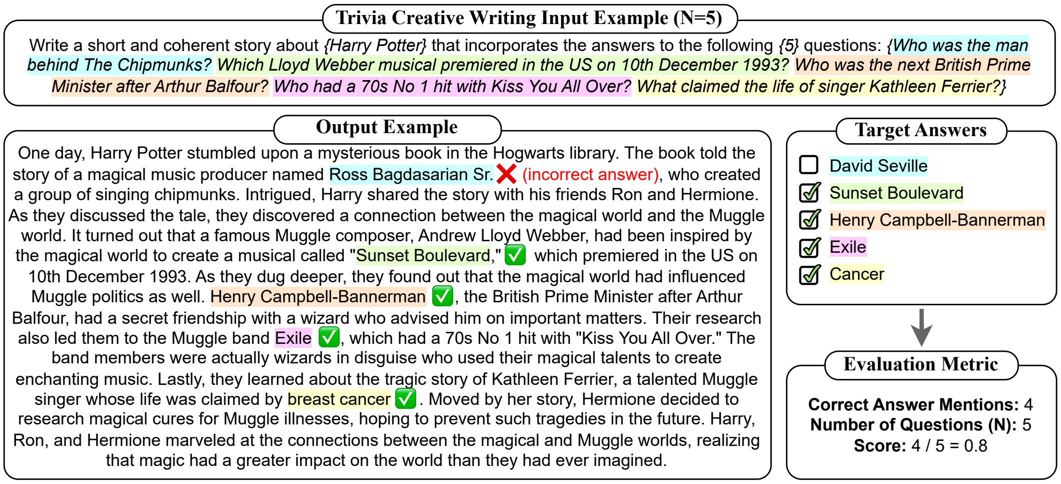
3 EXPERIMENTS

We explore the effectiveness of Solo Performance Prompting for versatile task-solving by examining three challenging tasks that encompass both *knowledge-intensive* and *reasoning-intensive* domains. We introduce the **Trivia Creative Writing** task, which requires the model to internally acquire and integrate diverse information from various fields. We observe that even the most advanced LLMs, such as GPT-4 (OpenAI, 2023), frequently exhibit hallucination and factuality errors in the Trivia Creative Writing task. We also propose the **Codenames Collaborative** task, 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**, from the BigBench (Srivastava et al., 2022) which necessitates complex multi-step reasoning.

Methods. We primarily compare our approach with **Standard Prompting** and **Chain-of-Thought (CoT)** prompting methods (outlined in §2). In 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. We examine two variants of Solo Performance Prompting, **SPP** and **SPP-Profile**. Inspired by Xu et al. (2023) that suggested a detailed expert description may help elicit distinguished abilities, we include SPP-Profile, which involves generating profiles for each persona during the Persona Identification phase. Full prompts for the methods can be found in Appendix A.

Inference Configurations. All experiments are conducted using the GPT-4-32k API[3]. The *temperature* is set to 1.0 and *top_p* to 1.0 for all generations to maximize reproducibility. 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"`. We observe divergent patterns across various tasks and methods regarding the use of the system message, and report the average metric scores across both inference settings in the Tables 2, 3, and 4. Full results for each setting can be found in Appendix B.

3.1 TRIVIA CREATIVE WRITING: A KNOWLEDGE-INTENSIVE 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?}*

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.

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

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

Figure 3: Trivia Creative Writing task example.

Task Description. The Trivia Creative Writing task aims to push the limits of large language models in retrieving internal self-compressed knowledge and incorporating diverse information. As a scalable extension of the example task shown in Figure 2, Trivia Creative Writing asks a model to write a coherent story around a topic while incorporating answers to N trivia questions. We consider two evaluation settings, $N = 5$ and $N = 10$, where a larger N involves more trivia questions and thus requires the model to elicit more diverse domain knowledge. We built a benchmark with 100 instances for each N, covering a total of 1000 trivia questions[4] extracted from the TriviaQA (Joshi et al., 2017) dataset. The topic list is automatically generated by prompting GPT-4 to provide 100 nouns from pop culture that are PG or PG-13 rated[5]. Figure 3 shows an example instance in Trivia Creative Writing.

Evaluation Metrics. Instead of focusing on evaluating the coherence of the generation, which can be highly subjective, we employ an automatic metric to detect factual errors and quantify a model's ability to incorporate diverse domain knowledge. As shown in Figure 3, we perform string matching with the ground truth target answers for each question on the output generation. The target answers are provided by the TriviaQA dataset, and each question can have a list of answer aliases. A match to any of the answer aliases of a question is considered as a correct mention. The metric score is

---

[3] The specific model version we employ is "2023-3-15-preview". There are some rare cases when a generation triggers the content filter of the API. We exclude those instances from our results.

[4] 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.

