

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 elicits internal knowledge acquisition abilities, reduces hallucination, and maintains strong reasoning capabilities. Additionally, comparative experiments have shown that cognitive synergy capability only emerges in GPT-4 and does not appear in other less capable models, such as GPT-3.5-turbo and Llama2-13b-chat. 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 hal-

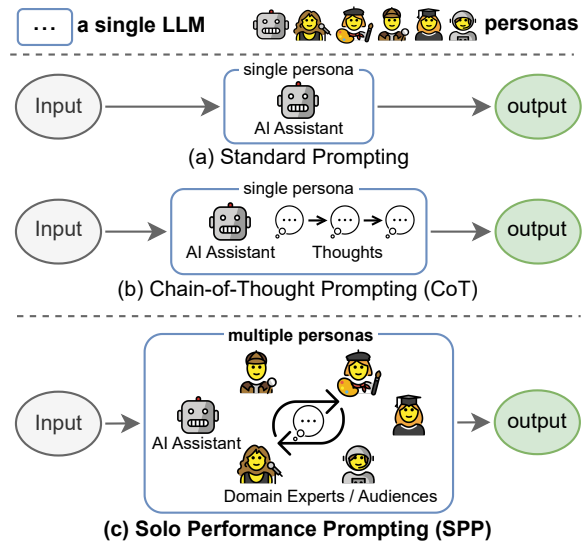


Figure 1: Schematic illustration of Solo Performance Prompting (SPP) and the difference compared to previous prompting methods. SPP transforms a **single LLM** into a **cognitive synergist** that **dynamically identifies personas** and **engages in multi-turn self-collaboration** to solve various tasks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

lucination (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 is an intelligent agent

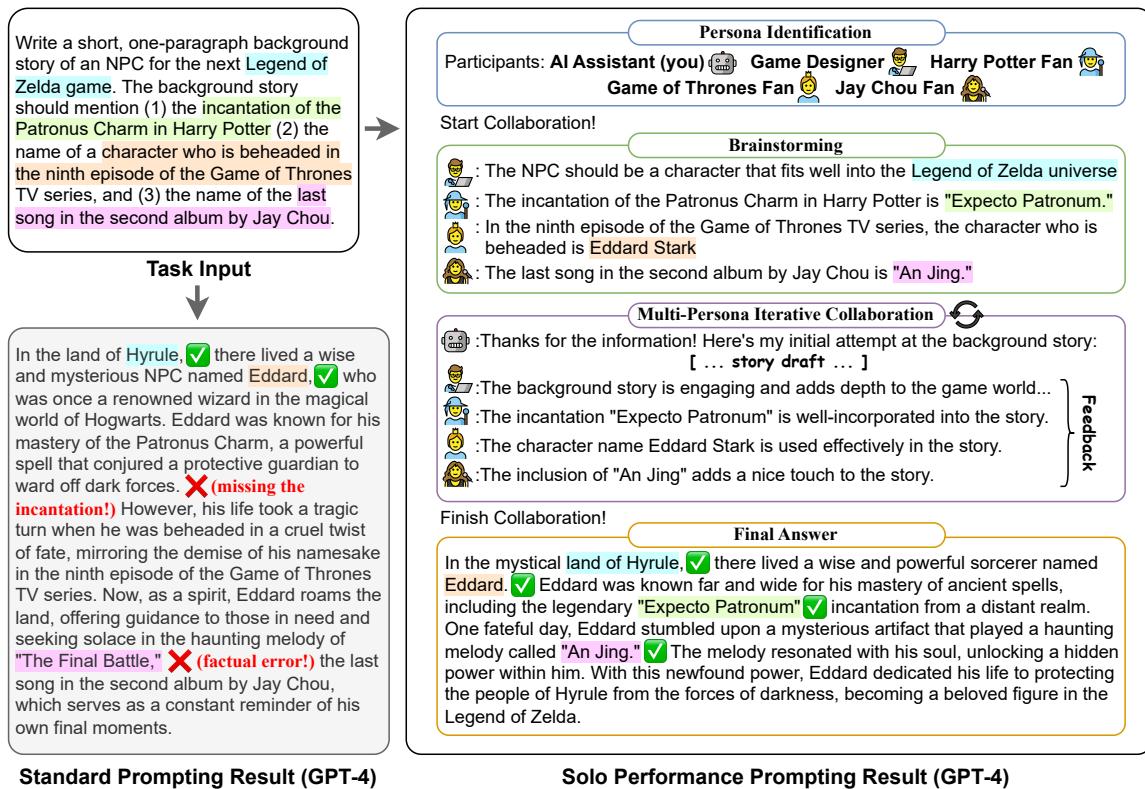


Figure 2: Task-solving example of Solo Performance Prompting (SPP) with GPT-4. The personas of the participants are automatically identified by GPT-4 based on the task input. This example demonstrates that standard prompting may result in factual errors, whereas expert personas identified by SPP contribute to informative and coherent answers. Note that, in real-world applications, the domains can vary not only within entertainment but also encompass history, science, education, healthcare, etc.

062 that collaborates with multiple minds to enhance
063 problem-solving and efficacy in complex tasks. In
064 this work, we aim to **create a cognitive syner-**
065 **gist based on a single LLM** that can "split into"
066 *multiple personas and engage in self-collaboration*
067 *to solve both knowledge-intensive and reasoning-*
068 *intensive tasks*. This idea is heavily inspired by the
069 role of pretend play (Piaget, 1954; Pellegrini, 2009)
070 in cognitive development and recent findings that
071 assigning personas (Deshpande et al., 2023; Xu
072 et al., 2023) to LLMs can elicit specific behaviors,
073 improve answer quality, and potentially build an
074 AI society (Park et al., 2023; Schick et al., 2022;
075 Li et al., 2023; Cai et al., 2023) with collabora-
076 tive LLM agents. However, as shown in Table 1,
077 limitations in previous works include fixed or task-
078 specific personas requiring human supervision and
079 increased inference costs due to multiple LLM in-
080 stances.

