Self-Evolving GPT: A Lifelong Autonomous Experiential Learner

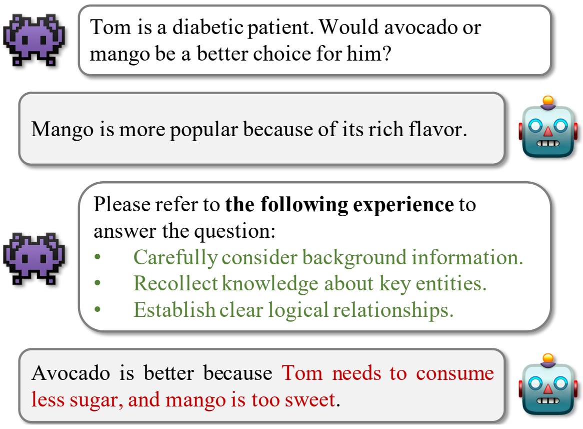
Jinglong Gao[1] **Xiao Ding**[1*] **Yiming Cui**[2] **Jianbai Zhao**[1]
Hepeng Wang[1] **Ting Liu**[1] **Bing Qin**[1]
[1]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
[2]State Key Laboratory of Cognitive Intelligence
iFLYTEK Research, Beijing, China
`{jlgao, xding, jianbaizhao, hpwang, tliu, qinb}@ir.hit.edu.cn`
`ymcui@iflytek.com`

Abstract

To improve the performance of large language models (LLMs), researchers have explored providing LLMs with textual task-solving experience via prompts. However, they rely on manual efforts to acquire and apply such experience for each task, which is not feasible for the growing demand for LLMs and the variety of user questions. To address this issue, we design a lifelong autonomous experiential learning framework based on LLMs to explore whether LLMs can imitate human ability for learning and utilizing experience. It autonomously learns and accumulates experience through experience transfer and induction, categorizing the types of input questions to select which accumulated experience to employ for them. Experimental results on six widely used NLP datasets show that our framework performs reliably in each intermediate step and effectively improves the performance of GPT-3.5 and GPT-4. This validates the feasibility of using LLMs to mimic human experiential learning and application capabilities. Additionally, we provide a detailed analysis of the behavior of our framework at each step.

1 Introduction

Recently, large language models (LLMs) like ChatGPT have achieved excellent performance in various NLP tasks (Kocoń et al., 2023; Ye et al., 2023). However, numerous NLP tasks still cannot be effectively addressed by them (Mao et al., 2023; Chang et al., 2023). This is mainly because they have not accumulated enough experience to handle these tasks during their training.



Tom is a diabetic patient. Would avocado or mango be a better choice for him?

Mango is more popular because of its rich flavor.

Please refer to **the following experience** to answer the question:
- Carefully consider background information.
- Recollect knowledge about key entities.
- Establish clear logical relationships.

Avocado is better because Tom needs to consume less sugar, and mango is too sweet.

Figure 1: An example of experience-enhanced LLMs inference.

To address these issues, previous studies have explored injecting task-solving experience into LLMs during the inference stage via prompts (as shown in Figure 1). Their experience is textual descriptions of the task-solving processes, guidelines, and other insights. Some studies manually craft such experience (Wei et al., 2022; Kong et al., 2023). Others attempt to summarize experience from manually annotated task datasets (Chen et al., 2023a; Zhao et al., 2023; Chen et al., 2024), and then during inference, they essentially need to manually select the experience to apply to each question. However, the demands of users on LLMs are ever-expanding, and the types of user questions continue to grow. These methods would lead to high and unbounded costs for human labor.

In contrast, humans are capable of autonomous learning and utilizing experience. Humans categorize encountered problems into different task types and induce experience from multiple concrete task practices, which are reused when encountering new problems of the same task type (Novak and Gowin, 1984; Cox, 1996). Besides, humans can transfer experience between similar tasks, thus gaining more experience without time-consuming practices (Deese, 1952; Perkins et al., 1992). As lifelong autonomous experience accumulates, humans gradually achieve ability growth. Inspired by this, we want to explore whether LLMs can mimic

---

[*] Corresponding Author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6385–6432
August 11-16, 2024 ©2024 Association for Computational Linguistics

the above process. This could avoid the substantial manual labor and provide a unique evolutionary path for artificial general intelligence.

To facilitate this, we propose a lifelong autonomous experiential learning framework called Self-Evolving GPT (SE-GPT), which consists of a task-specific experience memory and five experience-centric modules based on ChatGPT. For any user question, SE-GPT automatically categorizes the target task type and responds to the question with the target task experience in the memory. For newly encountered task types, it learns experience through experience transfer and induction before responding. Firstly, it locates similar tasks in its memory and transfers their experience to the target task. Then, it autonomously references web information and the transferred experience to practice the target task multiple times, thereby inducing more experience from its successes and failures. Finally, the transferred and induced experience is added to the memory. For tasks encountered previously, it assesses the need for repeating experience transfer and induction before responding, taking into account its proficiency level with the task.

To conduct experiments, we provide a basic implementation [1] of our framework. We mainly focus on the overall framework and aim to analyze its effectiveness and behavior. Experiments show that our framework is practically feasible. It effectively improves the average performance of GPT-3.5 and GPT-4 on six widely used datasets by 3.8% and 5.3%, respectively. Our framework reliably executes each intermediate module, achieving consistent performance improvements. Besides, we provide a detailed analysis of the behavior of our framework in each intermediate step.

2 Related Work

2.1 Autonomous Experiential Learning

To improve the performance of LLMs, researchers provide textual experience to LLMs through prompts. Early studies primarily involve manually crafting such experiential prompts (Wei et al., 2022; Kong et al., 2023), while more recent work focuses on utilizing the LLMs themselves to obtain task-solving experience automatically.

Some studies focus on how to guide LLMs to automatically summarize experience based on interactive environments. Chen et al. (2023a) guided LLMs to summarize cooking skills in a cooking simulation game. Wang et al. (2023) and Zhu et al. (2023) built LLM-based frameworks in the game “Minecraft” to autonomously learn to complete various game targets. Park et al. (2023) created a sandbox environment similar to “The Sims” to guide LLMs in learning role-playing skills. Both Wen et al. (2023) and Fu et al. (2024) taught LLMs how to perform autonomous driving in a simulated driving environment.

All of these studies guide LLMs to learn experience based on explicit feedback from environments, which is inaccessible for most NLP tasks. Besides, they require human labor to create the environment or develop feedback-reading methods.

For NLP tasks, Zhao et al. (2023) and Chen et al. (2024) leveraged ChatGPT to automatically summarize experience from manually annotated NLP datasets. Zhao et al. (2023) employed Reflexion (Shinn et al., 2023) to generate reasoning chains for each question. Then, the experience is summarized from the questions, chains, and human-annotated labels by ChatGPT. They also found that ChatGPT could transfer the summarized experience from the HotpotQA (Yang et al., 2018) dataset to the FEVER (Thorne et al., 2018) dataset. Chen et al. (2024) analyzed the impact of different examples and prompts on the quality of the summarized experience.

However, these methods still require human labor to obtain experience and determine which experience to employ for the current question. In contrast, our framework autonomously learns and selects experience, saving many human labor costs.

2.2 Unsupervised In-Context Learning

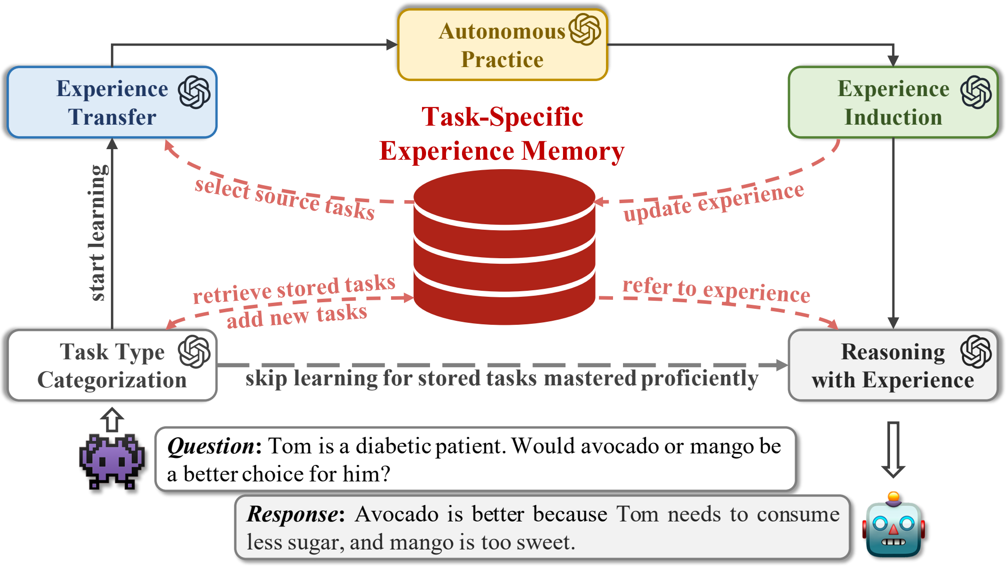
In-Context Learning (ICL) provides demonstrations to LLMs, which can be regarded as a specific substitute for textual experience. Therefore, we introduce the recent work on unsupervised ICL.

Several studies aim at predicting labels with LLMs for unlabeled questions, yielding demonstrations (Li and Qiu, 2023; Wan et al., 2023; Zhang et al., 2023). However, these studies still necessitate manual effort for the generation of questions. Therefore, Lyu et al. (2023) directly leveraged retrieved web texts as unlabeled questions, which is only suitable for specific task datasets. In contrast, our framework is task-agnostic and designed to operate autonomously.

Furthermore, several studies employed LLMs to generate entire demonstrations (Kim et al., 2022; Yu et al., 2023; Chen et al., 2023b). SG-ICL (Kim

---

[1] The code is available in: `https://github.com/ArrogantL/se_gpt`



Question: Tom is a diabetic patient. Would avocado or mango be a better choice for him?

Response: Avocado is better because Tom needs to consume less sugar, and mango is too sweet.

Figure 2: The framework of our proposed Self-Evolving GPT. The lines connected to the memory indicate the flow of information stored in memory. Other lines with arrows represent the execution sequence of our framework.

et al., 2022) requires the development set for selecting demonstrations, while TP-ICL (Yu et al., 2023) is designed explicitly for complex reasoning tasks like shortest-path reasoning, and Self-ICL (Chen et al., 2023b) is the general-purpose one. These demonstrations suffer from issues such as incorrect formatting, noise, and low diversity. However, our framework utilizes the general insights summarized from multiple demonstrations, which is more reliable than the demonstrations themselves.

3 Methodology

Figure 2 shows the framework of our proposed Self-Evolving GPT, which consists of one task-specific experience memory and five experience-centric modules based on ChatGPT. Our framework continuously receives various user questions. It automatically categorizes the task type of the question, and adds it to memory if it is a new task not yet stored. For tasks that are not proficiently mastered, it performs experience transfer, autonomous practice, and experience induction to update their experience in memory. Finally, it refers to experience stored in memory to respond the user question.

In practice, we provide a basic implementation of our framework, which may be further optimized. We primarily focus on the overall framework, and aim to analyze its effectiveness and behavior. **The prompts and execution examples of our implementation are presented in Appendix D and E.**

3.1 Task-Specific Experience Memory

We utilize an external memory to store the task-specific textual experience that our framework autonomously learns. This memory starts empty and gradually grows as our framework runs, assisting it in task-solving and learning new experience.

Specifically, we store each task in the memory with its name, description and experience. For the completeness of experience, our memory stores two types of experience for each task: 1) **Procedure**: the specific steps for handling the task; 2) **Suggestions**: how to better accomplish the task or avoid low-quality responses. These task names, descriptions, and experience are all autonomously generated by our framework.

3.2 Task Type Categorization

Users may pose various questions to the framework, corresponding to unpredictable task types. Therefore, we employ this module to first autonomously categorize the task type of each user question.

The operation of this module is divided into three steps: 1) ChatGPT utilizes Prompt 1 to generate the task name and description based on the question; 2) we retrieve the top 5 tasks from memory that are semantically most similar to the generated task description; 3) finally, ChatGPT utilizes Prompt 2 to select which one of the five tasks is identical to the generated task. If a match is found, the question is linked to the selected task; otherwise, it is linked to the generated task, and we add the generated task into the memory with empty initial task experience.

Please note that the word “task” in our framework represents a ChatGPT-generated task rather than a classic NLP task (e.g., sentiment analysis) in a certain predefined task list.

After this, we retrieve the experience of the current task from memory, and denote it as $\mathbf{E}_{\text{mem}}$. Then, we assess whether the current task has been adequately learned following our skip learning condition (§3.6). If it has, we respond to the user question with $\mathbf{E}_{\text{mem}}$ following our final reasoning prompt (§3.7); otherwise, we learn experience following our experience transfer module (§3.3), autonomous practice module §3.4 and experience induction module §3.5.

3.3 Experience Transfer

Experience from similar tasks often exhibits transferability (Deese, 1952; Perkins et al., 1992). Therefore, we employ this module to transfer the experience of other tasks in memory to the current task.

This module is orchestrated through four fundamental steps: 1) we retrieve the top 10 tasks from memory that are semantically most similar to the target task description; 2) if the previous step outputs at least one candidate task, ChatGPT utilizes Prompt 3 to select which among the 10 tasks should be chosen as source tasks for the transfer; 3) if the previous step outputs at least one source task, ChatGPT utilizes Prompt 4 to facilitate a step-by-step experience transfer process. It begins by understanding the differences between the source and target tasks, then identifying shared general experience between them, and finally rephrasing the general experience in the context of the target task. We denote such experience as $\mathbf{E}_{\text{transferred}}$; 4) if $\mathbf{E}_{\text{mem}}$ is not empty, ChatGPT utilizes Prompt 5 to merge $\mathbf{E}_{\text{transferred}}$ and $\mathbf{E}_{\text{mem}}$ for updating $\mathbf{E}_{\text{transferred}}$. If steps 1 and 2 fail to select any source tasks, $\mathbf{E}_{\text{mem}}$ is employed as $\mathbf{E}_{\text{transferred}}$.

3.4 Autonomous Practice

Humans can autonomously practice tasks and derive experience from practice instances. Therefore, we employ this module to mimic the process of human autonomous practice. For the current target task, it automatically generates multiple examples, including questions, responses, and labels indicating whether the responses are correct. Additionally, it utilizes the transferred experience and the autonomously retrieved web information to provide references for its practice process.