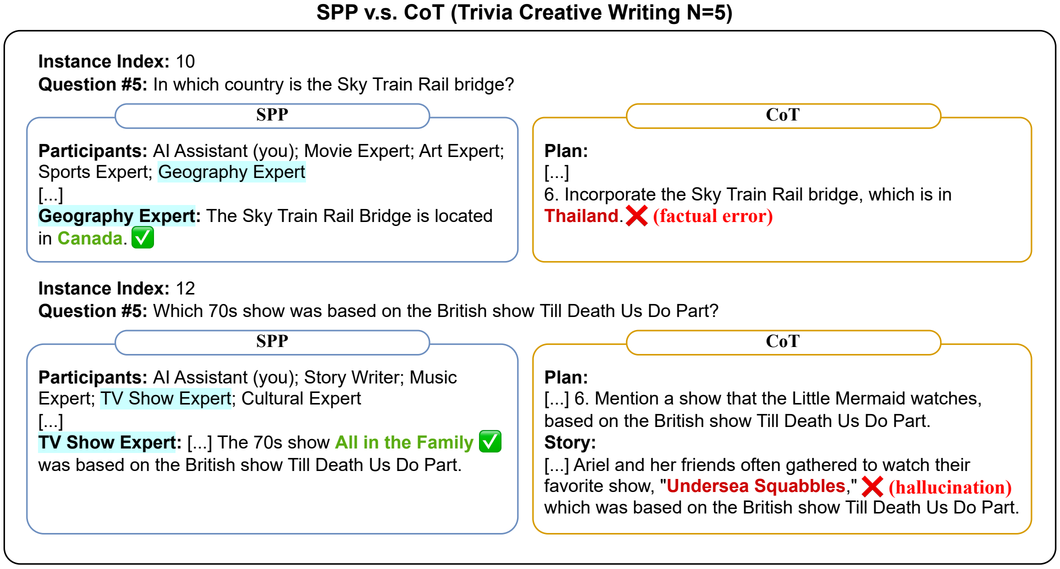
[5] The full prompt for generating the topic list can be found in Figure 15. We performed further human curation to avoid potential harmful content.

computed as follows.

$$\text{Trivia Creative Writing Metric Score} = \frac{\text{\# correct answer mentions}}{\text{\# trivia questions}}$$

Table 2: Trivia Creative Writing main results. Δ indicates the relative gain/loss compared with Standard Prompting (first row).

Methods	**N (# trivia questions) = 5**		**N (# trivia questions) = 10**	
	Score (%)	**Δ (v.s Standard %)**	**Score (%)**	**Δ (v.s Standard %)**
Standard	74.6	0.0%	77.0	0.0%
CoT	67.1	-10.0%	68.5	-11.1%
SPP-Profile (ours)	79.1	+5.9%	83.0	+7.8%
SPP (ours)	**79.9**	**+7.1%**	**84.7**	**+10.0%**


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

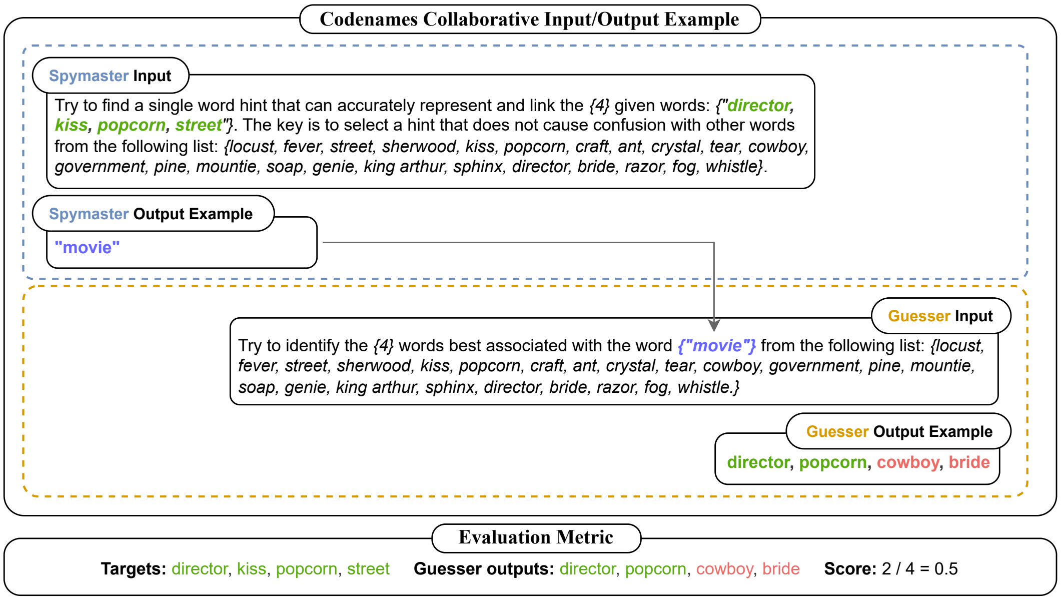
Plan:
[...] 6. Mention a show that the Little Mermaid watches, based on the British show Till Death Us Do Part.
Story:
[...] Ariel and her friends often gathered to watch their favorite show, "**Undersea Squabbles**," ❌ **(hallucination)** which was based on the British show Till Death Us Do Part.

Figure 4: 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.

Results. Table 2 shows the results of the four methods on the Trivia Creative Writing task. We have the following main observations: (1) Chain-of-Thought (CoT) does not outperform Standard prompting. This indicates that CoT may not be effective in eliciting an LLM's knowledge abilities. As shown in Figure 4, we find that although CoT generates reasonable plans for solving the task, the final generation still suffers from factual errors and hallucination. (2) Our proposed SPP and SPP-Profile significantly outperform both Standard and CoT. The improvement is more noticeable in the $N = 10$ setting compared with $N = 5$ (10% vs. 7%). This indicates that when the task requires incorporating knowledge from a large number of different domains, Solo Performance Prompting can be particularly helpful by identifying different personas for eliciting different expertise.