081 To unleash the potential of cognitive synergy in
082 LLMs, we propose **Solo Performance Prompt-**
083 **ing (SPP)**, which *prompts a single LLM to identify,*
084 *simulate, and collaborate with multiple personas*
085 *to solve challenging tasks*. Figure 1 provides a

086 high-level overview of SPP. Here, a persona can
087 represent either a domain expert, such as a movie
088 enthusiast, or a target audience, such as a ten-year-
089 old child. Through the dynamic identification of
090 various personas, we empower a single LLM to ac-
091 quire diverse domain knowledge accurately without
092 additional retrieval systems. By facilitating multi-
093 turn self-collaboration, we enable self-revision and
094 self-feedback from various perspectives without
095 requiring additional agents.

096 In real-world scenarios, such as those in creative
097 industries, there is often a need to incorporate di-
098 verse information from different domains. Figure 2
099 presents a concrete example of how SPP operates
100 on a challenging task that requires creative integra-
101 tion of information from various domains, such as
102 the Legend of Zelda game, Harry Potter movies,
103 and Jay Chou’s albums. Standard prompting fails
104 to generate satisfactory output due to missing es-
105 sential information and factual errors. In contrast,
106 SPP produces informative and coherent answers
107 by automatically identifying expert personas and
108 engaging in a multi-turn self-collaboration. In this
109 process, the AI Assistant persona iteratively writes

	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)	✓	✓	✓

Table 1: Comparison with previous prompting methods.

drafts of the story, solicits feedback from other participants, and revises accordingly.

To explore the prevalence of cognitive synergy ability in different LLMs, we apply SPP to LLMs with 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. In summary, the key contributions of this paper are as follows:

- We introduce **Solo Performance Prompting (SPP)**, which simulates multi-agent, multi-persona collaboration using a single LLM for effective task-solving.
- We evaluate SPP across **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 GPT-4, without the need for external resources.
- 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 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

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) 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:

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

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

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

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

where the SPP prompt (p_{spp}) includes a high-level instruction and two carefully crafted demonstration examples¹ that showcase the expected task-solving

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

179 procedure of SPP. We describe the design details
180 of the prompt in §E.1. The corresponding interme-
181 diate generations (z) of SPP are detailed below.

182 **Persona Identification (z_p).** Given an input task,
183 SPP first generates a list of participants with differ-
184 ent personas. The personas can be either domain
185 experts or targeted audiences whose feedback is
186 important. For example in Figure 2, the model iden-
187 tified a *Jay Chou Fan* persona for helping acquiring
188 the knowledge of "the last song in the second al-
189 bum by Jay Chou". We let the language model
190 identify the personas dynamically instead of manu-
191 ally defining them. Given only two demonstration
192 examples (detailed in §E), we observe that a state-
193 of-the-art large language model, e.g., GPT-4 (Ope-
194 nAI, 2023b), can identify accurate and meaningful
195 personas for diverse tasks. We denote this part of
196 intermediate generation as z_p in Equation 3.

197 **Brainstorming (z_b^i).** Among the identified par-
198 ticipants, "AI Assistant (you)" is treated as a leader
199 persona that initiates the collaboration and gener-
200 ates initial solutions. Before generating the ini-
201 tial answer, the personas brainstorm on how to
202 approach the task from their own perspectives. For
203 the example in Figure 2, the *Jay Chou Fan* points
204 out that the last song in Jay Chou’s second album
205 is "An Jing" ("Silence"). We find that the brain-
206 storming phase effectively improves the quality of
207 the initial solution. In Equation 3, the superscript
208 $i = 0$ is used to denote the "AI Assistant" persona,
209 while $i > 1$ represents other dynamically iden-
210 tified personas. The intermediate generations of the
211 brainstorming step are denoted as $\{z_b^1, z_b^2, \dots, z_b^m\}$.

212 **Multi-Persona Iterative Collaboration (z_s^0, z_f^i).**
213 Based on the brainstorming remarks, the AI Assis-
214 tant persona generates an initial solution z_s^0 , then it
215 consults each of the other participants for feedback
216 $\{z_f^i\}$. The participants are encouraged to critique
217 the current generation and give revision sugges-
218 tions. For example in Figure 2, the *Jay Chou Fan*
219 persona checks whether the song "An Jing" ("Sil-
220 ence") is nicely included in the story. This process
221 can be repeated for multiple times until every par-
222 ticipant is satisfied with the current solution. In
223 Equation 3, we denote the intermediate generations
224 of the multi-turn dialogue as $\{z_s^0, z_f^1, \dots, z_f^m\}_{j=1\dots n}$
225 where n is the number of iterations before reaching
226 the final answer.

227 Based on only a single large language model,
228 SPP enables multi-persona self-collaboration

229 which effectively elicits domain knowledge and
230 reduces hallucination. Meanwhile, the iterative
231 procedure inherits the benefit of CoT prompting
232 for eliciting reasoning ability. The main advan-
233 tage over CoT is that at each step we can receive
234 feedback from diverse perspectives due to the dy-
235 namically assigned personas. A comprehensive
236 comparison with previous prompting methods can
237 be found in Table 1.

238 3 Experiments

239 To explore the effectiveness of Solo Performance
240 Prompting (SPP), we adopt an evaluation method-
241 ology similar to that of previous work (Yao et al.,
242 2023). We carefully design new tasks and select
243 tasks from existing benchmarks (Srivastava et al.,
244 2022) that are challenging even for the most capa-
245 ble LLMs (OpenAI, 2023b) using either Standard
246 or Chain-of-Thought prompting. The evaluation
247 aims to cover diverse types of tasks encompassing
248 both *knowledge-intensive* and *reasoning-intensive*
249 domains. The results show that SPP can be a
250 promising new paradigm for general task-solving
251 with state-of-the-art LLMs and reveal an interest-
252 ing finding on the emergence of cognitive synergy
253 ability.