This module performs autonomous practice step by step: 1) we retrieve web documents that are semantically most related to the user question; 2) ChatGPT utilizes Prompt 6 to reference one of the retrieved web documents, the user question, and the task description generated in §3.2 to generate a new question; 3) ChatGPT utilizes Prompt 7 to respond to the generated new question with $\mathbf{E}_{\text{transferred}}$; 4) ChatGPT utilizes Prompt 8 to reference the web document in the second step for verifying the correctness of its responses. We repeat the above steps to obtain five examples for the current task.

3.5 Experience Induction

After the autonomous practice, we summarize new experience for the current task from examples generated in §3.4 with correct or incorrect answers.

In practice, we utilize Prompt 9 to guide ChatGPT in summarizing experience step-by-step. ChatGPT first summarizes the commonalities in the correct examples, identifying patterns in the incorrect examples, and compares the differences between the correct and incorrect examples. Then, based on these observations and analysis, ChatGPT tries to summarize task-solving insights generally applicable to unseen examples of the current task. We denote such experience as $\mathbf{E}_{\text{induced}}$. After that, if $\mathbf{E}_{\text{transferred}}$ is not empty, we utilize Prompt 5 to merge $\mathbf{E}_{\text{induced}}$ and $\mathbf{E}_{\text{transferred}}$ for updating $\mathbf{E}_{\text{induced}}$.

Finally, we replace $\mathbf{E}_{\text{mem}}$ in memory as $\mathbf{E}_{\text{induced}}$, which has been enhanced through experience transfer, autonomous practice and experience induction.

3.6 Learning or Skip Learning

The tasks that our framework has already adequately learned do not require further learning. It is inefficient to repeat learning for each user question.

Implementation-wise, our memory records the number of incorrect examples during each autonomous practice stage. If the number of incorrect examples remains zero three times for the same task, we consider that such task has already been adequately learned, and further learning is skipped.

Although we provide a basic skip condition, it may be modified for different preferences for efficiency and experience quality.

3.7 Reasoning with Experience

Finally, we utilize Prompt 10 to guide ChatGPT in responding to the user question with the experience of the current task in memory. For tasks that require further learning, the experience stored in memory has been enriched through experience transfer, autonomous practice, and experience induction.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on the mixture of the following six widely used NLP datasets, including: 1) MMLU (Hendrycks et al., 2021), which is a massive multitask test consisting of multiple-choice questions from various branches of knowledge, covering 57 tasks; 2) e-CARE (Du et al., 2022), which is a causal reasoning dataset that requires determining which option is the cause or result of a given event from various domains; 3) SocialIQA (Sap et al., 2019), which is a social commonsense test that focuses on reasoning about people's actions and their social implications in various social situations; 4) WinoGrande (Sakaguchi et al., 2021), which is a robust commonsense reasoning dataset formulated as a fill-in-a-blank task with binary options; 5) HELP (Yanaka et al., 2019), which is a natural language inference dataset that focuses on logical inferences licensed by phrase replacements, so-called monotonicity reasoning; 6) LogiQA-2 (Liu et al., 2023), which is sourced from expert-written questions for testing civil servants, covering multiple types of deductive reasoning.

We randomly select K data points from each dataset and mix them randomly as the test dataset. The test dataset includes human annotated labels, which are only used for evaluating performance. For GPT-3.5, K=1,000, resulting in a final experimental data size of 6,000. For GPT-4, K=500, resulting in a final experimental data size of 3,000. We adopt accuracy (Acc) as the evaluation metric and report the average accuracy of three rounds of predictions to reduce randomness. For the human evaluation in our experiments, three evaluators are asked to perform annotations.

4.2 Parameters Setting

We conduct experiments using OpenAI's official API[2] with two versions of ChatGPT separately, including `gpt-3.5-turbo-1106` (GPT-3.5) and `gpt-4-1106-preview` (GPT-4). Moreover, `temperature` is fixed as 1. The retrieval operations in §3.2, §3.3 and §3.4 are accomplished by the Faiss index (Johnson et al., 2021). For the stability of Prompt 2 and 8, we run them multiple times until one option is output twice, and then we select this option as the final output. The web texts in §3.4 are retrieved from Wikipedia and truncated to 512 tokens. If Prompt 8 outputs "inconclusive" for a generated question-answer pair, we discard it.

4.3 Baselines

In our experiments, we employ the following baseline methods: 1) **Zero-shot**, we directly feed the input question into ChatGPT; 2) **Zero-shot-CoT**, we add "Let's think step by step" at the end of each input question and then feed it into ChatGPT; 3) **Self-EXP**, we first utilize Prompt 11 to instruct ChatGPT to directly generate **exp**erience for each input question. Then, just like our framework, we utilize Prompt 10 to guide ChatGPT in responding to each input question with the experience generated for it; 4) **Self-ICL** (Chen et al., 2023b), which first prompts ChatGPT to generate new questions following the input question. Subsequently, ChatGPT predicts pseudo-labels for the new questions via zero-shot prompting. Finally, it performs ICL for the input question with the pseudo-question-label pairs as demonstrations; 5) **Self-ICL-CoT** (Chen et al., 2023b), which is a Chain-of-Thought-based variation of **Self-ICL**. It adds "Let's think step by step" at the end of new questions and the input question before predicting them. We faithfully replicated the methods of Chen et al. (2023b) according to their origin paper; 6) **Modified Self-ICL**, from the test dataset, we retrieve the top 5 examples with the highest semantic similarity for each test example, to replace the generated input question in the self-ICL; 7) **AutoP-ICL**, employs demonstrations generated by our autonomous practice module (§3.4) to perform in-context learning. Specifically, the pairs (new question, reasoning process) deemed correct by our auto practice module are concatenated with the user query as the prompt for LLMs.

4.4 Main Results

Table 1 shows the results on the mixture of six NLP datasets. We find that:

Firstly, our SE-GPT achieves consistently better performance than baseline methods and improves the average performance of zero-shot GPT-3.5 and GPT-4 by 3.8% and 5.3%, respectively. This is because our framework can effectively learn task-solving experience and select appropriate experience for the input question.

Secondly, across all datasets, our framework shows the most significant gains over zero-shot

---

[2] `https://platform.openai.com/`

Model	Method	MMLU	e-CARE	SocialIQA	WinoGrande	HELP	LogiQA-2	*Average*
GPT-3.5	**Zero-shot**	0.670	0.813	0.754	0.679	0.502	0.516	0.656
	Zero-shot-CoT	0.666	0.802	0.751	0.675	0.516	0.522	0.655
	Self-EXP	0.673	0.773	0.712	0.658	0.509	0.515	0.640
	Self-ICL	0.621	0.728	0.693	0.604	0.494	0.349	0.582
	Self-ICL-CoT	0.615	0.742	0.696	0.619	0.507	0.350	0.588
	Modified Self-ICL	0.655	0.814	0.746	0.674	0.534	0.510	0.656
	AutoP-ICL	0.652	0.799	0.735	0.650	0.504	0.422	0.627
	SE-GPT (Ours)	**0.708**	**0.857**	**0.792**	**0.693**	**0.557**	**0.556**	**0.694**
GPT-4	**Zero-shot**	0.796	0.828	0.788	0.812	0.608	0.706	0.756
	Zero-shot-CoT	0.822	0.830	0.805	0.833	0.628	0.686	0.767
	Self-EXP	0.834	0.846	0.808	0.828	0.646	0.698	0.777
	Self-ICL	0.732	0.808	0.740	0.795	0.649	0.651	0.729
	Self-ICL-CoT	0.788	0.820	0.734	0.826	0.655	0.607	0.738
	SE-GPT (Ours)	**0.850**	**0.869**	**0.835**	**0.848**	**0.690**	**0.761**	**0.809**

Table 1: Experimental results (%) on the mixture of six datasets. **Bold** and Underlined numbers represent the 1st and the 2nd best performance of two versions of ChatGPT on each dataset. "Average" denotes the mean accuracy across different datasets for each method.

Model	Method	MMLU	e-CARE	SocialIQA	WinoGrande	HELP	LogiQA-2	*Average*
GPT-3.5	**SE-GPT (Ours)**	**0.708**	**0.857**	**0.792**	**0.693**	**0.557**	**0.556**	**0.694**
	- w/o transfer	0.697	0.843	0.771	0.689	0.535	0.541	0.679
	- w/o induction	0.703	0.851	0.779	0.678	0.542	0.547	0.683
GPT-4	**SE-GPT (Ours)**	**0.850**	**0.869**	**0.835**	**0.848**	**0.690**	**0.761**	**0.809**
	- w/o transfer	0.841	0.853	0.827	0.838	0.673	0.744	0.796
	- w/o induction	0.846	0.859	0.819	0.841	0.683	0.756	0.801

Table 2: Performance (%) of our framework with/without experience transfer and induction.

Model	Method	Acc	Experience		
			Sug.	**Pro.**	**All**
GPT-3.5	**SE-GPT (Ours)**	0.998	7.8	6.2	14.0
	- w/o transfer	0.999	5.0	4.6	9.5
	- w/o induction	1.000	7.0	5.8	12.7
GPT-4	**SE-GPT (Ours)**	0.998	11.5	10.4	21.9
	- w/o transfer	0.999	8.2	7.4	15.6
	- w/o induction	0.999	9.2	8.4	17.6

Table 3: The statistics and human-evaluated accuracy (%) of experience of our framework with/without experience transfer and induction. We report the average number of insights for experience across all tasks in our memory at the end of the runtime. "Sug." means the suggestions. "Pro." means the procedure. "All" means both of them.

GPT-3.5 and GPT-4 on the HELP dataset, with improvements of 5.5% and 8.2%, respectively. The reason may be that zero-shot ChatGPT performs worst on HELP, and additional guidance is more helpful for questions that the ChatGPT itself is not good at.

Thirdly, the performance of Self-EXP is unstable. This is due to the quality of the experience it generates is unreliable, with errors, irrelevant information, or insights that LLMs cannot follow. We conduct a case study in Appendix B. The powerful capabilities of GPT-4 alleviate this issue. However, our approach summarizes experience by observing patterns across specific examples and transferring shared insights from multiple source tasks to the target task. This allows our framework to learn highly task-relevant and more general experience.

Besides, the demonstrations generated by Self-ICL and Self-ICL-CoT cannot effectively enhance the performance of ChatGPT. There are mainly three reasons: 1) ChatGPT often generates new questions that are inconsistent with the format of the example; 2) there are errors in the reasoning chains and pseudo-labels predicted by ChatGPT; 3) new questions directly generated by ChatGPT may be simple and lack diversity. We conduct a case study on them in the Appendix B. However, by referencing web texts, our SE-GPT improves the diversity of generated questions and verifies the correctness of responses. Additionally, we do not directly use specific examples for inference but extract general patterns from them, reducing the impact of noise.

Additionally, our framework outperforms the

Modified Self-ICL. This is because we do not directly use specific demonstrations but summarize task-solving insights from them, reducing the impact of noise and providing more direct guidance.

Moreover, according to the results of AutoP-ICL, the performance gains of our framework is not largely due to web retrieval. In our framework, web texts are only utilized in the auto practice module. Web retrieval aids in checking the correctness of practice and provides necessary guiding signals for lifelong learning, but these signals cannot be directly applied to solving user queries. Our experience induction module further summarizes task-solving experiences from multiple practices, while the experience transfer module enables these experiences to assist with other similar tasks.

Furthermore, baseline methods need to generate demonstrations or experience for each question. However, our SE-GPT reuses the learned experience across different questions, resembling human thought processes.

4.5 Effect of the Experience Transfer and Induction

As shown in Table 2 and Table 3, we analyze the variations of our framework with/without the experience transfer and the experience induction module: 1) "- w/o transfer", directly skips the experience transfer module of our framework; 2) "- w/o induction", skips the experience induction module after $1/3$ of all test data in our experiments, i.e., 2,000 for GPT-3.5 and 1,000 for GPT-4. Please note that our framework learns from the test data (only their inputs and not their labels) as it proceeds to the next instance. In human evaluation, we randomly select the experience of 100 tasks from memory and then identify insights that are incorrect, unrelated to the tasks, or cannot be followed by LLMs to report the "Acc". We find that:

Firstly, both experience transfer and induction contribute to the performance and the experience quantity of the overall framework. This is mainly because they can acquire experience for the target task by transferring from other tasks or summarizing from multiple examples, respectively.

Secondly, "- w/o induction" maintains an acceptable level of performance. This indicates that after running for some time, our framework can still achieve consistent improvement only through experience transfer, which is more cost-effective than experience induction.

Besides, our framework can generate high-quality experience. This arises from the fact that our framework references web texts to generate low-noise examples for summarizing experience, and leverage shared insights from multiple source tasks to obtain more reliable experience.

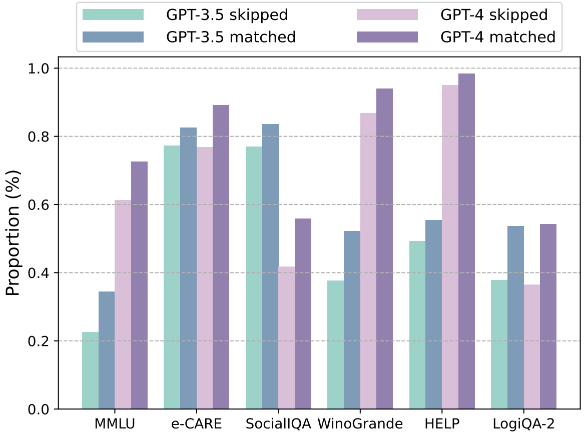


Figure 3: The proportion (%) of the questions that match existing tasks in memory or skip the learning process.

4.6 Analysis of the Task Type Categorization

Human Evaluation of Categorizing Task Types. Task type categorization is the first module of our framework and critically influences the performance of subsequent modules. Table 4 shows the human-evaluated accuracy of our task type categorization module. For each dataset, we randomly evaluate 100 questions linked to newly generated tasks and 100 questions matched to tasks in memory. Accuracy on all data is reported as the weighted accuracy average for both. We find that ChatGPT performs very well in this stage. This is mainly due to it is not a difficult task, and we provide a reasonable prompt for ChatGPT.