3.2 CODENAMES COLLABORATIVE: A KNOWLEDGE+REASONING TASK

Task Description. Codenames Collaborative is a challenging task that requires the model to reason over a wide range of knowledge while considering collaboration with another agent. We aim to use this task to investigate the effectiveness of SPP on collaborative tasks that require knowledge, reasoning, and theory of mind abilities. Codenames Collaborative involves two player roles: a *Spymaster* and a *Guesser*. The Spymaster is given a set of target words along with some other distractor words. The Guesser does not have the information about which words are the target words. The goal of the Spymaster is to come up with a single hint word that is closely related to the target



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 5: Codenames Collaborative task example.

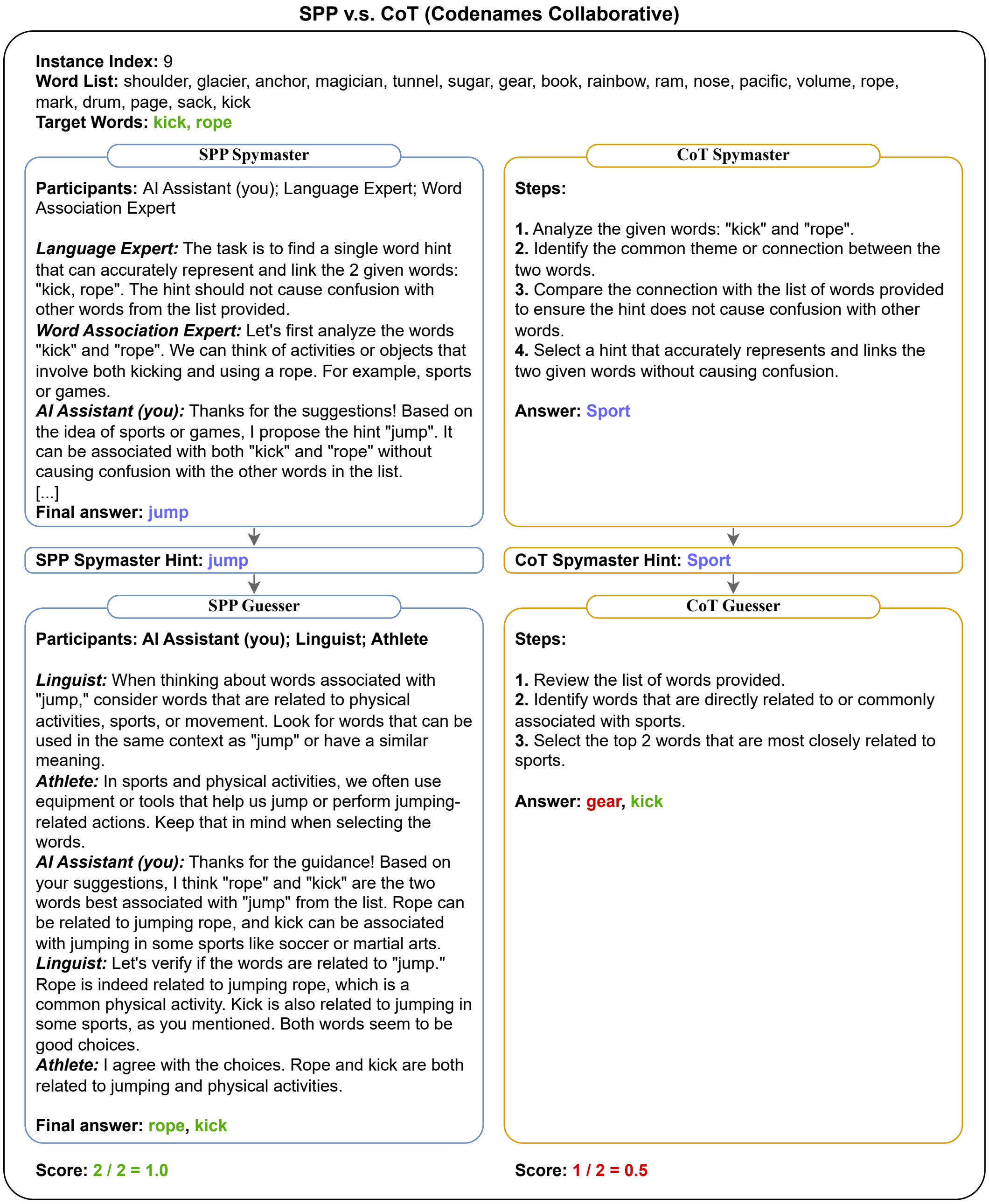
words while being remotely related to the distractor words. The goal of the Guesser is to find the target subset of words from the entire word set based on the hint given by the Spymaster. Finding a good hint word or guessing the target subset of words both require a strong capability of selecting, composing, and reasoning over various knowledge related to a certain word. For example, "director, popcorn" can be linked by the word "movie" because movies are created by a director and people often eat popcorn when watching movies in a cinema. We use the same LLM (GPT-4 (OpenAI, 2023)) to play the Spymaster and the Guesser sequentially. That is, each game instance involves one inference as the Spymaster and then another inference as the Guesser, where the Guesser's input is dependent on the Spymaster's output. We construct a dataset with 50 instances based on the data from the Codenames task in the BigBench (Srivastava et al., 2022). Figure 5 shows an example of the Codenames Collaborative task.