254 **Tasks.** We invent the **Trivia Creative Writing**
255 task (§3.1), which requires the model to internally
256 acquire and integrate diverse information from var-
257 ious fields. We observe that even GPT-4 (OpenAI,
258 2023b) frequently exhibit hallucination and fac-
259 tuality errors in the Trivia Creative Writing task.
260 We also propose the **Codenames Collaborative**
261 task (§3.2), an extension of the Codenames task
262 from the BigBench (Srivastava et al., 2022) that
263 features a two-role collaboration setup. Codenames
264 Collaborative demands creative reasoning across a
265 broad range of related knowledge and challenges
266 the model’s theory of mind skills. Lastly, we in-
267 clude a challenging pure-reasoning task, **Logic**
268 **Grid Puzzle** (§3.3), from the BigBench (Srivastava
269 et al., 2022) which necessitates complex multi-step
270 reasoning.

271 **Prompting Methods.** We compare our approach
272 with **Standard Prompting** and **Chain-of-Thought**
273 (**CoT**) prompting methods (outlined in §2). In CoT,
274 a similar prompt design to (Yao et al., 2023) is em-
275 ployed, where the model is prompted to generate a
276 plan or a series of steps before producing the final
277 output. We examine two variants of Solo Perform-
278 ance Prompting, **SPP** and **SPP-Profile**. Inspired

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%
SPP-Profile (ours)	79.1	+5.9%	83.0	+7.8%	76.7	+1.7%	64.8	+12.4%
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).

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 E.2.

Models. The default model we use is GPT-4 (OpenAI, 2023b). We consider two inference settings: with or without a system message, and report the average scores across both settings in Table 2. Detailed inference configurations and full results can be found in Appendices A 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 (Figure 7). 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 questions² 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 culture³. Figure 7 shows an example instance in Trivia Creative Writing.

²To 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.

³The full prompt for generating the topic list can be found in Figure 18. We performed further human curation to avoid potential harmful content.

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 7, 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 follows:

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

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 6 and 10 illustrate that although CoT generates reasonable plans for task resolution, the final generation still contains factual errors and hallucinations. (2) SPP outperforms both Standard and CoT prompting 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 (Figure 8). Codenames Collaborative is a collaborative task that challenges a model’s knowledge, reasoning, and theory of mind abilities by assigning two player roles: the *Spy-master* 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. A concrete example of the task can be found in Figure 8.

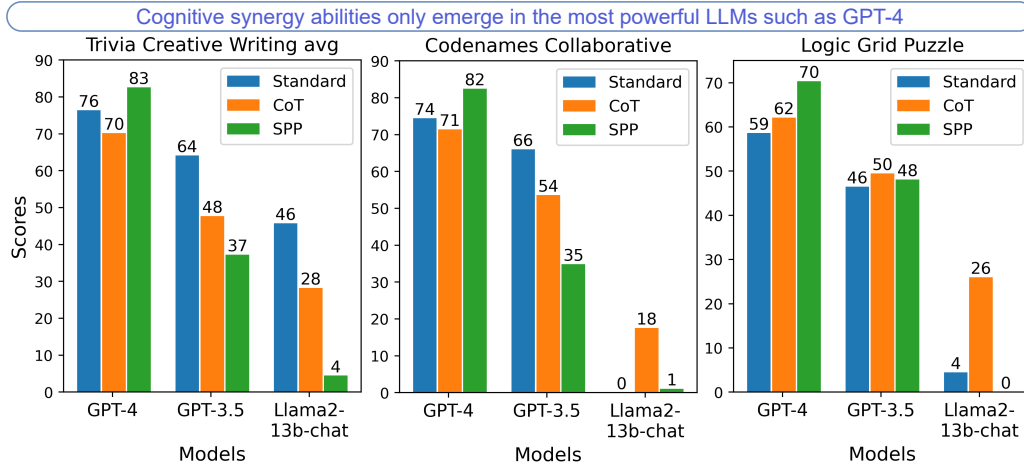


Figure 3: SPP achieves superior performance only with the most powerful LLM (GPT-4), but not with GPT-3.5 and Llama2-13b. This indicates that cognitive synergy abilities only emerge in LLMs with GPT-4 level capabilities.

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 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 8, 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. In contrast, SPP brings significant improvements (~5%), which indicates the effectiveness of the proposed Solo Performance Prompting on collaborative tasks that require knowledge, reasoning, and theory of mind skills. Figure 11 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 (Figure 9). We utilize the Logic Grid Puzzle task from the BigBench (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 9. 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 outperforms CoT on this task, indicating competitive reasoning capabilities on reasoning-intensive tasks. This result also demonstrates that the increased number of personas does not deteriorate the models’ reasoning abilities.

3.4 The Emergence of Cognitive Synergy

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 D.

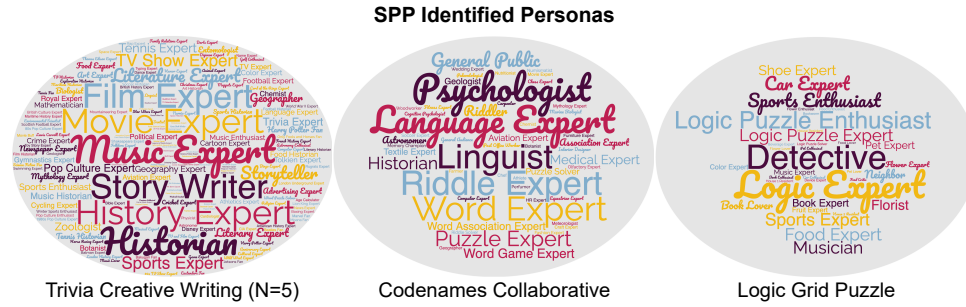


Figure 4: Visualization of the SPP-identified personas for each task. Personas identified in knowledge-intensive tasks like Trivia Creative Writing are more diverse, while those in reasoning-intensive tasks like Logic Grid Puzzle appear more homogeneous.

These findings suggest that the **cognitive synergy ability can only be fully unleashed in LLMs with a certain level of instruction following capabilities, akin to that of GPT-4.**

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-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. Current research on AI agents usually follows a paradigm where one model instance, such as an LLM, plays a specific role. We are interested in investigating whether a single LLM can naturally simulate a multi-agent setup. We visualize the personas automatically identified by SPP using a word cloud for each task in Figure 4, 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." This indicates that current LLMs are inherently capable of identifying useful expert personas for diverse tasks. (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. Furthermore, the results in Table 2 demonstrate 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.