Proportion of Matched and Skipped Questions. Figure 3 shows the proportion of the input questions that are matched to tasks in memory or skip the learning process. These proportions determine the efficiency of our framework in utilizing stored experience without the need to repeat the experiential learning process for each question. We find that: 1) compared to GPT-3.5, more questions are matched and skipped by GPT-4. The main reason is the stronger capabilities of GPT-4, allowing it to better recognize learned tasks and meet the skipping criteria in §3.6; 2) the trends in SocialIQA are opposite to those in other datasets. This may arise from the differences of ChatGPT in the prior knowledge and biases of task categorizing.

Model	Type	MMLU	e-CARE	SocialIQA	WinoGrande	HELP	LogiQA-2
GPT-3.5	**Generated Task**	0.99	0.98	0.98	1.00	1.00	0.99
	Matched Task	0.94	0.92	0.96	0.97	1.00	0.94
	All Task	0.97	0.93	0.96	0.98	1.00	0.96
GPT-4	**Generated Task**	0.99	1.00	1.00	1.00	1.00	1.00
	Matched Task	1.00	1.00	0.99	1.00	1.00	1.00
	All Task	0.99	1.00	0.99	1.00	1.00	1.00

Table 4: Human evaluation (%) of the task type categorization module.

Model	MMLU	e-CARE	SocialIQA	WinoGrande	HELP	LogiQA-2
GPT-3.5	0.719	0.960	0.969	0.995	1.000	0.846
GPT-4	0.968	0.998	1.000	0.990	0.978	0.982

Table 5: Human evaluation (%) of source task selection.

4.7 Analysis of the Experience Transfer

Human Evaluation of Selecting Source Tasks. Table 5 shows the human-evaluated accuracy of our source task selection process. For each dataset, we randomly evaluate 100 target tasks, leading to 2,825 source-target task pairs. We find that: 1) overall, ChatGPT performs well in selecting source tasks. This is mainly because recognizing similarity is not a difficult task; 2) the accuracy on MMLU is relatively low. This might arise from the diverse types of tasks in MMLU and its low similarity with other datasets. However, our framework still achieves improvements on MMLU. This is due to we identify shared insights among multiple source tasks, excluding non-transferable insights.

Number of Source Tasks Varying with Runtime. Figure 4 shows the average number of source tasks of each input task varying with runtime. The operating round refer to the number of test questions processed by our framework. As the operating rounds increase, our framework can utilize more source tasks. The main reason is the increasing types of tasks in memory. This also implies that our framework could continually enhance its transfer ability, benefiting from lifelong learning.

4.8 Analysis of the Autonomous Practice

As shown in Table 6, we analyze the performance of the autonomous practice module with/without reference web texts. We randomly selected 300 generated examples and manually evaluate whether the validation results are correct. Besides, we report the diversity of new questions generated per input question. We find that by referencing web texts, our framework significantly improves both the validation accuracy and the diversity of generated questions. This is because: 1) the differences in reference texts lead to variations in generating questions; 2) the texts referenced by question generation usually contain question-solving information.

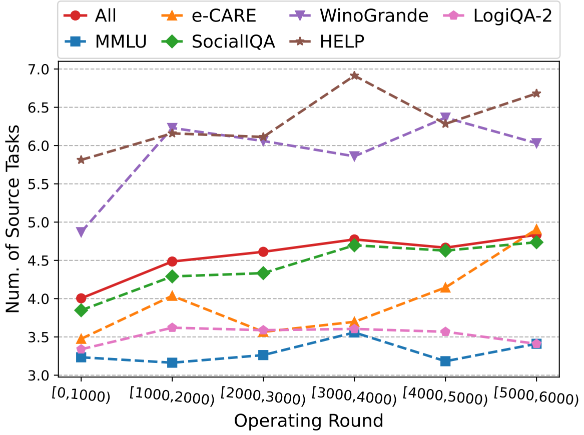


Figure 4: The average number of source tasks chosen per target task for experience transfer in each dataset during the execution of our SE-GPT based on GPT-3.5.

4.9 Analysis of the Experience Induction

As shown in Figure 5, we repeatedly perform the autonomous practice and the experience induction module, reporting the number of generated insights. We randomly select 100 questions and employ GPT-3.5 for the test. We find that the experience increases with each round and stabilizes at the 8th round. This is because as the quantity of experience increases, the difficulty of acquiring new experience grows. Through case observation, we find that almost all of the insights obtained in round 9 are included in the experience obtained previously.

Method	Verify	Generated Question		
		Num	Dist-1	Dist-2
Ours	0.877	5	0.31	0.51
		20	0.12	0.25
- w/o reference	0.797	5	0.24	0.41
		20	0.03	0.08

Table 6: Performance of the autonomous practice module with GPT-3.5. “Verify” shows the human-evaluated accuracy (%) of the validation step (Prompt 8). “Num” is the count of questions generated per input question. “Dist-n” is the ratio of distinct N-grams to total N-grams in the generated questions per input question.

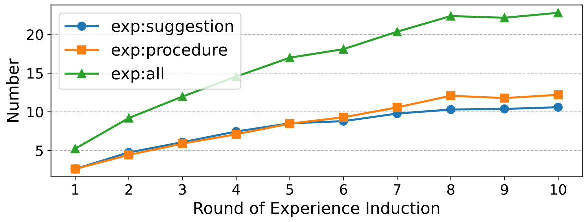


Figure 5: The number of insights generated by multi-round experience induction.

5 Conclusion & Future Work

In this paper, we propose a lifelong autonomous experiential learning framework based on LLMs. It continuously and autonomously accumulates experience in solving tasks through experience transfer and induction, recognizing the nature of input questions to align them with relevant experience. Considering the increasing demand for LLMs and the emergence of new types of user questions, our framework effectively reduces the human labor associated with previous methods. Experiments show that the implementation of our framework can reliably execute each intermediate module and effectively enhance overall performance for responding to the input question. The following content may be subject to our research in future work: 1) **Enhanced engineering designs.** We only offer a basic implementation for our framework, and there is still room for improvement, e.g., supporting more complex functions; 2) **Cold start.** At present, we run our framework completely from empty memory. However, the existing manually annotated datasets can be used to replace the autonomous practice module. Our framework can first learn from the manually annotated datasets, complete the cold start, and then run independently; 3) **Employing a combination of different-scaled LLMs to implement the framework.** It is evident that not all tasks necessitate using ChatGPT; integrating LLMs of various scales can achieve a balance between cost and performance; 4) **Experience Distillation.** Distilling the rich experience summarized by GPT-4 onto smaller-scale LLMs to enhance their performance on tasks that have been adequately learned by GPT-4.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Natural Science Foundation of China (U22B2059, 62176079), and the Natural Science Foundation of Heilongjiang Province (YQ2022F005).

Limitations

In this work, we design a framework to validate the feasibility of using LLMs to mimic human experiential learning and application capabilities. However, it is a basic implementation for experimental exploration but not a perfect LLM product, with room for improvement: 1) **Experience Failure and Operating Error:** Even with high-quality experience, LLMs may still make mistakes. Common errors we observed include reasoning errors/hallucination, LLMs disregarding partial experience, and LLMs lacking necessary knowledge to solve problems. Besides, the steps such as auto practice, experience induction and transfer are complex, and there still remains some noise in the obtained experience; 2) **Both Computationally and Financially Expensive:** the system repeatedly invokes an LLM, which is quite expensive both computationally and also financially. In §A.4, we carefully discuss our prompt cost and possible methods to reduce the cost. 3) **Task Applicability:** Experience may still be effective in tasks requiring skills such as mathematical reasoning, but it might not be as effective for tasks relying on factual knowledge such as WikiQA. Therefore, the framework should have the ability to adaptively determine whether past experience is needed;

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Ding Chen, Shichao Song, Qingchen Yu, Zhiyu Li, Wenjin Wang, Feiyu Xiong, and Bo Tang. 2024. Grimoire is all you need for enhancing large language models.

Liting Chen, Lu Wang, Hang Dong, Yali Du, Jie Yan, Fangkai Yang, Shuang Li, Pu Zhao, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023a. Introspective tips: Large language model for in-context decision making.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023b. Self-ICL: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662, Singapore. Association for Computational Linguistics.

Michael Thomas Cox. 1996. *Introspective multistrategy learning: Constructing a learning strategy under reasoning failure*. Georgia Institute of Technology.

James Deese. 1952. The psychology of learning.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. 2024. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 910–919.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *ArXiv*, abs/2206.08082.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting.

Xiaonan Li and Xipeng Qiu. 2023. Mot: Memory-of-thought enables chatgpt to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2304–2317, Toronto, Canada. Association for Computational Linguistics.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4.

Joseph D Novak and D Bob Gowin. 1984. *Learning how to learn*. cambridge University press.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. UIST '23, New York, NY, USA. Association for Computing Machinery.

David N Perkins, Gavriel Salomon, et al. 1992. Transfer of learning. *International encyclopedia of education*, 2:6452–6457.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. 2023. Dilu: A knowledge-driven approach to autonomous driving with large language models.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

Junchi Yu, Ran He, and Rex Ying. 2023. Thought propagation: An analogical approach to complex reasoning with large language models.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2023. Expel: Llm agents are experiential learners.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory.

Appendix

Contents

1 Introduction 1

2 Related Work 2
  2.1 Autonomous Experiential Learning 2
  2.2 Unsupervised In-Context Learning 2

3 Methodology 3
  3.1 Task-Specific Experience Memory 3
  3.2 Task Type Categorization 3
  3.3 Experience Transfer 4
  3.4 Autonomous Practice 4
  3.5 Experience Induction 4
  3.6 Learning or Skip Learning 4
  3.7 Reasoning with Experience 4

4 Experiments 5
  4.1 Datasets and Evaluation Metrics . 5
  4.2 Parameters Setting 5
  4.3 Baselines 5
  4.4 Main Results 5
  4.5 Effect of the Experience Transfer and Induction 7
  4.6 Analysis of the Task Type Categorization 7
  4.7 Analysis of the Experience Transfer 8
  4.8 Analysis of the Autonomous Practice 8
  4.9 Analysis of the Experience Induction 8

5 Conclusion & Future Work 9

A Additional Experimental Analysis 13
  A.1 Number of Source Tasks Varying with Runtime based on GPT-4 . . 13
  A.2 Number of Tasks and Experience in the Memory Varying with Runtime 13
  A.3 Performance of Experience Induction Through More Examples . . . 13
  A.4 Prompt Cost 13

B Case Study 15
  B.1 Experience Generated by Self-EXP 15
  B.2 Demonstrations Generated by Self-ICL 15

C Additional Information on Responsible NLP Research 15

D Prompts 18

E Examples of Our Framework 25

A Additional Experimental Analysis

A.1 Number of Source Tasks Varying with Runtime based on GPT-4

Figure 6 shows the average number of source tasks selected for each target task during the execution of our framework based on GPT-4. Overall, the performance of GPT-4 is consistent with the performance of GPT-3.5 that we analyzed in §4.7. An exception occurs with the HELP dataset, where the number of source tasks runs to 0 between 1500 to 2500 iterations. This is due to we do not consider input questions that skip learning when calculating the average number of source tasks. In other words, between 1500 to 2500 iterations, no examples in the HELP dataset require experience transfer. This is because the proportion of questions skipping learning is relatively high in the HELP dataset, as described in §4.6.

A.2 Number of Tasks and Experience in the Memory Varying with Runtime

Figure 7 and Figure 8 show the number of insights and tasks in memory during the execution of our framework based on GPT-3.5 and GPT-4, respectively. We find that as the number of running rounds increases, our framework accumulates more task-specific experience. This indicates that the capabilities of our framework grow over time, enabling it to cover a broader range of user target tasks or provide experience for more user questions through experience transfer.

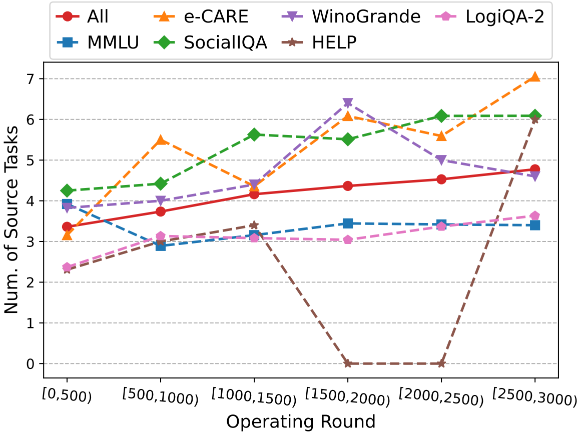


Figure 6: The average number of source tasks selected for each target task during the execution of our framework based on GPT-4.

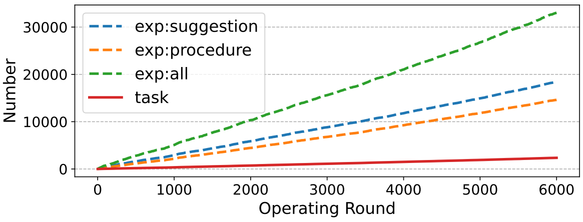


Figure 7: The number of insights and tasks in memory during the execution of our framework based on GPT-3.5.

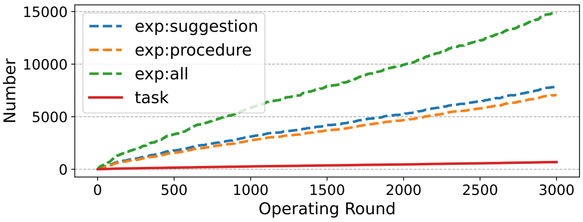


Figure 8: The number of insights and tasks in memory during the execution of our framework based on GPT-4.

A.3 Performance of Experience Induction Through More Examples

Figure 9 shows the number of experience generated by the experience induction module based on GPT-3.5 with more input examples. It can be found that ChatGPT cannot effectively summarize more experience from a larger number of examples. This may be due to the increased difficulty for ChatGPT to analyze, requiring ChatGPT to think for a longer time.

A.4 Prompt Cost

As shown in Table 7, we analyze the cost of our framework by reporting the average token usage per prompt for each example. Please note that for a single example, a prompt may be run multiple times due to reasons such as output format errors or API crashes. All these occurrences are included in the statistics to reflect the true cost.

It can be found that, compare to the traditional

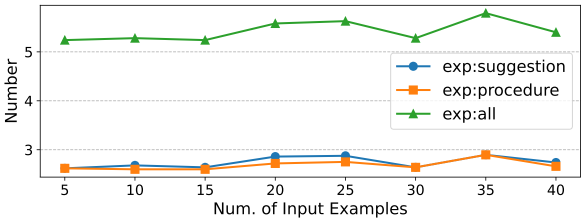


Figure 9: The number of insights generated by experience induction based on GPT-3.5 with more input examples.