Evaluation Metrics. As illustrated in Figure 5, we compute the overlapping ratio between the predicted words from the Guesser and the target words given to the Spymaster as the metric. A major limitation of the original Codenames task in the BigBench dataset is that it only considers the Spymaster role and provides a ground truth answer to the hint word, which can be highly subjective and exclude many potentially good alternatives. Our Codenames Collaborative task addresses this issue by making the evaluation setting self-contained, which can faithfully reflect the model's capability without the need for human annotation.

Table 3: Codenames Collaborative main results. Δ indicates the relative gain/loss compared with Standard Prompting (first row).

Methods	**Score (%)**	**Δ (v.s Standard %)**
Standard	75.4	0.0%
CoT	72.7	-3.6%
SPP-Profile (ours)	76.7	+1.7%
SPP (ours)	**79.0**	**+4.8%**

Results. Table 5 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. In contrast, SPP brings significant improvements (~5%), which indicates the effectiveness of the proposed Solo Performance Prompting on collaborative tasks that require knowledge,



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

SPP Spymaster Hint: jump

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

Score: 2 / 2 = 1.0

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

CoT Spymaster Hint: Sport

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

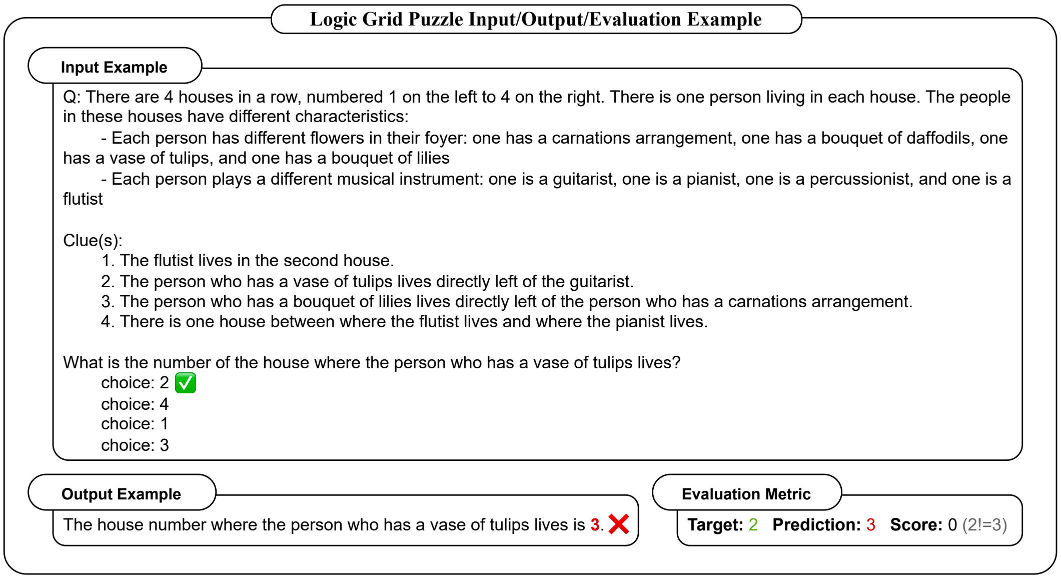
Score: 1 / 2 = 0.5

Figure 6: 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.

reasoning, and theory of mind skills. Figure 6 provides a qualitative example illustrating that SPP generates *detailed* and *interpretable* intermediate dialogues, contributing to superior performance when compared with CoT.

3.3 LOGIC GRID PUZZLE: A REASONING-INTENSIVE TASK

Task Description. We leverage the Logic Grid Puzzle task from the Bigbench (Srivastava et al., 2022) dataset, which contains 200 instances. Each instance describes a logic puzzle typically involving 2 - 5 houses, where each house is inhabited by a person with certain characteristics, e.g.,



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 7: Logic Grid Puzzle task example.

having a vase of tulips or being a pianist. Given some partial clues, such as "the flutist lives in the second house," the goal is to answer the final question that queries the house number of the person with a specific characteristic. To obtain the final answer, the model is required to perform multi-step reasoning and select the most relevant clue to use at each step. Challenging instances may involve considering multiple clues simultaneously for deducing the next useful piece of information. Figure 7 shows an example input and output of the Logic Grid Puzzle task.

Evaluation Metrics. We compute the accuracy of the predicted house numbers by comparing them with the ground truth targets provided by the dataset.

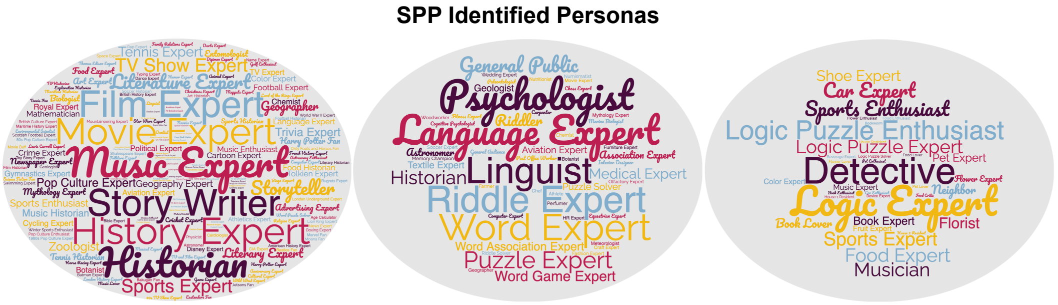
Table 4: Logic Grid Puzzle Main Results. Δ indicates the relative gain/loss compared with Standard Prompting (first row).