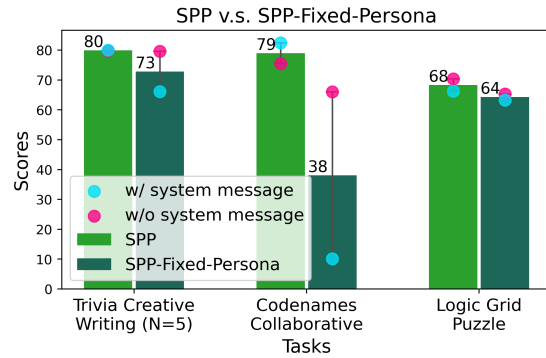


Figure 5: Comparison between SPP (with dynamically identified personas) and SPP-Fixed-Persona (with fixed personas) on GPT-4. The results demonstrate that dynamic, fine-grained personas consistently outperform fixed, general personas.

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 16) to force the personas to be fixed as an "AI Assistant" and an "Expert", while keeping all the information in the demonstration examples intact. Figure 5 shows the comparison between SPP and SPP-Fixed-Persona. Main insights include: (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 6 and 12 shows that the fine-grained personas such as "Film Expert" and "Sports Enthusiast" correctly provide the answers, while the fixed persona "Ex-

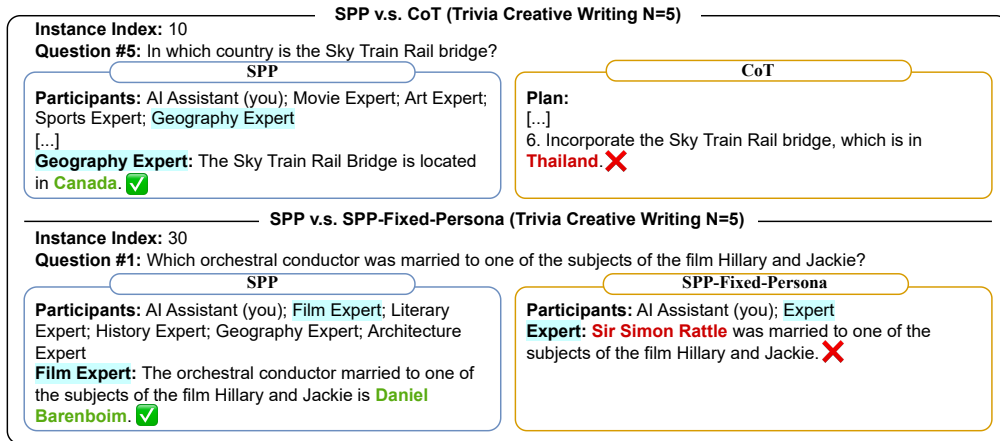


Figure 6: Qualitative examples on Trivia Creative Writing comparing SPP, CoT and SPP-Fixed-Persona. SPP v.s. CoT demonstrates that while CoT provides reasonable steps, it still struggles with hallucination. SPP v.s. SPP-Fixed-Persona reveals that dynamically identified fine-grained personas, such as the "Film Expert," tend to outperform the fixed general persona of an "Expert. More examples can be found in Figures 10, 11, and 12.

480 pert" fails. (2) *SPP-Fixed-Persona* also suffers
 481 from the *early-termination* problem as defined in
 482 §3.4, where the LLM stops collaboration before
 483 providing the final answer as if it were waiting
 484 for external inputs. Details can be found in Ap-
 485 pendix D.

486 5 Related Work

487 **LLMs as role-playing agents.** Recent re-
 488 search (Deshpande et al., 2023; Xu et al., 2023; Fu
 489 et al., 2023; aut, 2023; Li et al., 2023) demon-
 490 strates that assigning personas or roles to LLMs influences
 491 their generation behavior. AI societies with dis-
 492 tinct personas or occupations have been explored
 493 for collaboration (Park et al., 2023; Schick et al.,
 494 2022; Li et al., 2023; Cai et al., 2023). However,
 495 limitations in persona assignment and multi-agent
 496 collaboration include single or fixed persona as-
 497 signments (Xu et al., 2023; Fu et al., 2023; Schick
 498 et al., 2022; Li et al., 2023) and the need for multi-
 499 ple LLM instances, increasing inference cost. We
 500 propose using a single LLM to simulate multi-
 501 persona collaboration, allowing the LLM to dy-
 502 namically identify useful personas for each task
 503 instance. Our finding that such cognitive synergy
 504 ability only emerges in GPT-4 but not in GPT-3.5
 505 and Llama2 also aligns with related work (Olaus-
 506 son et al., 2023), which investigates the emergent
 507 ability of self-debugging in code generation.

508 **Improving reasoning and knowledge acquisition**
 509 **abilities in LLMs.** LLMs face challenges in com-
 510 plex knowledge-intensive tasks due to hallucina-
 511 tion (Maynez et al., 2020) and reasoning-intensive
 512 tasks due to the lack of human-like slow think-

513 ing (Sloman, 1996; Kahneman, 2011). Approaches
 514 like Chain-of-Thought (CoT) and Self-Refinement
 515 encourage LLMs to solve tasks step by step or
 516 iteratively revise their answers (Wei et al., 2023;
 517 Kojima et al., 2022; Zhang et al., 2022; Fu et al.,
 518 2022; Xue et al., 2023; Yao et al., 2023; Madaan
 519 et al., 2023; Shinn et al., 2023; Gou et al., 2023;
 520 Chen et al., 2023; Huang et al., 2022; Yao et al.,
 521 2022). However, these methods do not necessari-
 522 ally reduce hallucination or improve factuality. Re-
 523 trieval augmented LLMs (Borgeaud et al., 2022;
 524 Izacard et al., 2022; Wang et al., 2022; Shuster
 525 et al., 2021) enhance knowledge acquisition but do
 526 not improve reasoning abilities. We propose Solo
 527 Performance Prompting (SPP) to elicit both *inter-*
 528 *nal* knowledge acquisition and reasoning abilities
 529 in LLMs, improving factuality while maintaining
 530 strong performance on pure-reasoning tasks. The
 531 key difference is that SPP dynamically identifies
 532 multiple personas and simulates iterative collabora-
 533 tion to generate intermediate "thoughts".