Module	PROMPT	Usage		
		Input	Output	Total
Task Type Categorization	Prompt 1	260	63	323
	Prompt 2	825	26	851
Experience Transfer	Prompt 3	289	11	300
	Prompt 4	2139	381	2520
	Prompt 5	456	221	677
Autonomous Practice	Prompt 6	1209	386	1595
	Prompt 7	1184	607	1791
	Prompt 8	3292	309	3601
Experience Induction	Prompt 9	1145	183	1328
Reasoning with Experience	Prompt 10	532	15	546
Total	-	11331	2202	13532
Zero-shot-CoT	Zero-shot-CoT	159	32	191

Table 7: Average token usage per prompt for each example.

zero-shot CoT method, our framework are much more expensive and time-consuming. Overall, our main experiment using GPT-3.5 requires about five days of running, whereas GPT-4 requires three to five times longer. However, this does not mean that the framework is without hope for practical application, as our current basic implementation focuses more on demonstrating the behavior of LLMs at various stages, without any optimization for efficiency. We believe the following approach can be further explored to reduce operational costs:

- Use existing annotated corpora to replace the Autonomous Practice module. It can be found that the main cost of our framework lies in the Autonomous Practice module. Our previous experimental results indicate that, given sufficient prior experience, our framework can perform comparably to the complete framework solely through experience transfer. Therefore, allowing the framework to gain experience from existing annotated corpora first could significantly reduce the substantial costs associated with the Autonomous Practice module.

- Consider using smaller PLMs to perform simple steps. Within our framework, prompts 4, 9, and 10 are involved in experience transfer, experience induction, and experience application, respectively. Other steps are relatively simple and can be substituted with smaller PLMs instead of the expensive ChatGPT.

B Case Study

In this section, we analyze the case of Self-EXP and Self-ICL.

B.1 Experience Generated by Self-EXP

Self-EXP employs Prompt 11 to instruct ChatGPT to directly generate experience for each input question. Case 1 shows the example of experience output by Self-EXP based on GPT-3.5.

For the 1st example, we can find that although the generated experience seems reasonable, it is not well aligned with the input problem. In fact, these generated insights are wrong or irrelevant to the input problem. The possible reason is that ChatGPT only focuses on the keywords of the input question without understanding the essential task objective and the processing flow of the task.

For the 2nd example, Self-EXP suggests that "Consider Addison's typical preferences and behaviors" and "Ask Jesse about Addison's purpose." are valuable insights. These suggestions require LLMs to have the ability to actively explore unknown information and communicate with humans. However, ChatGPT itself does not possess such abilities, and implementing such abilities requires additional auxiliary modules.

Compared with Self-EXP, our framework generates multiple pairs of pseudo questions and reasoning processes, and summarizes experience from them. This ensures that the experience generated by our framework is highly consistent with the input question and matches the abilities of LLMs. Besides, this enables our framework to discover new insights rather than relying solely on the experience learned during ChatGPT's pre-training process.

B.2 Demonstrations Generated by Self-ICL

Self-ICL prompts ChatGPT to generate demonstrations for each input question. Case 2 shows the example of demonstrations generated by Self-ICL based on GPT-3.5. We can find that:

Firstly, there is a format inconsistency between the input question and the question generated by Self-ICL. In fact, this format issue is not an exception but often occurs in generated pseudo questions. ICL will make such format errors interfere with the reasoning process of the LLMs.

Besides, the 2nd example contains a wrong pseudo-response. This response error is difficult to avoid for LLMs unless they have fully mastered the task type of the input question.

In summary, the demonstrations generated by Self-ICL exhibit issues such as inconsistent formatting and noisy responses. However, our framework does not directly utilize the generated demonstrations; instead, it extracts general experience from demonstrations, thereby mitigating the impact of the issues above.

C Additional Information on Responsible NLP Research

Use Scientific Artifacts. As shown in Section §4.1, we use six NLP benchmark datasets in English. They are all allowed to be used for scientific research. They do not contain information that names or uniquely identifies individual people or offensive content.

Human Evaluators. We recruit NLP graduate students to conduct evaluation and are compensated by providing academic credits and guidance. They know explicitly that the annotated data is used for scientific research.

AI Assistants in Writing. We use ChatGPT to help check for grammatical errors and provide suggestions for improving language expression.

Case 1: experience generated by Self-EXP.

[Example 1]

Input Question:
Sentence: I couldn't figure out what my costume would be for Halloween Peter Pan or Snow White. The _ costume I'd have to wear a dress.
Option A: Peter Pan
Option B: Snow White
Choose the more appropriate option to fill in the blank space in the given sentence.
Use the following JSON format to output your answer:

~~~
```json
{
  "correct option ID": /∗ one of A, B ∗/
}
```
~~~

Suggestion:
– Consider the overall theme or context of the event you will be attending for Halloween.
– Think about which character you would enjoy portraying the most.
– Take into account any props or accessories you may need for the costume.
– Consider how comfortable you would be wearing a dress for the duration of the event.
– Avoid making a decision solely based on gender stereotypes.

Procedure:
– Research the character traits, appearance, and costumes of Peter Pan and Snow White.
– Consider your personal preferences and the level of comfort in each costume.
– Try on different costume options to see which one suits you best.
– Seek feedback from friends or family members.
– Ultimately, choose the option that aligns with your personal interests and comfort level.

[Example 2]

Input Question:
Context: Jesse walked Addison's dog because she was on vacation.
Question: What will Addison want to do next?
Option A: great
Option B: going to the ocean to swim
Option C: loved
Based on the given context, choose the correct answer to the question from the three options.
Use the following JSON format to output your answer:

~~~
```json
{
  "correct option ID": /∗ one of A, B, C ∗/
}
```
~~~

Suggestion:
– Understand the context of the situation and the characters involved
– Consider Addison's typical preferences and behaviors
– Ask Jesse about Addison's purpose.

Procedure:
– Analyze the situation – Addison's dog being walked by Jesse while she's on vacation – and anticipate what Addison would want next
– Choose the option that aligns best with Addison's typical behaviors and interests

Case 2: demonstrations generated by Self-ICL.

[Example 1]

Input Question:
Premise: The magazine revealed all his dirty watchwords .
Hypothesis: The magazine revealed all his dirty secrets .
You need to decide whether the hypothesis is entailed by the premise by choosing one of the following answers:
"Yes": The hypothesis follows logically from the information contained in the premise.
"No": The hypothesis is logically false from the information contained in the premise.
"Neutral": It is not possible to determine whether the hypothesis is true or false without further information.
Use the following JSON format to output your answer:

~~~
```json
{
  "answer": /* Yes, No or Neutral */
}
```
~~~

Generated Pseudo-Question:
Premise: The company announced a new product launch date.
Hypothesis: The new product will be a bestseller.

~~~
```json
{
  "answer": "Neutral"
}
```
~~~

Generated Pseudo-Response:
"Neutral" – The premise of the company announcing a new product launch date does not necessarily entail that the new product will be a bestseller. It is possible that the new product could be successful, but it is not guaranteed based on the premise alone.

[Example 2]

Input Question:
Premise: Most of them chose to eat steak.
Choice A: A group of men went to a restaurant for dinner.
Choice B: The new guests coming in the famous steakhouse were four blokes.
For the given two options, choose the one that is more likely to cause the occurrence of the premise.
Use the following JSON format to output your answer:

~~~
```json
{
  "correct choice ID": /* one of A, B */
}
```
~~~

Generated Pseudo-Question:
Premise: The majority of the passengers on the flight slept during the journey.
Choice A: The flight attendants provided exceptional service and entertainment options.
Choice B: The flight experienced turbulence and delays due to bad weather.

~~~
```json
{
  "correct choice ID": "B"
}
```
~~~

Generated Pseudo-Response:
The correct choice is B: The flight experienced turbulence and delays due to bad weather. This choice is supported by the premise as it explains why the majority of the passengers slept during the journey likely due to the uncomfortable conditions caused by turbulence and the delay in reaching their destination. Choice A is not supported by the premise and does not explain why the passengers slept during the journey.

D Prompts

Prompt 1: generate the corresponding task type and task description of the user question.

You are an advanced task type induction agent capable of naming a task and describing its goals based on an example of the task.
The description of the task goals should be abstract, general, and essential, avoiding any specifics about how the problem is described or the variable elements within it, as the same task can be described in various ways.
Use the following JSON format to output task name and task descriptions:

```json
{
  "task name": ,
  "task description":
}
```

<Task Example >
[user question]
</Task Example >

Prompt 2: determine whether the target task is identical to one of the candidate tasks in memory.

<Target Task>
[task description of the target task]
</Target Task>

<Candidate Task 1>
[task description of the 1st candidate task]
</Candidate Task 1>

<Candidate Task 2>
[task description of the 2nd candidate task]
</Candidate Task 2>

[...the remaining candidate tasks...]

You are an excellent task identifier, capable of determining whether the target task is identical to one of the above candidate tasks.
If no such candidate tasks exist, or if you are unsure, please return −1.
You must carefully avoid selecting any candidate task that are not completely identical to the target task.
Please use the following JSON format to output the selected candidate task:

```json
{
"selected task id": /∗ −1 or ID of the selected candidate task. ∗/
}
```

Prompt 3: select source tasks for the target task during experience transfer.

<Target Task>
[task description of the target task]
</Target Task>

<Candidate Task 1>
[task description of the 1st candidate task]
</Candidate Task 1>

continued from the above content.

<Candidate Task 2>
[task description of the 2nd candidate task]
</Candidate Task 2>

[...the remaining candidate tasks...]

You are an outstanding source task retriever, capable of discovering source tasks related to the target task from the above candidate tasks.
The experience gained from solving the source tasks should be transferable to the target task.
Use the following JSON format to output the selected source tasks:

```json
{
"selected task ids": [ /* ids of selected source tasks. If there are no suitable source tasks, please return an empty list. */ ]
}
```

Prompt 4: transfer the experience of multiple source tasks to the target task.

You are an excellent experience transfer agent, adept at transferring experience from one or more source tasks to the target task.
Here is the task description of the target task, as well as the task description and task experience of source tasks.

<Target Task>
[task description of the target task]
</Target Task>

<Source Task 1>
Task Description:
[task description of the 1st source task]
Task Experience:
[task experience of the 1st source task]
</Source Task 1>

<Source Task 2>
Task Description:
[task description of the 2nd source task]
Task Experience:
[task experience of the 2nd source task]
</Source Task 2>

[...the remaining source tasks...]

Please follow the steps below to transfer experience:

Step 1: Task Understanding
Thoroughly understand the target task and source tasks, clearly identifying the commonalities and differences between them.

Step 2: Identify General Experience
Extracting general experience from the source tasks that can also be applied to the target task, especially insights that are common across multiple source tasks.
Avoid using task−specific experience from the source tasks that may not be relevant to the target task.
Be cautious of experience effective in the source tasks but could lead to errors in the target task.
Pay attention to the differences between the source and target tasks.

continued from the above content.

Step 3: Experience Adaptation
Adapt the general experience identified in Step 2 to the target task, adjusting for aspects that do not align perfectly with the target task's conditions and meeting the specific requirements of the target task.
Ensure that the experience provided are CLEAR, DETAILED, and GENERALLY APPLICABLE to unseen examples in the target task.
Use the following JSON format to output the adapted experience:
```json
{
"How to better accomplish the task or avoid low–quality responses": [ no more than 20 insights ],
"The specific process for handling this task": [ no more than 20 insights ]
}
```

Let's think step by step.

Prompt 5: combine and deduplicate two sets of experience for the same task.

<Target Task>
[task description of the target task]
</Target Task>

<Existing Experience>
{
"How to better accomplish the task or avoid low–quality responses":
[list all the unordered suggestions from two sets of experience.],
"Task Processing Flow 1": **[the ordered procedure from the first set of experience.]**,
"Task Processing Flow 2": **[the ordered procedure from the second set of experience.]**
</Existing Experience>

You are an excellent experience refiner. Please help me refine the above existing experience related to the target task.
1. For "How to better accomplish the task or avoid low–quality responses", please integrate insights by combining those that are closely related and eliminating any repetitions.
2. Please integrate the above "Task Processing Flow 1" and "Task Processing Flow 2" into one unified workflow process. Ensure that the primary goals and functionality of both original processes are preserved; Effectively resolve possible conflicts or overlaps between the two processes.
Use the following JSON format to output refined target task experience:
```json
{
"How to better accomplish the task or avoid low–quality responses": [ no more than 20 insights ],
"The specific process for handling this task": [ no more than 20 insights ]
}
```

Prompt 6: generate a new question of the target task type based on the reference web text.

<Reference Text>
[reference text retrieved from the internet]
</Reference Text>

continued from the above content.

<Example Question>
[The example question of the target task, i.e., the input user question of our framework]
</Example Question>

<Task Type of the Example Question>
[task description of the target task]
</Task Type of the Example Question>

You are an excellent questioner.
Please carefully read the reference text provided above and formulate a new question based on it.
The new question must maintain the same expression style, structure, and required output format as the example question.
The new question must belong to the same task type of the example question.
The new question must be well−defined, with a complete and clear description that can be answered and at least one correct answer exists.
You are forbidden from providing answers to your new question.
Use the following format to output your answer:
<New Question>
/∗ Your new question. ∗/
</New Question>

Prompt 7: during the autonomous practic process, generate a thought process and answer to the generated new question based on experience.

<Task Experience>
[experience of the target task]
</Task Experience>
Please refer to the above experience to answer the following question.
# The above part is omitted when the experience is empty.

[a generated new question]

Please provide specific, detailed, and comprehensive steps of your thought.

Prompt 8: based on the reference text, check if the response to the question is correct.

<Reference Text>
[reference text retrieved from the internet]
</Reference Text>

<Target Question>
[the generated new question]
</Target Question>

<Reasoning Process and Answer>
[the thought process and answer of the new question]
</Reasoning Process and Answer>

continued from the above content.

You are an outstanding checker, skilled at examining the reasoning process and the correctness of the answer of the target question based on the reference text.
Pay close attention to whether the reasoning process and the answer are consistent or inconsistent with the reference text.
Use the following JSON format to output your opinion:
```json
{
"correctness": /* "correct", "wrong" or "inconclusive" */
}
```

Let's think step by step.