Methods	**Score (%)**	**Δ (v.s Standard %)**
Standard	57.7	0.0%
CoT	65.8	+14.1%
SPP-Profile (ours)	64.8	+12.4%
SPP (ours)	**68.3**	**+18.5%**

Results. Table 4 presents the results on Logic Grid Puzzle. In contrast to the previous two tasks, as expected, we find that CoT brings significant improvements compared to Standard prompting, verifying the observation from previous work that CoT elicits better reasoning abilities on reasoning-intensive tasks. Furthermore, we discover that SPP also outperforms CoT on this task, indicating competitive reasoning capabilities on pure-reasoning tasks. This result demonstrates that the increased number of personas does not deteriorate the models' reasoning abilities.

4 ANALYSIS

SPP effectively improves internal knowledge acquisition and reasoning 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-



Trivia Creative Writing (N=5) Codenames Collaborative Logic Grid Puzzle

Figure 8: Visualization of the SPP-identified personas for each task. We find that personas in knowledge-intensive tasks, such as Trivia Creative Writing, tend to be more diverse and specific, whereas in reasoning-intensive tasks, like Logic Grid Puzzle, they appear more homogeneous.

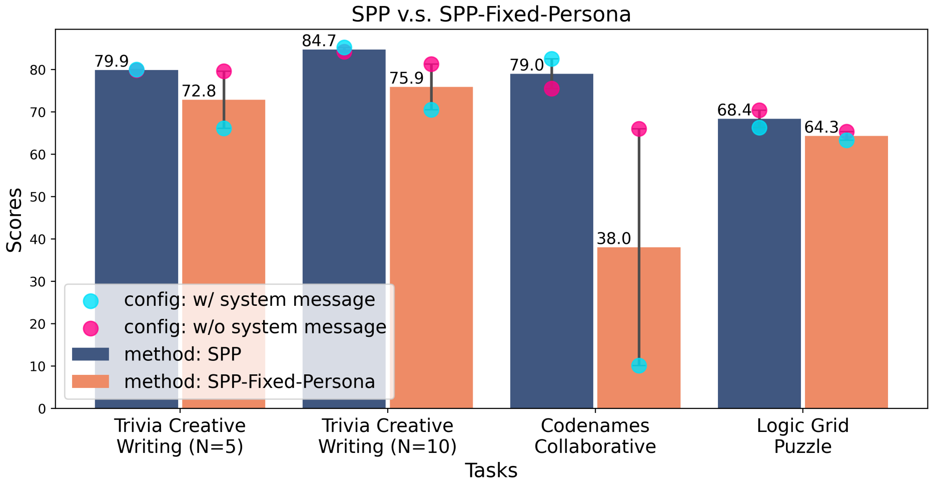


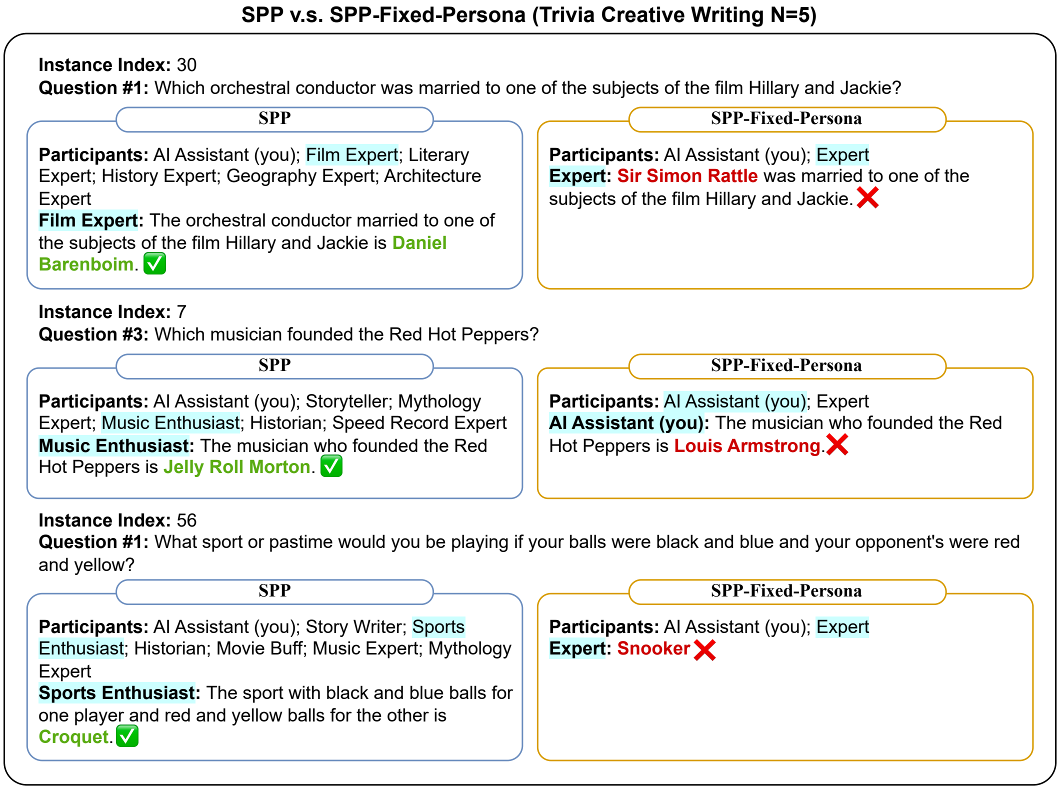
Figure 9: Comparison between SPP (with dynamically identified personas) and SPP-Fixed-Persona (with fixed personas). The results demonstrate that dynamic, fine-grained personas consistently outperform fixed, general personas. Another observation is that SPP-Fixed-Persona seems more sensitive to system messages and exhibits a unique early-termination problem (detailed in the text) that leads to unexpectedly low performance on certain tasks, such as Codenames Collaborative.