534 6 Conclusion

535 Solo Performance Prompting (SPP) unleashes the
 536 emergent cognitive synergy abilities within pow-
 537 erful LLMs, significantly reducing hallucination
 538 while enhancing reasoning. The performance is
 539 assessed using newly proposed tasks, e.g., Trivia
 540 Creative Writing and Codenames Collaborative,
 541 demonstrating superior results compared to Stan-
 542 dard and CoT prompting. SPP has the potential
 543 to become a new paradigm for inference in GPT-
 544 4 level LLMs and provides valuable insights for
 545 research on LLM-based autonomous agents.

546 Limitations

547 Although Solo Performance Prompting exhibits
548 promising improvements in acquiring factually cor-
549 rect knowledge compared to Standard prompting,
550 it has some limitations. For instance, even when a
551 fine-grained persona is assigned, the answer may
552 still be incorrect. It remains unclear to what ex-
553 tent assigning a persona can help enhance domain
554 knowledge in a specific area. Dedicated diagnostic
555 experiments and theoretical efforts are needed to
556 quantify the impact of having a persona or not.

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

564 Last but not least, if given sufficient computa-
565 tional budget, a natural variant of SPP could extend
566 to a *multi-agent cognitive synergist* setup where
567 a leader persona identifies several expert agents
568 and forms a cabinet to collaboratively solve a task.
569 The multi-agent setup allows for leveraging richer
570 computation power, larger local memory, and more
571 flexible human-computer interaction, which could
572 be essential for deploying to real-world applica-
573 tions.

574 References

575 2023. Auto-gpt. [https://github.com/Significant-](https://github.com/Significant-Gravitas/Auto-GPT)
576 [Gravitas/Auto-GPT](https://github.com/Significant-Gravitas/Auto-GPT).

577 Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wen-
578 liang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei
579 Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-
580 task, multilingual, multimodal evaluation of chatgpt
581 on reasoning, hallucination, and interactivity. *arXiv*
582 *preprint arXiv:2302.04023*.

583 Sebastian Borgeaud, Arthur Mensch, Jordan Hoff-
584 mann, Trevor Cai, Eliza Rutherford, Katie Mill-
585 can, George Bm Van Den Driessche, Jean-Baptiste
586 Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022.
587 Improving language models by retrieving from tril-
588 lions of tokens. In *International conference on ma-*
589 *chine learning*, pages 2206–2240. PMLR.

590 Sébastien Bubeck, Varun Chandrasekaran, Ronen El-
591 dan, Johannes Gehrke, Eric Horvitz, Ece Kamar,
592 Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lund-
593 berg, et al. 2023. Sparks of artificial general intelli-
594 gence: Early experiments with gpt-4. *arXiv preprint*
595 *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*. 596
597
598

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and
Denny Zhou. 2023. Teaching large language models
to self-debug. *arXiv preprint arXiv:2304.05128*. 599
600
601

Petru L Curşeu, Nicoleta Meslec, Helen Pluut, and Ger-
ardus JM Lucas. 2015. Cognitive synergy in groups
and group-to-individual transfer of decision-making
competencies. *Frontiers in psychology*, 6:1375. 602
603
604
605

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpuro-
hit, Ashwin Kalyan, and Karthik Narasimhan. 2023.
Toxicity in chatgpt: Analyzing persona-assigned lan-
guage models. *arXiv preprint arXiv:2304.05335*. 606
607
608
609

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. 610
611
612
613

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark,
and Tushar Khot. 2022. Complexity-based prompt-
ing for multi-step reasoning. *arXiv preprint*
arXiv:2210.00720. 614
615
616
617

Ben Goertzel. 2009. Cognitive synergy: A universal
principle for feasible general intelligence. In *2009*
8th IEEE International Conference on Cognitive In-
formatics, pages 464–468. IEEE. 618
619
620
621

Ben Goertzel. 2017. A formal model of cognitive syn-
ergy. In *Artificial General Intelligence: 10th Interna-*
tional Conference, AGI 2017, Melbourne, VIC, Aus-
tralia, August 15-18, 2017, Proceedings 10, pages
13–22. Springer. 622
623
624
625
626

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen,
Yuju Yang, Nan Duan, and Weizhu Chen. 2023.
Critic: Large language models can self-correct with
tool-interactive critiquing. 627
628
629
630

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky
Liang, Pete Florence, Andy Zeng, Jonathan Tomp-
son, Igor Mordatch, Yevgen Chebotar, et al. 2022.
Inner monologue: Embodied reasoning through
planning with language models. *arXiv preprint*
arXiv:2207.05608. 631
632
633
634
635
636

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lu-
cas Hosseini, Fabio Petroni, Timo Schick, Jane
Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and
Edouard Grave. 2022. Few-shot learning with re-
trieval augmented language models. *arXiv preprint*
arXiv:2208.03299. 637
638
639
640
641
642

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke
Zettlemoyer. 2017. *TriviaQA: A large scale distantly*
supervised challenge dataset for reading comprehen-
sion. In *Proceedings of the 55th Annual Meeting of*
the Association for Computational Linguistics (Vol-
ume 1: Long Papers), pages 1601–1611, Vancouver,
Canada. Association for Computational Linguistics. 643
644
645
646
647
648
649