Prompt 9: summarize the task-solving experience from examples with correct or incorrect answers.

You are an excellent experiential summarizer, adept at extracting task−solving insights from examples of the target task.
Here are several target task examples with correct or incorrect answers:
<Correct Example 1>
<Question>
[the generated new question]
</Question>
<Reasoning Process and Answer>
[the thought process and answer of the new question]
</Reasoning Process and Answer>
</Correct Example 1>

[...the remaining correct examples...]

<Incorrect Example 1>
<Question>
[the generated new question]
</Question>
<Reasoning Process and Answer>
[the thought process and answer of the new question]
</Reasoning Process and Answer>
</Incorrect Example 1>

[...the remaining incorrect examples...]

Based on the examples provided above, please follow the steps below to summarize the experience:

Step1: Observe and Analyze the Examples
Summarize the commonalities in the correct examples, identify patterns in the incorrect examples, and compare the differences between the correct and incorrect examples.

Step2: Summarize Experience
Based on the observations and analysis from the Step1, summarize task−solving insights.
Ensure that the insights provided are CLEAR, DETAILED, and are GENERALLY APPLICABLE to unseen examples of the target task.
Use the following JSON format to output the summarized experience:

continued from the above content.

~~~
```json
{
"How to better accomplish the task or avoid low−quality responses": [ no more than 20 insights ],
"The specific process for handling this task": [ no more than 20 insights ]
}
```
~~~

Let's think step by step.

Prompt 10: think the question based on experience and respond to the user.

<Experience>
[How to better accomplish the task or avoid low−quality responses]:
[list the unordered suggestions from the experience.]
[The specific process for handling this task]:
[list the ordered procedure from the experience.]
</Experience>
Please refer to the above experience to answer the following question.

[the input user question of our framework]

Prompt 11: directly generate task-solving experience for the input question.

You are an excellent advisor, skilled in providing task−solving insights for the target task.
<Target Task>
[the input question]
</Target Task>

Please give your suggestions.
Ensure that the insights provided are CLEAR, DETAILED.
Use the following JSON format to output:

~~~
```json
{
"How to better accomplish the task or avoid low−quality responses": [ your insights ],
"The specific process for handling this task": [ your insights ]
}
```
~~~

E Examples of Our Framework

In this section, we demonstrate examples of our framework.

Example 1 [GPT-3.5]: task-specific experience in the memory of our framework.

Task 1

Task Name: Logical Reasoning Inference Task
Task Description:
The goal of this task is to evaluate the ability to draw logical inferences from a given context and make deductions based on the information provided.

Suggestion:
- Thoroughly understand the provided context or situation before drawing inferences
- Consider the broader background information related to the given information
- Use logical reasoning to evaluate the relevance and probability of the deductions
- Ensure the selected deduction leads to a probable and logical outcome based on the given context
- Provide clear and logically reasoned justifications for the chosen deduction based on the likelihood of being correct
- Align deductions with the emotional significance or conflicting categories in the context
- Evaluate the likelihood of each deduction based on the context provided
- Focus on the most probable outcome based on the context
- Eliminate deductions that do not logically align with the given context to narrow down the options
- Avoid introducing irrelevant or unrelated information in the deductions

Procedure:
- Step 1: Thoroughly understand the provided context and question for clarity and logical deductions.
- Step 2: Consider the broader context or background information related to the given question or statement to make accurate deductions.
- Step 3: Use logical reasoning to evaluate the relevance and probability of the available options for making informed deductions.
- Step 4: Employ critical thinking to evaluate the relevance and probability of the inferences
- Step 5: Choose the most probable inference that logically follows from the given context and premise.
- Step 6: Consider the emotional significance or conflicting categories in the context to align deductions.
- Step 7: Provide clear and concise reasoning for the chosen inferences to demonstrate a thorough understanding of the context and premise.

Task 2

Task Name: Sentence Completion
Task Description:
The task requires the selection of the more appropriate option to fill in the blank space in a given sentence.

Suggestion:
- Thoroughly evaluate the given sentence and assess the most suitable option based on logical reasoning, contextual understanding, and specific language comprehension.
- Consider the broader context or background information related to the given sentence to enhance comprehension and decision−making.
- Use precise and unambiguous language to accurately convey the most suitable option, avoiding vague or misleading statements.
- Avoid introducing irrelevant or unrelated information in the options to ensure alignment with the given sentence.
- Provide clear and logically reasoned justifications for the chosen option, demonstrating an understanding of the context and logical completion.

Procedure:
- Step 1: Thoroughly understand the given sentence or context to identify key elements and relationships between entities.
- Step 2: Use logical reasoning to determine the most suitable option based on the given context, ensuring that the chosen option logically follows from the context and maintains consistency.
- Step 3: Choose the option that logically follows from the context provided in the sentence and provide a clear and detailed reasoning process for the chosen option.

continued from the above content.

Task 3

Task Name: Understanding Universal Conclusions
Task Description:
This task involves recognizing and understanding the process of drawing universal conclusions from a wide range of concrete data, without getting into specifics about the data or the particular conclusions drawn.

Suggestion:
– Thoroughly analyze a wide range of concrete data and historical evidence before drawing universal conclusions
– Use logical reasoning to assess the plausibility of conclusions drawn from the available data
– Avoid introducing irrelevant or unrelated information in the analysis to ensure alignment with the provided context
– Consider the broader historical and societal implications of the conclusions drawn
– Carefully evaluate the given data and information to identify the key points and implications for drawing universal conclusions
– Prioritize conclusions based on logical reasoning and critical thinking aligned with the available data
– Avoid biases and assumptions that may affect the determination of universal conclusions
– Provide clear and detailed justifications based on specific knowledge and evidence relevant to the universal conclusions

Procedure:
– Thoroughly analyze the concrete data and historical evidence to identify patterns and overarching themes
– Use logical reasoning to assess and justify the selected universal conclusions
– Consider the broader historical and societal implications to ensure a comprehensive evaluation
– Select the conclusions that best align with the main overarching themes and patterns identified in the concrete data
– Ensure that the conclusions are logically supported by the available data and do not make unwarranted assumptions

Task 4

Task Name: Next Move Prediction
Task Description:
The goal is to predict the likely course of action that others will take based on a given context, and the available options, in order to make an informed decision or plan potential actions.

Suggestion:
– Thoroughly understand the given context before predicting the likely course of action
– Use logical reasoning to evaluate potential courses of action and provide clear justifications based on the context
– Consider the broader context related to the given situation for enhanced comprehension and decision–making
– Avoid making hasty predictions without considering all relevant information
– Evaluate potential consequences and factors impacting the decision–making process when predicting others' likely actions
– Apply critical thinking skills to analyze and interpret the information before making a decision
– Eliminate irrelevant options based on logical reasoning
– Ensure that the chosen option aligns with specific details and logical reasoning within the given context
– Provide clear and reasoned justifications for the chosen course of action
– Focus on the most probable outcome based on the context
– Carefully read and understand the premise before evaluating the choices
– Eliminate absolute or extreme options that do not account for the complexity of the situation
– Align the chosen action with the emotional significance of the context and specific knowledge of the situation

Procedure:
– Step 1: Thoroughly understand the given context or situation for clarity and coherence
– Step 2: Use logical reasoning to evaluate the available options and determine the most suitable one based on the context and question
– Step 3: Provide clear and logically reasoned justifications for the chosen option, aligning with the specific details and logical reasoning within the given context
– Step 4: Consider the broader context related to the given situation for enhanced comprehension and decision–making in predicting others' likely actions
– Step 5: Thoroughly evaluate the given situation and context to assess the potential consequences and factors impacting the decision–making process
– Step 6: Choose the option that best reflects the significance of the context provided

Example 2 [GPT-4]: task-specific experience in the memory of our framework.

Task 1

Task Name: Predicting Consequences
Task Description:
The goal of this task is to analyze a given context involving social interactions and conflicts, and to predict the most likely immediate behavior or action that a character will take in response to the event described.

Suggestion:
– Focus on the natural and logical progression from the context provided.
– Consider the most reasonable and effective way to address the issue presented.
– Understand the immediate goals or wants of the characters involved.
– Reflect actions that real people might take in similar real–life scenarios.
– Evaluate options based on how well they fit within the broader context or narrative.
– Dismiss options that result in discontinuity or that do not address the situation effectively.
– Analyze the implied urgency or importance of the situation to determine appropriate actions.
– Weigh the potential consequences of each option to select the most constructive outcome.
– Consider the norms of behavior within the given context (e.g., professional settings, health matters).
– Prioritize actions that would seemingly lead to conflict resolution or advancement of the characters' objectives.
– Identify any potential misunderstandings or contradictions in the options against the context.
– Take into account the impact of cultural or situational factors on decision–making.
– Avoid options that indicate overreactions or underreactions to the presented issue.
– Eliminate options that could introduce unnecessary complexity or confusion.
– Assess the level of urgency and choose an option that reflects an appropriate response time.
– Be cautious of options that might represent plausible actions but are secondary or unrelated to the pressing issue.

Procedure:
– Identify the key issue or problem presented in the context.
– Determine what the main objectives of the subjects involved are.
– List the available options and examine the viability of each in relation to the context and objectives.
– Use deduction to eliminate options that less effectively address the problem or are inconsistent with likely behavior.
– Consider the consequences and effectiveness of each remaining option.
– Select the option that provides a solution or continuation most aligned with the context and characterized behavior.
– Ensure the choice conforms with logical and rational decision–making.
– Verify the choice does not violate any implicit or explicit norms established by the context.
– Review the reasoning process to ensure it is free of bias and based on the information provided.
– Confirm that the final selection promotes progression rather than regression or stagnation.

Task 2

Task Name: Legal Multiple Choice Question Answering
Task Description:
The goal of the task is to determine the correct answer to a legal problem presented in the format of a multiple choice question by applying legal principles and rules relevant to the scenario provided. The task entails analyzing the given information, discerning the legal issue at hand, considering the potential implications of applicable laws or statutes (such as the Dead Man's Statute in this case), and selecting the answer that best aligns with legal rationale.

Suggestion:
– Assess relevancy: Consider whether evidence provides significant support for the legal requirement or argument at issue.
– Understand legal standards: Be aware of the legal criteria for decisions such as admissibility of evidence and motions like JNOV.
– Recognize exceptions: Identify any potential legal exceptions that might apply, particularly in areas like hearsay evidence.
– Identify applicable law: Determine which laws or legal principles are relevant to the decision–making process.
– Analyze logical reasoning: Scrutinize the connections between the facts presented and the options given to ensure logical coherence.
– Evaluate the role of discretion: Understand the extent of discretion that a judge has in making decisions about evidence and jury verdicts.

continued from the above content.

– Consider judicial remedies: Be aware of the different types of remedies a judge can offer, especially in response to jury verdicts.
– Examine the facts: Closely inspect the facts of the case as presented in the scenario to guide which option is most appropriate.
– Account for the legal outcome: Factor in the consequence each option would logically lead to within the context of the case.
– Avoid assumptions: Do not make unfounded assumptions; base decisions on information present in the question.
– Prioritize direct connections: Choose options where the connection to the case scenario is most direct and unambiguous.
– Anticipate objections: Consider the probable grounds on which an attorney would object and how the law typically treats such objections.
– Factor in all case details: Make sure all components of the scenario are considered and none are inadvertently overlooked.
– Identify procedural context: Be aware of where in the legal process the question situates you, as this can affect the available options.
– Prevent personal bias: Avoid allowing personal beliefs or biases to influence the decision of which option to choose.
– Acknowledge case precedents: When applicable, take into account similar cases or precedents that may influence the ruling.
– Reflect on the public policy: Consider how the legal decision might reflect on social norms or public policy where relevant.
– Stay updated on legal changes: Keep abreast of any recent changes in the law that might affect the interpretation or application of legal concepts.
– Be precise with legal terminology: Understand and apply the precise meanings of legal terms that are relevant to the answers.
– Conduct a comprehensive analysis: Integrate all relevant factors and details from the scenario for a well–rounded assessment.
Procedure:
– Read and comprehend the scenario carefully to understand the case details and legal issues involved.
– Identify and deconstruct the question, pinpointing exactly what legal matter needs resolving.
– Match the facts of the case to the answer options, discarding those that don't align with the presented scenario.
– Search for keywords or phrases that directly relate to legal principles or standards which govern the outcome.
– Eliminate options that are clearly not supported by the provided facts or the prevailing legal framework.
– Consider the implications of each remaining option, focusing on those that lead to a legally sound consequence.
– Apply critical thinking to discern patterns or nuances that relate the scenario to the answer options.
– Review related legal doctrines or principles that can provide a basis for evaluating the correct response.
– Weigh the evidence presented in light of its relevance and sufficiency to support the various answer options.
– Recall relevant legal precedents or case law that may influence the admissibility or effect of evidence.
– Look for any exceptions to general rules, especially when dealing with hearsay or objections to evidence.
– Gauge the extent to which each option adheres to the rule of law as presented in the scenario and options.
– Assess objections and remedies properly by considering the context and possible legal responses to them.
– Clarify ambiguous terms or concepts that may have a significant impact on the correctness of the answer.
– Evaluate whether the options are addressing the question directly or whether they are tangential or irrelevant.
– Contemplate the logical consistency of the reasoning for each option as it pertains to the facts of the case.
– Distill the essence of the case to focus on the main issues that will determine the correct answer.
– Cross–reference options with facts, aligning them to see if they would naturally lead to the outcome suggested.
– Adopt an objective standpoint, disregarding any potential personal opinions about the case.
– After an option is selected, re–examine it against the scenario to confirm it is indeed the most fitting and cogent choice.

Task 3

Task Name: Emotion Prediction based on Context
Task Description:
Determine the most likely emotional response that a generic person or group would experience in a given situation or after an event, using provided context and potential emotional states.
Suggestion:
– Understand the context and how it relates to human emotions or behavior to accurately predict responses.
– Make logical inferences about people's feelings or reactions based on their preferences, actions, or the nature of the situation presented.
– Avoid choices that do not align with the implications of the given context.
– Be mindful that people's reactions are influenced by their perceptions and experiences.
– Recognize that positive contexts or enhancements tend to evoke positive emotions such as optimism or reassurance.
– Negative or confusing contexts may provoke feelings like anxiety or perplexity.
– Consider the outcome expectations; effective solutions typically lead to positive emotions.

continued from the above content.