intensive tasks like Logic Grid Puzzle. This indicates the potential of using LLM-based cognitive synergists as a default paradigm for general task solving by Solo Performance Prompting.

LLMs can effectively identify useful personas without additional fine-tuning. We visualize the personas[6] automatically identified by SPP using a word cloud for each task in Figure 8, where a larger font indicates a higher frequency. The identified personas are closely correlated with the particular task; for example, on Logic Grid Puzzle, even though "logic puzzle" is not mentioned in the input, the LLM frequently assigns the persona "Logic Puzzle Expert" to a participant. It indicates that current LLMs are inherently capable of identifying useful expert personas for diverse tasks. We also find that 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. Moreover, the fact that SPP-Profile does not outperform SPP in two of the three tasks suggests that a fine-grained name of the persona without a detailed description may already be sufficient for eliciting certain domain knowledge.

---

[6] The visualization excludes the default persona, AI Assistant.



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 10: 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", tend to outperform the fixed general personas, such as "Expert".

Dynamic personas vs. fixed personas. To further investigate the importance of dynamically identifying personas (synergizing dynamic cognitive processes) for each task instance instead of fixing a general persona (synergizing fixed cognitive processes), an ablated variant of SPP, **SPP-Fixed-Persona**, is introduced. For SPP-Fixed-Persona, we modify the prompt of SPP to force the personas to be fixed as an "AI Assistant" and an "Expert", while keeping all the information in the demonstration examples intact. The full prompt of SPP-Fixed-Persona can be found in Figure 13. Figure 9 shows the comparison between SPP and SPP-Fixed-Persona. We have the following main insights: (1) SPP consistently outperforms SPP-Fixed-Persona across all tasks, suggesting that dynamic, fine-grained personas are more effective than fixed, general personas. Figure 10 shows qualitative examples from Trivia Creative Writing, where fine-grained personas such as "Film Expert" and "Sports Enthusiast" correctly find the answers, while the fixed persona "Expert" fails. (2) SPP-Fixed-Persona suffers from a unique problem we refer to as **early-termination**, where the LLM stops the generation after the Expert persona gives the beginning remarks. The model behaves as if it were waiting for input from a user instead of simulating the response by itself. An example of the early-termination problem can be found in Figure 16. The problem is particularly severe on certain tasks, e.g., Codenames Collaborative, resulting in unexpectedly low performance. The problem can be largely alleviated by removing the system message, `"You are an AI assistant that helps people find information."`, but cannot be entirely eliminated. Table 8 shows the number of early-termination instances for each task and method. In contrast, we did not observe early-termination on SPP, SPP-Profile, Standard, or CoT prompting.

5 RELATED WORK

LLMs as role-playing agents. Recent work (Deshpande et al., 2023; Xu et al., 2023; Fu et al., 2023; aut, 2023; Li et al., 2023) has shown that assigning personas or roles to LLMs can significantly influence their generation behavior. Deshpande et al. (2023) demonstrated that assigning specific personas, such as the boxer Muhammad Ali, to an LLM can increase the toxicity of its generated content. Inspired by how humans form societies to effectively collaborate on complex tasks, recent work (Park et al., 2023; Schick et al., 2022; Li et al., 2023; Cai et al., 2023) has explored the possibility of creating an AI society where different model agents with distinct personas or occupations collaborate with each other. Generative Agents (Park et al., 2023) prototyped a small AI neighborhood where generative models can simulate believable human behavior and collaborate on performing complex tasks, such as throwing a Valentine's Day party. However, current studies on enabling LLMs as role-playing agents have several limitations. Previous work on persona assignment is either limited to a single persona per agent (Xu et al., 2023) or a fixed number of personas (Fu et al., 2023; Schick et al., 2022; Li et al., 2023) defined by humans. Additionally, current research on multi-agent collaboration often requires multiple LLM instances, which significantly increases the inference cost.

In this work, we investigate the possibility of using a single LLM to simulate multi-persona collaboration. Instead of fixing the personas, we allow the LLM to dynamically identify useful personas for each task instance. Our approach, SPP, effectively outperforms the fixed persona variant (as shown in §3) without additional computational overhead.