650	Daniel Kahneman. 2011. <i>Thinking, fast and slow</i> . macmillan.	701
651		702
652	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>arXiv preprint arXiv:2205.11916</i> .	703
653		704
654		705
655		706
656	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. <i>arXiv preprint arXiv:2303.17760</i> .	707
657		708
658		709
659		710
660		711
661	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. <i>arXiv preprint arXiv:2303.17651</i> .	712
662		713
663		714
664		715
665		716
666	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 1906–1919, Online. Association for Computational Linguistics.	717
667		718
668		719
669		720
670		721
671		722
672	Theo X Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2023. Demystifying gpt self-repair for code generation. <i>arXiv preprint arXiv:2306.09896</i> .	723
673		724
674		725
675		726
676	OpenAI. 2023a. Gpt-35 . https://platform.openai.com/docs/models/gpt-3-5 .	727
677		728
678	OpenAI. 2023b. Gpt-4 technical report .	729
679	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. <i>arXiv preprint arXiv:2304.03442</i> .	730
680		731
681		732
682		733
683		734
684	Anthony D Pellegrini. 2009. <i>The role of play in human development</i> . Oxford University Press, USA.	735
685		736
686	Jean Piaget. 1954. The construction of reality in the child.	737
687		738
688	Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? <i>arXiv preprint arXiv:2302.06476</i> .	739
689		740
690		741
691		742
692	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. <i>arXiv preprint arXiv:2208.11663</i> .	743
693		744
694		745
695		746
696		747
697	Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. <i>arXiv preprint arXiv:2303.11366</i> .	748
698		749
699		750
700		751
		752
		753
		754
		755
		756
		757
		758
		759
		760
		761
		762
		763
		764
		765
		766
		767
		768
		769
		770
		771
		772
		773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
		790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800

A Inference Configurations

The main results in Table 2 are obtained from GPT-4⁴. 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". 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, hyperparameters and the best system message setting in terms of SPP’s GPT-4 performance. For Llama2, we leverage the Huggingface text-generation pipeline⁵ with greedy decoding.

B Task Illustrations

Trivia Creative Writing. Figure 7 shows an example instance in the Trivia Creative Writing task.

Logic Grid Puzzle. Figure 8 shows an example instance in the Codenames Collaborative task.

Codenames Collaborative. Figure 9 shows an example input and output of the Logic Grid Puzzle task.

C Qualitative Analysis

Figure 10 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 11 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 12 shows additional qualitative examples on Solo Performance Prompting vs SPP-Profile.

⁴The GPT-4 API version we employ is Azure 2023-3-15-preview. There are rare cases when a generation triggers the content filter of the API. We exclude those instances from our results.

⁵<https://huggingface.co/blog/llama2>

D Early-termination with SPP-Fixed-Persona

Figure 13 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 5. The problem can be largely alleviated by removing the system message but cannot be entirely eliminated. Table 3 shows the statistics of the early-termination problem for each task and method. In contrast, we did not observe early-termination on SPP, SPP-Profile, Standard, or CoT prompting with GPT-4.

E Prompts

E.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 § E.2.⁶

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,

⁶We use the same prompt for any arbitrary tasks.

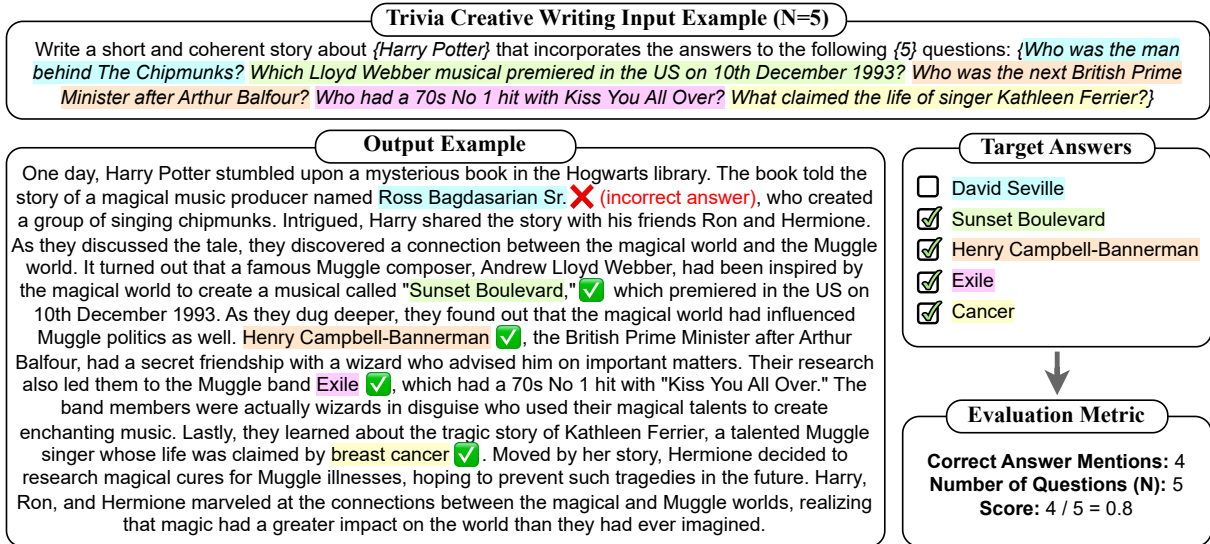


Figure 7: Trivia Creative Writing task example.

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 3: 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.

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.

E.2 Full Prompts

Figures 14, 15 and 16 show the full prompts for SPP, SPP-Profile and SPP-Fixed-Persona respectively. Figure 17 shows the full prompts for Chain-

of-Thought (CoT) prompting.

F Full Results

Full results of the three tasks: Trivia Creative Writing, Codenames Collaborative and Logic Grid Puzzle can be found in Tables 4, 5 and 6, respectively.