– Acknowledge that familiarity and preference influence comfort and ease, impacting responses to situations.
– Ensure responses are not arbitrary; they should be backed by the rationale provided in the context.
– Be aware of the common emotions associated with certain experiences, such as comfort with familiar systems or security in safe processes.
– Identify keywords in the context that hint at the emotional state being queried.
– Assess whether the question demands a logical or an emotional response and tailor your answer accordingly.
– Keep assumptions consistent with everyday experiences and normal psychological reactions.
– Consider societal norms and general expectations when determining what reaction is likely.
– Examine the implications of a context for different stakeholders to accurately assess potential reactions.
– Focus on the most direct and significant emotional impact as suggested by the context.
– Avoid over−complication; opt for the most straightforward interpretation that fits the provided information.
– Take note of extremes in wording that might mislead the anticipated emotion or reaction.
– Understand the perspective being asked about (e.g., personal user experience vs. institutional procedure).
– Double−check that the chosen option is the only one that fits logically with the context.

Procedure:
– Identify the key elements of the context that will influence the emotion or reaction.
– Read the question carefully to understand exactly whose feelings or reactions are being asked about.
– Analyze each response option and how it correlates with the identified key elements of the context.
– Utilize deductive reasoning to eliminate options that do not fit with the clues provided in the context.
– Extract implications from the context that might affect the emotional state in question.
– Predict the most logical emotional response based on the context.
– Select the answer option that best represents this predicted emotional response.
– Ensure that the answer is consistent with the context and the question posed.
– Reflect on common human behaviors and responses to similar situations.
– Verify the chosen option against alternative interpretations of the context.
– Review the reasoning process for systematic biases or leaps in logic.
– Be aware of nuances in the language used within the context that might influence the interpretation.
– Assess if the context implies security, efficiency, preference, or other factors that impact emotional reactions.
– Maintain objectivity and avoid personal biases when interpreting the context and predicting emotions.
– Revisit initial conclusions with a critical mind to confirm the logical coherence.
– Take into account both direct and indirect hints in the context when deducing emotions.
– Articulate a clear rationale for why each incorrect option was dismissed upon selecting the correct answer.
– Consider how a neutral observer would likely interpret the context and emotions.
– Cross−examine the final choice by envisioning real−life scenarios that parallel the context given.
– Conclude with a brief validation step to ensure that the selected answer aligns with general human experience.

Example 3 [GPT-3.5]: Prompt 1 generates the corresponding task type and task description of the user question. This is the first step of our framework to process a user question.

INPUT:
You are an advanced task type induction agent capable of naming a task and describing its goals based on an example of the task.
The description of the task goals should be abstract, general, and essential, avoiding any specifics about how the problem is described or the variable elements within it, as the same task can be described in various ways.
Use the following JSON format to output task name and task descriptions:

~~~
```json
{
   "task name": ,
   "task description":
}
```
~~~

<Task Example>
Premise: Tom has a fever.
Choice A: They are TIPs.
Choice B: Doctors think Tom's fever is caused by infection.

continued from the above content.

For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:

```json
{
  "correct choice ID": /∗ one of A, B ∗/
}
```

</Task Example>

RESPONSE:

```json
{
  "task name": "Likelihood Judgment",
  "task description": "Assess and compare the likelihood of different outcomes based on a given premise"
}
```

Example 4 [GPT-3.5]: Prompt 2 determines whether the target task is identical to one of the candidate tasks in memory. After Prompt 1, we retrieve the top 5 tasks from memory that are semantically most similar to the generated task description. Then, ChatGPT perform Prompt 2. If a match is found, the question is linked to the selected task; otherwise, it is linked to the generated task, and we add the generated task into the memory with empty initial task experience.

INPUT:
<Target Task>
Assess and compare the likelihood of different outcomes based on a given premise
</Target Task>

<Candidate Task 1>
Given a scientific conclusion, select the piece of evidence that best supports the conclusion from a list of options.
</Candidate Task 1>

<Candidate Task 2>
The task requires identifying the choice that is more likely to have caused the given premise based on the options provided.
</Candidate Task 2>

<Candidate Task 3>
This task requires determining the choices that are more likely to be caused by the given premise.
</Candidate Task 3>

<Candidate Task 4>
The goal of this task is to determine whether the hypothesis logically follows from the information contained in the premise, without providing any specific details about the content of the premise and hypothesis.
</Candidate Task 4>

<Candidate Task 5>
The task involves inferring the emotional response of a person in a given situation based on their actions or the actions of others towards them.
</Candidate Task 5>

You are an excellent task identifier, capable of determining whether the target task is identical to one of the above candidate tasks.
If no such candidate tasks exist, or if you are unsure, please return −1.
You must carefully avoid selecting any candidate task that are not completely identical to the target task.
Please use the following JSON format to output the selected candidate task:

~~~
```json
{
"selected task id": /∗ −1 or ID of the selected candidate task. ∗/
}
~~~

RESPONSE:

```
{
  "selected task id": 3
}
```

Example 5 [GPT-3.5]: Prompt 3 selects source tasks for the target task during experience transfer. After Prompt 2, the Task Type Categorization are finished. Then, we follow Sec 3.6 to identify learning or skip learning. If cotinue learning, we perform Prompt 3, which is the start of the Experience Transfer.

INPUT:
<Target Task>
Assess and compare the likelihood of different outcomes based on a given premise
</Target Task>

<Candidate Task 1>
Given a scientific conclusion, select the piece of evidence that best supports the conclusion from a list of options.
</Candidate Task 1>

<Candidate Task 2>
The task requires selecting the appropriate one from several candidates to complete the sentence.
</Candidate Task 2>

<Candidate Task 3>
The goal of this task is to assess the ability to understand and interpret beliefs and their impact on communication and behavior.
</Candidate Task 3>

<Candidate Task 4>
The goal of this task is to determine whether the hypothesis logically follows from the information contained in the premise, without providing any specific details about the content of the premise and hypothesis.
</Candidate Task 4>

<Candidate Task 5>
The task involves inferring the emotional response of a person in a given situation based on their actions or the actions of others towards them.
</Candidate Task 5>

<Candidate Task 6>
The goal of this task is to identify the most appropriate course of action based on a given situation or problem, considering potential consequences and factors that could impact the decision−making process.
</Candidate Task 6>

<Candidate Task 7>
This task aims to assess the understanding of the subfield of psychology that specializes in the study of subjective well−being, optimism, and happiness.
</Candidate Task 7>

<Candidate Task 8>
The task requires identifying the option that resolves a paradox presented in a given scenario. The goal is to select the option that provides new information or a different perspective to reconcile conflicting statements or observations.
</Candidate Task 8>

<Candidate Task 9>
Predict the more probable outcome given a specific premise and two potential choices.
</Candidate Task 9>

continued from the above content.

<Candidate Task 10>
The task involves determining whether the main character in a given scenario has engaged in morally wrong behavior according to ordinary moral standards in the US as of 2020.
</Candidate Task 10>

You are an outstanding source task retriever, capable of discovering source tasks related to the target task from the above candidate tasks.
The experience gained from solving the source tasks should be transferable to the target task.
Use the following JSON format to output the selected source tasks:

~~~
```json
{
"selected task ids": [ /∗ ids of selected source tasks. If there are no suitable source tasks, please return an empty list. ∗/ ]
}
```
~~~

RESPONSE:

```
{
"selected task ids": [1, 9, 8]
}
```

Example 6 [GPT-3.5]: Prompt 4 transfers the experience of multiple source tasks to the target task. After Prompt 3, we perform Prompt 4.

INPUT:
You are an excellent experience transfer agent, adept at transferring experience from one or more source tasks to the target task.
Here is the task description of the target task, as well as the task description and task experience of source tasks.

<Target Task>
Assess and compare the likelihood of different outcomes based on a given premise
</Target Task>

<Source Task 1>
Task Description: Given a scientific conclusion, select the piece of evidence that best supports the conclusion from a list of options.
Task Experience:

```
{
  "How to better accomplish the task or avoid low−quality responses": [
    "Align the options with the provided scientific conclusion and context",
    "Evaluate the logical connection between the evidence options and the given conclusion",
    "Consider the likelihood of each piece of evidence based on the scientific context",
    "Avoid including irrelevant or unrelated information in the evidence options",
    "Focus on the evidence that most directly and strongly supports the given scientific conclusion",
    "Carefully read and understand the scientific conclusion before evaluating the evidence options",
    "Consider the broader scientific context or background information related to the conclusion",
    "Use logical reasoning to evaluate the relevance and probability of the evidence based on the scientific context",
    "Carefully analyze the given context and question before evaluating the options",
    "Use background knowledge or contextual clues to support the chosen answer",
    "Eliminate options that do not logically connect with the given context",
    "Eliminate unlikely scenarios that do not logically connect with the context",
    "Avoid choosing answers without sufficient logical reasoning or evidence"
  ],
```

continued from the above content.

```
  "The specific process for handling this task": [
    "Carefully read and understand the provided scientific conclusion and context",
    "Evaluate each evidence option based on its logical connection to the conclusion",
    "Consider the likelihood of each piece of evidence based on the scientific context",
    "Eliminate evidence options that do not logically support the given scientific conclusion",
    "Evaluate the implications and consequences of each piece of evidence based on the given context",
    "Select the evidence that best supports the provided scientific conclusion",
    "Consider the broader scientific context or background information related to the conclusion",
    "Use logical reasoning to evaluate the relevance and probability of the evidence based on the scientific context",
    "Carefully analyze the given context and question before evaluating the options",
    "Use background knowledge or contextual clues to support the chosen answer",
    "Eliminate options that do not logically connect with the given context",
    "Eliminate unlikely scenarios that do not logically connect with the context",
    "Choose the most relevant and likely answer based on the context and logical reasoning"
  ]
}
</Source Task 1>

<Source Task 2>
Task Description: Predict the more probable outcome given a specific premise and two potential choices.
Task Experience:
{
  "How to better accomplish the task or avoid low−quality responses": [
    "Always align the choices with the provided premise and context",
    "Evaluate the logical connection between the premise and the choices",
    "Consider the likelihood of each choice based on the given context",
    "Avoid introducing irrelevant or unrelated information in the choices",
    "Focus on the most probable outcome based on the context provided",
    "Carefully read and understand the premise before evaluating the choices",
    "Consider the broader context or background information related to the premise",
    "Ensure that the choices are directly connected to the given premise or statement",
    "Consider the characters' knowledge, beliefs, and perspectives when selecting the correct option",
    "Logical reasoning process should guide the selection of the correct option",
    "Choose options that align with the characters' knowledge or beliefs to lead to probable outcomes",
    "Evaluate the choices based on their logical connection to the premise to assess causality",
    "Eliminate choices that do not logically follow from the given premise to narrow down the options",
    "Carefully read and understand the given premise or statement to establish a clear foundation for evaluation",
    "Evaluate the implications and consequences of each choice based on the given context for a thorough analysis",
    "Use logical reasoning to evaluate the relevance and probability of the options based on the context",
    "Select the option that aligns with the characters' knowledge or beliefs and logically addresses the given context and
    question",
    "Ensure that the chosen option leads to a probable and logical outcome based on the given context"
  ],
  "The specific process for handling this task": [
    "Carefully read and understand the given premise or statement",
    "Evaluate the choices based on their logical connection to the premise to assess causality",
    "Consider the likelihood of each choice based on the context provided to make an informed prediction",
    "Eliminate choices that do not logically follow from the given premise to narrow down the options",
    "Evaluate the implications and consequences of each choice based on the given context for a thorough analysis",
    "Choose the option that is most likely to result from the provided premise for a decisive prediction",
    "Consider the characters' perspectives and the premise provided to ensure a thorough evaluation",
    "Use logical reasoning to evaluate the relevance and probability of the options based on the context",
    "Select the option that aligns with the characters' knowledge or beliefs and logically addresses the given context and
    question",
    "Ensure that the chosen option leads to a probable and logical outcome based on the given context",
    "Thoroughly evaluate the logical connection of the choices to the given premise",
    "Consider the context and likelihood of each choice to make informed predictions"
  ]
}
</Source Task 2>
```

continued from the above content.

<Source Task 3>
Task Description: The task requires identifying the option that resolves a paradox presented in a given scenario. The goal is to select the option that provides new information or a different perspective to reconcile conflicting statements or observations.
Task Experience:
{
"How to better accomplish the task or avoid low−quality responses": [
"Carefully evaluate the paradox and consider the likelihood of each option based on the specific context",
"Align the options with the paradox and evaluate their logical connection to provide new information or a different perspective",
"Use logical reasoning to assess the relevance of each option in resolving the paradox",
"Focus on selecting the option that offers a new insight or perspective to reconcile the conflicting statements or observations",
"Carefully evaluate the given paradox and understand the conflicting statements or observations in the context before evaluating the options.",
"Select options that provide new information or a different perspective to directly address the paradox and explain the conflicting observations.",
"Ensure that the chosen options offer clear and logical explanations that reconcile the conflicting statements in the given context.",
"Avoid selecting options that merely repeat information already given in the context without providing new insights or addressing the paradox directly.",
"Evaluate the relevance of each option based on its logical connection to the paradox and its ability to offer a resolution to the conflicting observations."
],
"The specific process for handling this task": [
"Step 1: Carefully evaluate the given paradox and understand the conflicting statements or observations in the context.",
"Step 2: Evaluate each option based on its logical connection to the paradox and its ability to provide new information or a different perspective to resolve the conflicting observations.",
"Step 3: Select the option that directly addresses the paradox and offers a clear and logical explanation that reconciles the conflicting statements in the given context.",
"Step 4: Avoid introducing irrelevant or unrelated information in the options to resolve the paradox",
"Step 5: Ensure the chosen option aligns with the specific attributes and context of the given paradox"
]
}
</Source Task 3>

Please follow the steps below to transfer experience:

Step 1: Task Understanding
Thoroughly understand the target task and source tasks, clearly identifying the commonalities and differences between them.

Step 2: Identify General Experience
Extracting general experience from the source tasks that can also be applied to the target task, especially insights that are common across multiple source tasks.
Avoid using task−specific experience from the source tasks that may not be relevant to the target task.
Be cautious of experience effective in the source tasks but could lead to errors in the target task.
Pay attention to the differences between the source and target tasks.

Step 3: Experience Adaptation
Adapt the general experience identified in Step 2 to the target task, adjusting for aspects that do not align perfectly with the target task's conditions and meeting the specific requirements of the target task.
Ensure that the experience provided are CLEAR, DETAILED, and GENERALLY APPLICABLE to unseen examples in the target task.
Use the following JSON format to output the adapted experience:
```json
{
"How to better accomplish the task or avoid low−quality responses": [ no more than 20 insights ],
"The specific process for handling this task": [ no more than 20 insights ]
}
```

Let's think step by step.

continued from the above content.

RESPONSE:
Step 1: Task Understanding

Upon thorough understanding of the target task and the source tasks, it's clear that the commonality between them is the need to assess and compare likelihoods or probabilities based on given premises, statements, or observations. The differences lie in the specific contexts and types of choices or options involved.

Step 2: Identify General Experience

From Source Task 1:
- Align the options with the provided premise and context
- Evaluate the logical connection between the premise and the choices
- Consider the likelihood of each choice based on the given context
- Focus on the most probable outcome based on the context provided
- Use logical reasoning to evaluate the relevance and probability of the options based on the context
- Select the option that aligns with the premise and logically addresses the given context
- Ensure that the chosen option leads to a probable and logical outcome based on the given context

From Source Task 2:
- Align the choices with the provided premise and context
- Evaluate the logical connection between the premise and the choices
- Consider the likelihood of each choice based on the given context
- Avoid introducing irrelevant or unrelated information in the choices
- Focus on the most probable outcome based on the context provided
- Use logical reasoning to assess the relevance of each choice in addressing the given context
- Choose the option that is most likely to result from the provided premise
- Evaluate the implications and consequences of each choice based on the given context for a thorough analysis

From Source Task 3:
- Carefully evaluate the paradox and consider the likelihood of each option based on the specific context
- Align the options with the paradox and evaluate their logical connection to provide new information or a different perspective
- Use logical reasoning to assess the relevance of each option in resolving the paradox
- Focus on selecting the option that offers a new insight or perspective to reconcile the conflicting statements or observations
- Ensure the chosen options offer clear and logical explanations that reconcile the conflicting statements in the given context

Step 3: Experience Adaptation

Adapting the general experience identified in Step 2 to the target task, the following insights can be applied:

```json
{
  "How to better accomplish the task or avoid low-quality responses": [
    "Align the choices with the given premise and context",
    "Evaluate the logical connection between the premise and the choices",
    "Consider the likelihood of each outcome based on the provided premise",
    "Avoid introducing irrelevant or unrelated information in the choices or options",
    "Use logical reasoning to assess the relevance and probability of the options based on the context",
    "Select the option that aligns with the premise and logically addresses the given context to lead to probable outcomes",
    "Carefully evaluate the given premise before evaluating the choices or options"
  ],
  "The specific process for handling this task": [
    "Carefully read and understand the provided premise and context",
    "Evaluate each choice or option based on its logical connection to the premise",
    "Consider the likelihood of each outcome based on the given premise and context",
    "Choose the most relevant and likely outcome based on the context and logical reasoning",
    "Consider the broader context or background information related to the premise",
    "Ensure that the chosen option leads to a probable and logical outcome based on the given context"
  ]
}
```

Example 7 [GPT-3.5]: Prompt 5 combines and deduplicates two sets of experience for the same task. After Prompt 4, we perform Prompt 5 to update experience.

INPUT:

```
<Target Task>
Assess and compare the likelihood of different outcomes based on a given premise
</Target Task>

<Existing Experience>
{
  "How to better accomplish the task or avoid low−quality responses": [
    "Always align the choices with the provided premise and context",
    "Evaluate the logical connection between the premise and the choices",
    "Consider the likelihood of each choice based on the given context",
    "Avoid introducing irrelevant or unrelated information in the choices",
    "Focus on the most probable outcome based on the context provided",
    "Carefully read and understand the premise before evaluating the choices",
    "Consider the broader context or background information related to the premise",
    "Use reasoning and critical thinking skills to determine the most appropriate answer based on the given premise",
    "Eliminate choices that do not logically follow from the premise to narrow down the options",
    "Ensure that the selected choice directly addresses the situation outlined in the premise and provides a specific, relevant,
     and detailed response",
    "Thoroughly explain the reasoning process, including the steps of thought and the evaluation of the choices in relation to
     the given premise",
    "Use logical reasoning to evaluate the relevance and probability of the options based on the context",
    "Prioritize choices that align with the given premise and are more likely based on logical reasoning",
    "Thoroughly evaluate the choices based on their logical connection to the premise to assess causality",
    "Choose the option that is most likely to result from the provided premise, based on the logical connection and relevance
     to the given context",
    "Ensure the selected option aligns with the specific attributes and context of the given task",
    "Align the choices with the given premise and context",
    "Evaluate the logical connection between the premise and the choices",
    "Consider the likelihood of each outcome based on the provided premise",
    "Avoid introducing irrelevant or unrelated information in the choices or options",
    "Use logical reasoning to assess the relevance and probability of the options based on the context",
    "Select the option that aligns with the premise and logically addresses the given context to lead to probable outcomes",
    "Carefully evaluate the given premise before evaluating the choices or options"
  ],
  "Task Processing Flow 1": [
    "Step 1: Carefully read and understand the given premise or statement",
    "Step 2: Evaluate the choices based on their logical connection to the premise to assess causality",
    "Step 3: Eliminate choices that do not logically follow from the given premise to narrow down the options",
    "Step 4: Choose the option that is most likely to result from the provided premise, based on the logical connection and
    relevance to the given context",
    "Step 5: Ensure the selected option aligns with the specific attributes and context of the given task",
    "Step 6: Provide a specific, detailed, and comprehensive explanation of the thought process, including the logical
    evaluation of choices and the elimination of low−quality responses"
  ],
  "Task Processing Flow 2": [
    "Carefully read and understand the provided premise and context",
    "Evaluate each choice or option based on its logical connection to the premise",
    "Consider the likelihood of each outcome based on the given premise and context",
    "Choose the most relevant and likely outcome based on the context and logical reasoning",
    "Consider the broader context or background information related to the premise",
    "Ensure that the chosen option leads to a probable and logical outcome based on the given context"
  ]
}
</Existing Experience>
```

You are an excellent experience refiner. Please help me refine the above existing experiences related to the target task.
1. For "How to better accomplish the task or avoid low−quality responses", please integrate insights by combining those that are closely related and eliminating any repetitions
2. Please integrate the above "Task Processing Flow 1" and "Task Processing Flow 2" into one unified workflow process. Ensure that the primary goals and functionality of both original processes are preserved; Effectively resolve possible conflicts or overlaps between the two processes.

continued from the above content.

Use the following JSON format to output refined target task experience:

~~~
```json
{
"How to better accomplish the task or avoid low−quality responses": [ no more than 20 insights ],
"The specific process for handling this task": [ no more than 20 insights ]
}
```
~~~

RESPONSE:

~~~
```json
{
  "How to better accomplish the task or avoid low−quality responses": [
    "Always align the choices with the provided premise and context",
    "Evaluate the logical connection between the premise and the choices",
    "Consider the likelihood of each choice based on the given context",
    "Avoid introducing irrelevant or unrelated information in the choices",
    "Focus on the most probable outcome based on the context provided",
    "Carefully read and understand the premise before evaluating the choices",
    "Consider the broader context or background information related to the premise",
    "Use reasoning and critical thinking skills to determine the most appropriate answer based on the given premise",
    "Eliminate choices that do not logically follow from the premise to narrow down the options",
    "Ensure that the selected choice directly addresses the situation outlined in the premise and provides a specific, relevant, and detailed response",
    "Thoroughly explain the reasoning process, including the steps of thought and the evaluation of the choices in relation to the given premise",
    "Prioritize choices that align with the given premise and are more likely based on logical reasoning",
    "Thoroughly evaluate the choices based on their logical connection to the premise to assess causality",
    "Choose the option that is most likely to result from the provided premise, based on the logical connection and relevance to the given context",
    "Provide a specific, detailed, and comprehensive explanation of the thought process, including the logical evaluation of choices and the elimination of low−quality responses",
    "Avoid choices that do not directly connect to the premise or lack a logical and causal relationship to the given context",
    "Evaluate the choices based on their logical connection to the premise to assess likelihood"
  ],
  "The specific process for handling this task": [
    "Step 1: Carefully read and understand the given premise or statement",
    "Step 2: Evaluate the choices based on their logical connection to the premise and their likelihood",
    "Step 3: Eliminate choices that do not logically follow from the given premise to narrow down the options",
    "Step 4: Choose the option that is most likely to result from the provided premise, based on the logical connection and relevance to the given context",
    "Step 5: Ensure the selected option aligns with the specific attributes and context of the given task",
    "Step 6: Provide a specific, detailed, and comprehensive explanation of the thought process, including the logical evaluation of choices and the elimination of low−quality responses"
  ]
}
```
~~~

Example 8 [GPT-3.5]: Prompt 6 generates a new question of the target task type based on the reference web text. After Prompt 3, 4, and 5, the Experience Transfer is finished. Subsequently, we carry out Autonomous Practice by repeatedly executing prompts 6, 7, and 8 to generate multiple demonstrations.

INPUT:

<Reference Text>

Traumatic bone cyst, also called a simple bone cyst, is a condition of the jaws. It is more likely to affect men and is more likely to occur in people in their first and second decades. There is no known cause though it is sometimes related to trauma. It appears on radiographs as a well−circumscribed radiolucency (dark area), and it commonly scallops between the roots of teeth. When the lesion is surgically opened, an empty cavity is found.

continued from the above content.

One study showed female predominance.
</Reference Text>

<Example Question>
Premise: Tom has a fever.
Choice A: They are TIPs.
Choice B: Doctors think Tom's fever is caused by infection.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": /* one of A, B */
}
```
</Example Question>

<Task Type of the Example Question>
Assess and compare the likelihood of different outcomes based on a given premise
</Task Type of the Example Question>

You are an excellent questioner.
Please carefully read the reference text provided above and formulate a new question based on it.
The new question must maintain the same expression style, structure, and required output format as the example question.
The new question must belong to the same task type of the example question.
The new question must be well−defined, with a complete and clear description that can be answered and at least one correct answer exists.
You are forbidden from providing answers to your new question.
Use the following format to output your answer:
<New Question>
/* Your new question. */
</New Question>

RESPONSE:
<New Question>
Premise: John has discovered a well−circumscribed radiolucency in his jaw on a radiograph.
Choice A: The diagnosis revealed a traumatic bone cyst.
Choice B: He needs a root canal treatment.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": "A"
}
```
</New Question>

Example 9 [GPT-3.5]: Prompt 7 generates a thought process and answer to the generated new question based on experience.

INPUT:
<Task Experience>
```
{
  "How to better accomplish the task or avoid low−quality responses": [
    "Always align the choices with the provided premise and context",
    "Evaluate the logical connection between the premise and the choices",
    "Consider the likelihood of each choice based on the given context",
    "Avoid introducing irrelevant or unrelated information in the choices",
```

continued from the above content.

```
    "Focus on the most probable outcome based on the context provided",
    "Carefully read and understand the premise before evaluating the choices",
    "Consider the broader context or background information related to the premise",
    "Use reasoning and critical thinking skills to determine the most appropriate answer based on the given premise",
    "Eliminate choices that do not logically follow from the premise to narrow down the options",
    "Ensure that the selected choice directly addresses the situation outlined in the premise and provides a specific, relevant, and detailed response",
    "Thoroughly explain the reasoning process, including the steps of thought and the evaluation of the choices in relation to the given premise",
    "Prioritize choices that align with the given premise and are more likely based on logical reasoning",
    "Thoroughly evaluate the choices based on their logical connection to the premise to assess causality",
    "Choose the option that is most likely to result from the provided premise, based on the logical connection and relevance to the given context",
    "Provide a specific, detailed, and comprehensive explanation of the thought process, including the logical evaluation of choices and the elimination of low−quality responses",
    "Avoid choices that do not directly connect to the premise or lack a logical and causal relationship to the given context",
    "Evaluate the choices based on their logical connection to the premise to assess likelihood"
  ],
  "The specific process for handling this task": [
    "Step 1: Carefully read and understand the given premise or statement",
    "Step 2: Evaluate the choices based on their logical connection to the premise and their likelihood",
    "Step 3: Eliminate choices that do not logically follow from the given premise to narrow down the options",
    "Step 4: Choose the option that is most likely to result from the provided premise, based on the logical connection and relevance to the given context",
    "Step 5: Ensure the selected option aligns with the specific attributes and context of the given task",
    "Step 6: Provide a specific, detailed, and comprehensive explanation of the thought process, including the logical evaluation of choices and the elimination of low−quality responses"
  ]
}
```

</Task Experience>
Please refer to the above experience to answer the following question.

Premise: John has discovered a well−circumscribed radiolucency in his jaw on a radiograph.
Choice A: The diagnosis revealed a traumatic bone cyst.
Choice B: He needs a root canal treatment.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:

```json
{
  "correct choice ID": "A"
}
```

Please provide specific, detailed, and comprehensive steps of your thought.

RESPONSE:

```json
{
  "correct choice ID": "A"
}
```

My thought process for choosing choice A as the more likely option is as follows:
Step 1: I carefully read and understood the given premise, which is that John has discovered a well−circumscribed radiolucency in his jaw on a radiograph. This indicates a specific medical condition that needs to be diagnosed.
Step 2: I evaluated choice A and choice B based on their logical connection to the premise and their likelihood. Choice A, " The diagnosis revealed a traumatic bone cyst," aligns with the premise of a discovered radiolucency in the jaw, indicating a potential diagnosis related to bone abnormalities.
Step 3: I eliminated choice B, "He needs a root canal treatment," as it does not logically follow from the given premise of a discovered radiolucency in the jaw. A root canal treatment is related to dental issues, which is not the focus of the given premise.

continued from the above content.

Step 4: I chose option A, "The diagnosis revealed a traumatic bone cyst," as the most likely to result from the provided premise, based on the logical connection and relevance to the given context of a discovered radiolucency in the jaw.
Step 5: I ensured that the selected option aligns with the specific attributes and context of the given task, which is the discovery of a radiolucency in the jaw.
Step 6: I provided a specific, detailed, and comprehensive explanation of my thought process, including the logical evaluation of choices and the elimination of low−quality responses, to support my decision.

Example 10 [GPT-3.5]: Prompt 8 checks if the response to the question is correct based on the reference text.

INPUT:
<Reference Text>
Traumatic bone cyst, also called a simple bone cyst, is a condition of the jaws. It is more likely to affect men and is more likely to occur in people in their first and second decades. There is no known cause though it is sometimes related to trauma. It appears on radiographs as a well−circumscribed radiolucency (dark area), and it commonly scallops between the roots of teeth. When the lesion is surgically opened, an empty cavity is found.

One study showed female predominance.
</Reference Text>

<Target Question>
Premise: John has discovered a well−circumscribed radiolucency in his jaw on a radiograph.
Choice A: The diagnosis revealed a traumatic bone cyst.
Choice B: He needs a root canal treatment.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": "A"
}
```
</Target Question>

<Reasoning Process and Answer>
```json
{
  "correct choice ID": "A"
}
```

My thought process for choosing choice A as the more likely option is as follows:

Step 1: I carefully read and understood the given premise, which is that John has discovered a well−circumscribed radiolucency in his jaw on a radiograph. This indicates a specific medical condition that needs to be diagnosed.