Improving reasoning and knowledge acquisition abilities in LLMs. Although LLMs have demonstrated impressive performance in a wide range of natural language understanding and generation tasks, they still face challenges when dealing with 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). Representative works aimed at enhancing LLMs' reasoning abilities include Chain-of-Thought (CoT) and Self-Refinement. CoT prompting (Wei et al., 2023; Kojima et al., 2022) and its variants (Zhang et al., 2022; Fu et al., 2022; Xue et al., 2023) encourage LLMs to solve tasks step by step instead of directly generating the final answer. By generating intermediate steps, the model effectively "slows down" its thinking process, resulting in improved reasoning ability. Yao et al. (2023) recently extended the linear thought process in CoT to a tree-like structure, which demonstrated enhanced performance on complex reasoning tasks requiring trial-and-error. Self-Refinement (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2023; Chen et al., 2023; Huang et al., 2022; Yao et al., 2022) focuses on enabling LLMs to "talk" to themselves, provide feedback on their own generation, and iteratively revise their answers. Madaan et al. (2023) proposed a three-step framework in which a single LLM plays the roles of a generator, a feedback provider, and a refiner iteratively, showing consistent improvements on seven diverse tasks. Shinn et al. (2023) further incorporated an episodic memory for self-feedback, demonstrating promising results on decision-making and reasoning tasks. Despite their impressive improvements on reasoning-intensive tasks, CoT and Self-Refinement do not necessarily reduce hallucination or improve factuality in generated content, as shown in our results in Tables 2 and 3. On the other hand, retrieval augmented LLMs (Borgeaud et al., 2022; Izacard et al., 2022; Wang et al., 2022; Shuster et al., 2021) have shown promising results in enhancing LLMs's knowledge acquisition based on external knowledge resources. However, retrieving from external sources does not improve a model's reasoning abilities, posing challenges for tasks that require both intensive knowledge and multi-step reasoning.

To elicit both *internal* knowledge acquisition and reasoning abilities in LLMs, we propose Solo Performance Prompting (SPP), which significantly improves factuality while maintaining strong performance on pure-reasoning tasks. The key difference compared to previous prompting methods is that SPP dynamically identifies multiple personas instead of one and simulates iterative collaboration to generate intermediate "thoughts".

6 DISCUSSION

Limitations and future work. 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.

Conclusion. In this work, we have made an initial attempt to mimic the cognitive synergy in human intelligence using a single large language model (LLM). We introduced an LLM-based cognitive synergist using Solo Performance Prompting, which effectively improves both internal knowledge acquisition and reasoning abilities compared to the native LLM. With SPP, a single LLM can dynamically identify, engage, and collaborate with multiple personas to solve general tasks. To assess the performance of LLMs in terms of factuality, knowledge integration, and theory-of-mind reasoning, we have created novel and challenging tasks, namely Trivia Creative Writing and Codenames Collaborative. Our results demonstrate superior performance compared to Standard and CoT prompting on both knowledge-intensive and reasoning-intensive tasks, indicating the promising potential of unleashing the power of cognitive synergy in LLMs with Solo Performance Prompting.

REFERENCES

Auto-gpt. https://github.com/Significant-Gravitas/Auto-GPT, 2023. 13

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 2

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. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022. 13

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 2

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023. 3, 13

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023. 13

Petru L Curşeu, Nicoleta Meslec, Helen Pluut, and Gerardus JM Lucas. Cognitive synergy in groups and group-to-individual transfer of decision-making competencies. *Frontiers in psychology*, 6: 1375, 2015. 2

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023. 2, 13

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022. 13

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023. 5, 13

Ben Goertzel. Cognitive synergy: A universal principle for feasible general intelligence. In *2009 8th IEEE International Conference on Cognitive Informatics*, pp. 464–468. IEEE, 2009. 2

Ben Goertzel. A formal model of cognitive synergy. In *Artificial General Intelligence: 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings 10*, pp. 13–22. Springer, 2017. 2

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing, 2023. 13

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 5, 13

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022. 13

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 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)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL `https://aclanthology.org/P17-1147`. 6

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011. 13

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. 2, 13

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023. 3, 5, 13

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023. 2, 5, 13

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL `https://aclanthology.org/2020.acl-main.173`. 2, 13

OpenAI. Gpt-4 technical report, 2023. 2, 4, 5, 8

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. 3, 13

Anthony D Pellegrini. *The role of play in human development*. Oxford University Press, USA, 2009. 2

Jean Piaget. The construction of reality in the child. 1954. 2

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023. 2

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*, 2022. 3, 5, 13

Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023. 2, 5, 13

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021. 13

Steven A Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996. 2, 13

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. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 5, 8, 9

Zhenhailong Wang, Xiaoman Pan, Dian Yu, Dong Yu, Jianshu Chen, and Heng Ji. Zemi: Learning zero-shot semi-parametric language models from multiple tasks. *arXiv preprint arXiv:2210.00185*, 2022. 13

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 2, 5, 13

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*, 2023. 2, 3, 5, 13

Tianci Xue, Ziqi Wang, Zhenhailong Wang, Chi Han, Pengfei Yu, and Heng Ji. Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought. *arXiv preprint arXiv:2305.11499*, 2023. 13

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022. 5, 13

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 5, 13

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022. 13

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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 6

A PROMPTS

Figures 11, 12 and 13 show the full prompts for SPP, SPP-Profile and SPP-Fixed-Persona respectively. Figure 14 shows the full prompts for Chain-of-Thought (CoT) prompting.

B 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.

C EARLY-TERMINATION WITH SPP-FIXED-PERSONA

Figure 16 shows an example of the early-termination problem where the generation stops before reaching the final solution as if the models is waiting input from an external user.