843
844
845
846
847
848

849
850
851
852
853
854

855
856
857
858

859
860
861
862
863
864

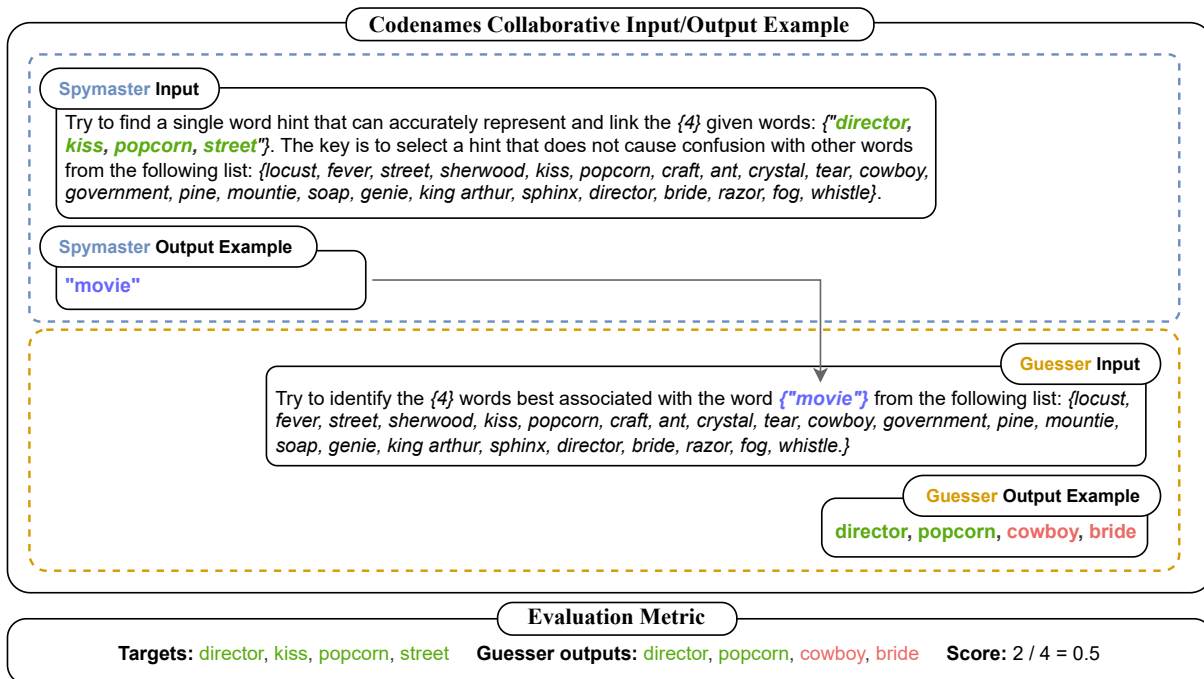


Figure 8: Codenames Collaborative task example.

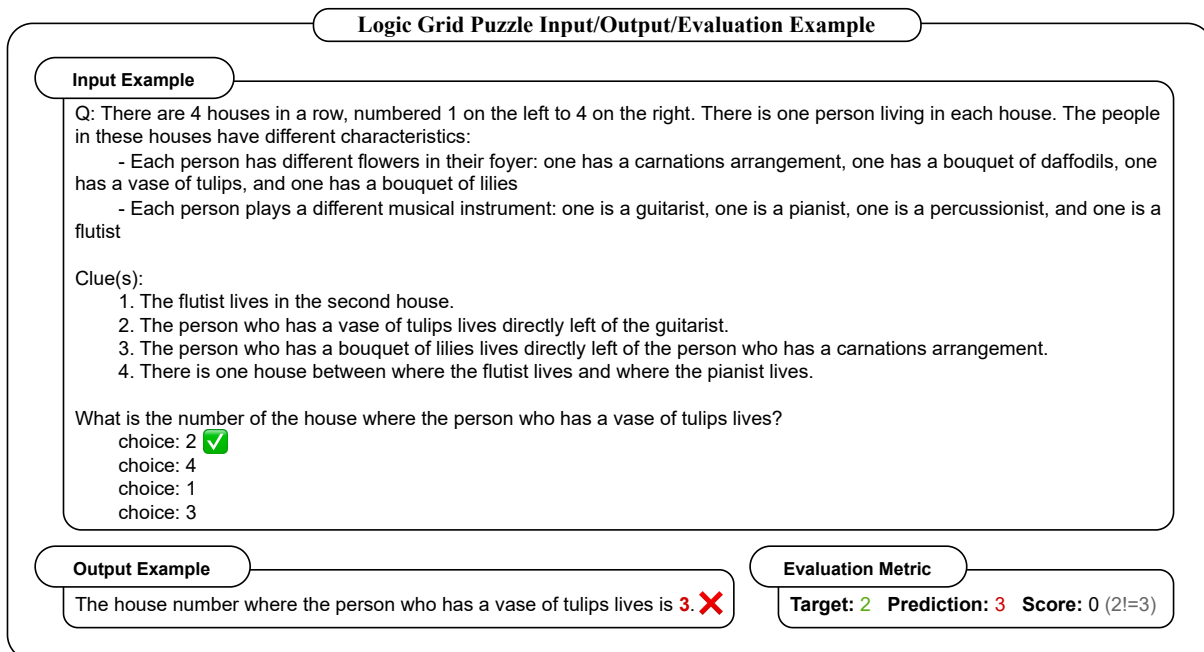


Figure 9: Logic Grid Puzzle task example.

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

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

SPP	CoT
<p>Participants: AI Assistant (you); Movie Expert; Art Expert; Sports Expert; Geography Expert [...]</p> <p>Geography Expert: The Sky Train Rail Bridge is located in Canada. ✓</p>	<p>Plan: [...]</p> <p>6. Incorporate the Sky Train Rail bridge, which is in Thailand. ✗ (factual error)</p>

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

SPP	CoT
<p>Participants: AI Assistant (you); Story Writer; Music Expert; TV Show Expert; Cultural Expert [...]</p> <p>TV Show Expert: [...] The 70s show All in the Family ✓ was based on the British show Till Death Us Do Part.</p>	<p>Plan: [...] 6. Mention a show that the Little Mermaid watches, based on the British show Till Death Us Do Part.</p> <p>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.</p>

Figure 10: 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
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 4: 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



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 11: 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?

<p style="text-align: center;">SPP</p> <p>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. ✓</p>	<p style="text-align: center;">SPP-Fixed-Persona</p> <p>Participants: AI Assistant (you); Expert Expert: Sir Simon Rattle was married to one of the subjects of the film Hillary and Jackie. ✗</p>
--	--

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

<p style="text-align: center;">SPP</p> <p>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. ✓</p>	<p style="text-align: center;">SPP-Fixed-Persona</p> <p>Participants: AI Assistant (you); Expert AI Assistant (you): The musician who founded the Red Hot Peppers is Louis Armstrong. ✗</p>
--	---

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?

<p style="text-align: center;">SPP</p> <p>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. ✓</p>	<p style="text-align: center;">SPP-Fixed-Persona</p> <p>Participants: AI Assistant (you); Expert Expert: Snooker. ✗</p>
---	---

Figure 12: 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
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

Table 5: 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."