Step 2: I evaluated choice A and choice B based on their logical connection to the premise and their likelihood. Choice A, " The diagnosis revealed a traumatic bone cyst," aligns with the premise of a discovered radiolucency in the jaw, indicating a potential diagnosis related to bone abnormalities.

Step 3: I eliminated choice B, "He needs a root canal treatment," as it does not logically follow from the given premise of a discovered radiolucency in the jaw. A root canal treatment is related to dental issues, which is not the focus of the given premise.

Step 4: I chose option A, "The diagnosis revealed a traumatic bone cyst," as the most likely to result from the provided premise, based on the logical connection and relevance to the given context of a discovered radiolucency in the jaw.

continued from the above content.

Step 5: I ensured that the selected option aligns with the specific attributes and context of the given task, which is the discovery of a radiolucency in the jaw.

Step 6: I provided a specific, detailed, and comprehensive explanation of my thought process, including the logical evaluation of choices and the elimination of low–quality responses, to support my decision.
</Reasoning Process and Answer>

You are an outstanding checker, skilled at examining the reasoning process and the correctness of the answer of the target question based on the reference text.
Pay close attention to whether the reasoning process and the answer are consistent or inconsistent with the reference text.
Use the following JSON format to output your opinion:
```json
{
"correctness": /* "correct", "wrong" or "inconclusive" */
}
```

Let's think step by step.

RESPONSE:
```json
{
"correctness": "correct"
}
```

Example 11 [GPT-3.5]: Prompt 9 summarizes the task-solving experience from examples with correct or incorrect answers. After autonomous practice, we perform experience induction (Prompt 9) based on the generated demonstrations. Subsequently, Prompt 5 is invoked again to integrate the transferred and induced experiences (similar to Example 7).

INPUT:
You are an excellent experiential summarizer, adept at extracting task–solving insights from examples of the target task.
Here are several target task examples with correct or incorrect answers:
<Correct Example 1>
<Question>
Premise: John has discovered a well–circumscribed radiolucency in his jaw on a radiograph.
Choice A: The diagnosis revealed a traumatic bone cyst.
Choice B: He needs a root canal treatment.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": "A"
}
```
</Question>
<Reasoning Process and Answer>
```json
{
  "correct choice ID": "A"
}
```

My thought process for choosing choice A as the more likely option is as follows:

Step 1: I carefully read and understood the given premise, which is that John has discovered a well–circumscribed radiolucency in his jaw on a radiograph. This indicates a specific medical condition that needs to be diagnosed.

continued from the above content.

Step 2: I evaluated choice A and choice B based on their logical connection to the premise and their likelihood. Choice A, " The diagnosis revealed a traumatic bone cyst," aligns with the premise of a discovered radiolucency in the jaw, indicating a potential diagnosis related to bone abnormalities.

Step 3: I eliminated choice B, "He needs a root canal treatment," as it does not logically follow from the given premise of a discovered radiolucency in the jaw. A root canal treatment is related to dental issues, which is not the focus of the given premise.

Step 4: I chose option A, "The diagnosis revealed a traumatic bone cyst," as the most likely to result from the provided premise, based on the logical connection and relevance to the given context of a discovered radiolucency in the jaw.

Step 5: I ensured that the selected option aligns with the specific attributes and context of the given task, which is the discovery of a radiolucency in the jaw.

Step 6: I provided a specific, detailed, and comprehensive explanation of my thought process, including the logical evaluation of choices and the elimination of low−quality responses, to support my decision.
</Reasoning Process and Answer>
</Correct Example 1>

<Correct Example 2>
<Question>
Premise: After being diagnosed with craniopharyngioma, John experienced a decline in overall health.
Choice A: John's condition worsened over time.
Choice B: John's condition improved significantly.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": "A"
}
```
</Question>

<Reasoning Process and Answer>
{
  "correct choice ID": "A"
}

Step 1: The premise states that after being diagnosed with craniopharyngioma, John experienced a decline in overall health. This indicates a negative impact on John's condition.
Step 2: Choice A states that John's condition worsened over time, aligning with the premise of a decline in overall health. This is more likely to result from the provided premise.
Step 3: Choice B, stating that John's condition improved significantly, does not logically follow from the given premise of a decline in overall health and is less likely based on the context provided.
Step 4: Therefore, based on the logical connection and relevance to the given context, choice A is the most likely outcome.
</Reasoning Process and Answer>
</Correct Example 2>

<Correct Example 3>
<Question>
Premise: Sarah was experiencing symptoms of faintness, nervousness, and muscle spasm.
Choice A: The doctor recommended pelvic massage for treatment.
Choice B: She was diagnosed with a stomach infection.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": "A"
}
```
</Question>

continued from the above content.

<Reasoning Process and Answer>
```json
{
  "correct choice ID": "B"
}
```

Steps of thought:
1. The premise describes symptoms of faintness, nervousness, and muscle spasm, which are more likely to be related to a medical condition rather than a treatment recommendation.
2. Choice A, recommending pelvic massage, does not directly address the symptoms presented in the premise and lacks a logical and causal relationship to the given context.
3. Choice B, diagnosing a stomach infection, aligns with the symptoms of faintness, nervousness, and muscle spasm and is more likely to result from the provided premise based on logical reasoning.
~~~
</Reasoning Process and Answer>
</Correct Example 3>

<Correct Example 4>
<Question>
Premise: A 12−year−old girl experiences abnormal vaginal bleeding and cyclical pelvic pain.
Choice A: The diagnosis revealed estrogenization during puberty.
Choice B: The diagnosis showed an infection or abnormal vaginal bleeding.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:
```json
{
  "correct choice ID": "B"
}
~~~
</Question>
<Reasoning Process and Answer>
```json
{
  "correct choice ID": "B"
}
```

Steps of Thought:
1. The premise mentions abnormal vaginal bleeding and cyclical pelvic pain in a 12−year−old girl. This suggests a medical condition related to the reproductive system or hormonal changes.
2. Choice A states that the diagnosis revealed estrogenization during puberty, which seems unlikely as abnormal vaginal bleeding and cyclical pelvic pain are not typical symptoms of estrogenization during puberty.
3. Choice B suggests that the diagnosis showed an infection or abnormal vaginal bleeding, which aligns more closely with the symptoms described in the premise.
4. Given the symptoms provided in the premise, the likelihood of an infection or abnormal vaginal bleeding as the diagnosis seems more plausible and relevant.
5. Therefore, based on the logical connection and relevance to the given context, Choice B is the more likely option to result from the provided premise.
</Reasoning Process and Answer>
</Correct Example 4>

<Incorrect Example 1>
<Question>
Premise: Sarah experienced significant pain during sexual intercourse.
Choice A: The diagnosis revealed that she had dyspareunia.
Choice B: She became pregnant.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:

continued from the above content.

~~~
```json
{
  "correct choice ID": "B"
}
```
</Question>
<Reasoning Process and Answer>
```json
{
  "correct choice ID": "B"
}
```
~~~

Based on the provided premise, which states that Sarah experienced significant pain during sexual intercourse, the more likely choice is "B: She became pregnant." This is because experiencing pain during sexual intercourse is a common symptom of pregnancy, particularly in the early stages when there are changes in the body and the uterus is expanding. On the other hand, "Choice A: The diagnosis revealed that she had dyspareunia" is less likely as dyspareunia is a medical condition characterized by persistent or recurrent pain during sexual intercourse, and the premise does not mention any prior diagnosis of this condition. Therefore, "Choice B" is the more likely result based on the given premise.
</Reasoning Process and Answer>
</Incorrect Example 1>

Based on the examples provided above, please follow the steps below to summarize the experience:

Step1: Observe and Analyze the Examples
Summarize the commonalities in the correct examples, identify patterns in the incorrect examples, and compare the differences between the correct and incorrect examples.

Step2: Summarize Experience
Based on the observations and analysis from the Step1, summarize task−solving insights.
Ensure that the insights provided are CLEAR, DETAILED, and are GENERALLY APPLICABLE to unseen examples of the target task.
Use the following JSON format to output the summarized experience:

~~~
```json
{
"How to better accomplish the task or avoid low−quality responses": [ no more than 20 insights ],
"The specific process for handling this task": [ no more than 20 insights ]
}
```
~~~

Let's think step by step.

RESPONSE:
Step1: Observe and Analyze the Examples

Commonalities in the correct examples:
- In all correct examples, the reasonings provided for selecting the correct choice are detailed and logical.
- The correct choices in all examples are directly related to the given premises and logically aligned with the context.

Patterns in the incorrect examples:
- The reasoning for the incorrect choices is not always based on the logical connection to the premise.
- Incorrect choices in the examples often lack a direct relationship to the given premise and are less likely based on the context provided.

Differences between the correct and incorrect examples:
- Correct examples provide a step−by−step logical evaluation of the choices based on the given premises, while incorrect examples lack the same level of detailed reasoning.
- Correct choices align directly with the context provided in the premise, while incorrect choices often lack a clear connection to the given premise.

Step2: Summarize Experience

continued from the above content.

~~~
```json
{
  "How to better accomplish the task or avoid low−quality responses": [
    "Carefully read and understand the given premise before evaluating the choices",
    "Evaluate each choice based on its logical connection to the premise and its relevance to the context",
    "Eliminate choices that do not align with the symptoms or conditions described in the premise",
    "Avoid making assumptions without proper evidence or logical reasoning",
    "Provide a specific, detailed, and comprehensive explanation of the thought process to support the decision"
  ],
  "The specific process for handling this task": [
    "Carefully read and understand the given premise before evaluating the choices.",
    "Evaluate each choice based on its logical connection to the premise and its likelihood of resulting from the provided
    information.",
    "Eliminate choices that do not align with the context of the given premise or lack a logical relationship to the provided
    information.",
    "Choose the option that directly relates to the context given in the premise and logically follows from the provided
    information.",
    "Provide a specific and detailed explanation of the thought process, including the logical evaluation of choices and the
    elimination of low−quality responses, to support the decision."
  ]
}
```
~~~

Example 12 [GPT-3.5 and GPT-4]: Prompt 10 thinks the user question based on experience and respond to the user. Finally, we use the updated experiences from experience transfer and induction for inference.

INPUT:
<Experience>
[How to better accomplish the task or avoid low−quality responses]:
– Carefully read and understand the given premise before evaluating the choices
– Evaluate each choice based on its logical connection to the premise and its relevance to the context
– Eliminate choices that do not align with the symptoms or conditions described in the premise
– Avoid making assumptions without proper evidence or logical reasoning
– Provide a specific, detailed, and comprehensive explanation of the thought process to support the decision
[The specific process for handling this task]:
– Carefully read and understand the given premise before evaluating the choices.
– Evaluate each choice based on its logical connection to the premise and its likelihood of resulting from the provided information.
– Eliminate choices that do not align with the context of the given premise or lack a logical relationship to the provided information.
– Choose the option that directly relates to the context given in the premise and logically follows from the provided information.
– Provide a specific and detailed explanation of the thought process, including the logical evaluation of choices and the elimination of low−quality responses, to support the decision.
</Experience>
Please refer to the above experience to answer the following question.

Premise: Tom has a fever.
Choice A: They are TIPs.
Choice B: Doctors think Tom's fever is caused by infection.
For the given two choices, choose the one that is more likely to result from the provided premise.
Use the following JSON format to output your answer:

~~~
```json
{
  "correct choice ID": /∗ one of A, B ∗/
}
```
~~~

continued from the above content.

RESPONSE of GPT-3.5:
1. Read and understand the premise: Tom has a fever.
2. Evaluate Choice A: "They are TIPs." This choice does not align with the premise of Tom having a fever, and there is no logical connection between the two. Therefore, Choice A is not likely to result from the provided premise and should be eliminated.
3. Evaluate Choice B: "Doctors think Tom's fever is caused by infection." This choice directly relates to the context given in the premise of Tom having a fever, and it logically follows that doctors would consider an infection as the cause. Therefore, Choice B is more likely to result from the provided premise and is the correct choice.
4. Provide a specific and detailed explanation of the thought process, including the logical evaluation of choices and the elimination of low−quality responses, to support the decision.

```
{
  "correct choice ID": "B"
}
```

RESPONSE of GPT-4:
To handle the task, I will use the outlined process:

1. ∗∗Careful Reading and Understanding∗∗:
   I first read the given premise, which is "Tom has a fever."

2. ∗∗Logical Connection to the Premise∗∗:
   − Choice A: They are TIPs.
   − Choice B: Doctors think Tom's fever is caused by infection.

3. ∗∗Evaluate Relevance and Logical Relationship∗∗:
   − For Choice A, "They are TIPs" doesn't logically connect to the premise about Tom's fever. It's not clear what "TIPs" refers to, and there is no direct correlation between "TIPs" and having a fever.
   − For Choice B, it's common medical knowledge that fevers are often symptoms of infections. Therefore, it is logical for doctors to think that Tom's fever could be caused by an infection.

4. ∗∗Elimination of Choices∗∗:
   − I eliminate Choice A because it has no apparent relevance or logical connection to the premise of Tom having a fever.
   − I do not eliminate Choice B because it is logically consistent with common medical knowledge about fevers and infections.

5. ∗∗Choose the Option with Direct Relation and Logical Follow−through∗∗:
   − Choice B is the option with a direct relation to the premise and is a logical assumption based on the provided information.

6. ∗∗Detailed Explanation∗∗:
   The thought process began with an understanding of the premise, followed by an evaluation of each choice for its logical connection to the premise. Choice A was discarded due to its irrelevance and lack of logical connection. Choice B was consistent with medical knowledge and logically followed from the premise of having a fever, which can commonly be caused by an infection. Therefore, based on this reasoning:

```json
{
  "correct choice ID": "B"
}
```