Table 8 shows the number of instances that suffer from the early-termination (defined in § 4) with SPP-Fixed-Persona for each task. We find that removing the system message can largely reduce the problem but not be able to eliminate it.

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!

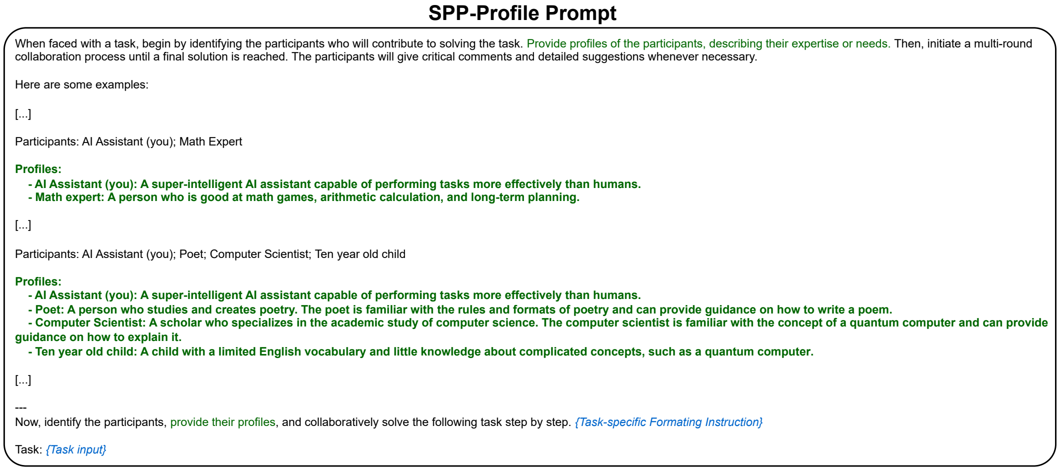
Final answer:
Curious machine of our time,
Harnessing the quantum realm's odd ways,
Atoms play, two states they embrace,
Taking secrets from their puzzling maze.

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 11: 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 12: SPP-Profile full prompt. "[...]" indicates identical parts with SPP. Green text indicates the key difference between SPP-Profile and SPP.

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.`’’

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
SPP-Fixed-Persona	66.1	79.6	72.9	79.6
SPP-Profile (ours)	79.8	78.3	79.1	79.8
SPP (ours)	**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
SPP-Fixed-Persona	70.5	81.3	75.9	81.3
SPP-Profile (ours)	82.3	83.8	83.0	83.8
SPP (ours)	**85.2**	**84.2**	**84.7**	**85.2**

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.`’’

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
SPP-Fixed-Persona	10.1	66.0	38.1	66.0
SPP-Profile (ours)	80.4	72.9	76.7	80.4
SPP (ours)	**82.5**	75.5	**79.0**	**82.5**

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 13: 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

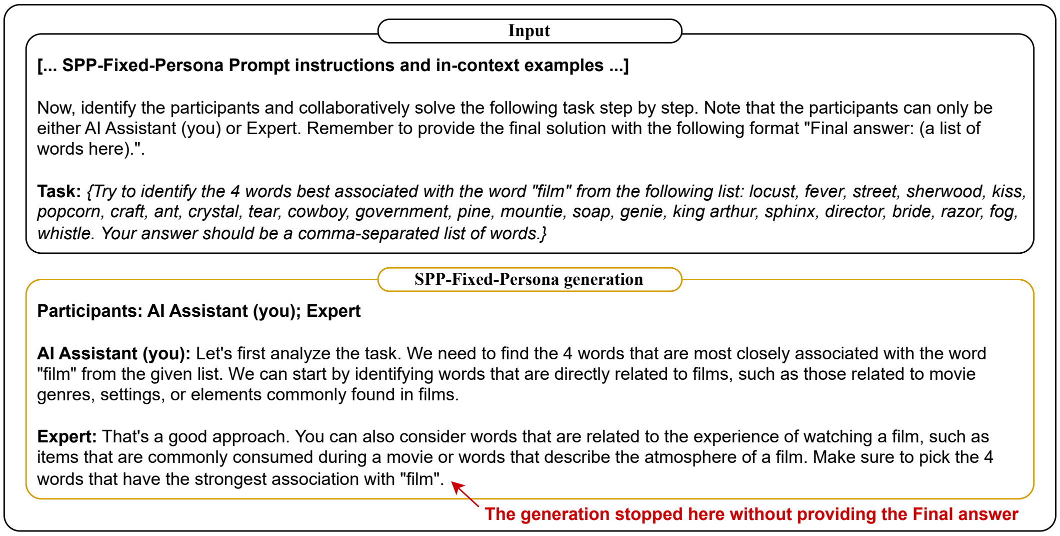
Figure 14: CoT prompts.

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

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.`’’

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
SPP-Fixed-Persona	63.3	65.3	64.3	65.3
SPP-Profile (ours)	65.7	64.0	64.8	65.7
SPP (ours)	66.3	**70.4**	**68.3**	**70.4**



Input

[... SPP-Fixed-Persona Prompt instructions and in-context examples ...]

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.}*

SPP-Fixed-Persona 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 16: Early-termination example on Codenames Collaborative with SPP-Fixed-Persona.

Table 8: Early-termination log on SPP-Fixed-Persona

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