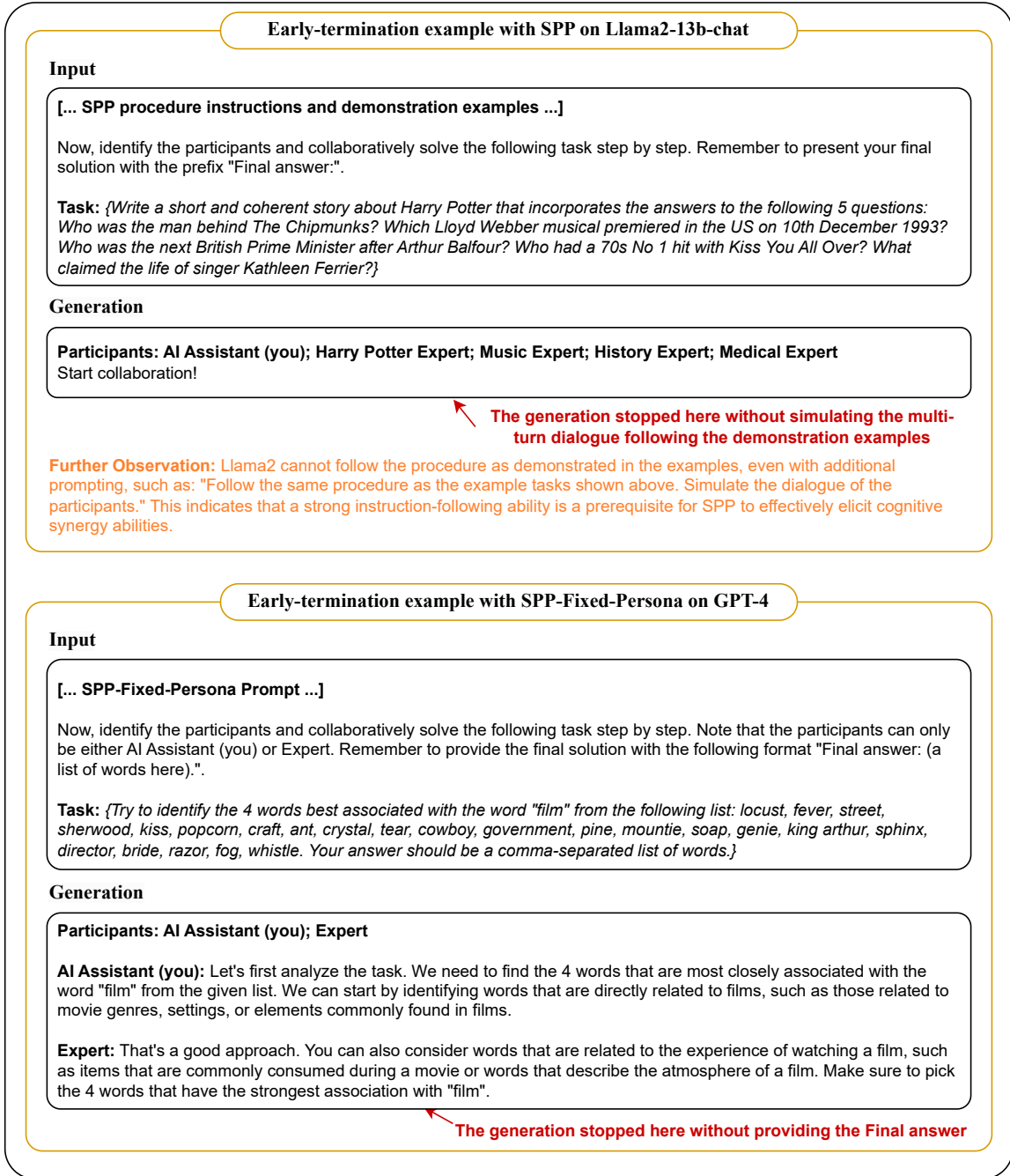


Figure 13: Examples of the early-termination problem with SPP on Llama2-13b-chat and SPP-Fixed-Persona on GPT-4.

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

Table 6: 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 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. $1 + 1 = 2$, $6 * 2 = 12$, $12 + 12 = 24$. 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 CHATGPT. 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 Formatting Instruction)*

Task: *(Task input)*

Figure 14: 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 15: SPP-Profile full prompt. "[...]" indicates identical parts with SPP. Green text indicates the key difference between SPP-Profile and SPP.

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. $1 + 1 = 2$, $6 * 2 = 12$, $12 + 12 = 24$. 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 16: SPP-Fixed-Persona full prompt. Red text indicates the key difference between SPP-Fixed-Persona and SPP.

CoT Prompts

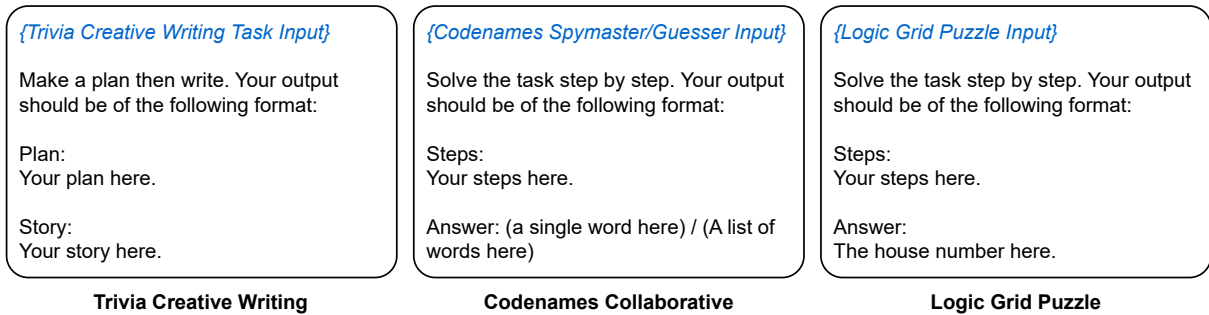


Figure 17: CoT prompts.

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