STEP-BACK Prompting

⚽ We introduced a simple prompt technique called "STEP-BACK Prompting", which enables large language models (LLMs) to abstract and derive high-level concepts and basic principles from instances containing specific details. By using these concepts and principles to guide reasoning, the ability of large language models to find solutions along the correct reasoning path has been significantly improved. We conducted experiments on the PaLM-2L, GPT-4, and Llama2-70B models with "step back prompting" and observed significant performance improvements on various challenging intensive reasoning tasks including STEM, knowledge questioning, and multi-hop reasoning. For example, "step back prompting" improved the performance of PaLM-2L on MMLU (physics and chemistry) by 7% and 11%, respectively, on TimeQA by 27%, and on MuSiQue by 7%.

Inspired by **the fact that humans often take a step back to abstract thinking when facing challenging tasks** , in order to derive high-level principles to guide the problem-solving process, we propose the "STEP-BACK PROMPTING" method, which aims to reduce the possibility of errors in intermediate reasoning steps through **abstract reasoning** .

Introduction

This study explores how large language models (LLMs) handle complex tasks involving many low-level details through a two-step process of abstraction and inference.

The first step is to demonstrate to LLMs how to "step back" through contextual learning - using prompts to deduce high-level abstract concepts and principles for specific examples.

The second step is to use reasoning ability to reason on top of high-level concepts and principles. We use a small number of examples to demonstrate "STEP-BACK PROMPTING" on LLMs.

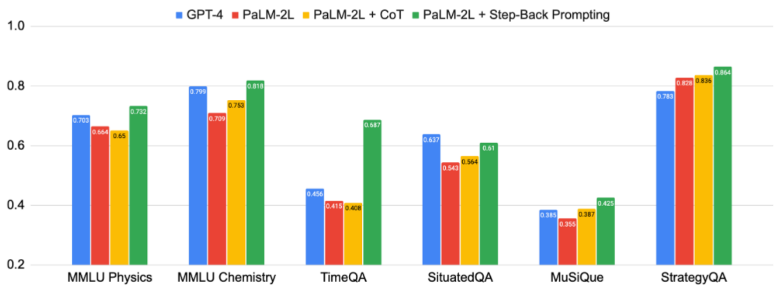

Figure 1: Strong Performance of STEP-BACK PROMPTING: our proposed Abstraction-and-Reasoning scheme leads to a substantial improvement in a wide range of challenging tasks in STEM, Knowledge QA and Multi-Hop Reasoning requiring complex (often multi-hop) reasoning.

Figure 1 summarizes all the key results presented in this article. Some tasks are very challenging: on TimeQA and MuSiQue, the accuracy of PaLM-2L and GPT-4 is only about 40%. Chain-of-Thought prompting only brings slight improvements on some tasks, while "Step Back Prompting" comprehensively improves the performance of PaLM-2L: 7% and 11% respectively on MMLU physics and chemistry, 27% on TimeQA, and 7% on MuSiQue.

We conducted various analyses and found that the "step back prompt" brought significant performance improvements (up to 36%) compared to the thought chain prompt (CoT) (Wei et al., 2022b) and the "deep breath" (TDB) prompt (Yang et al., 2023). We conducted qualitative evaluations and found that the "step back prompt" corrected most of the errors in the basic model (up to about 40%), while introducing fewer new errors (up to about 12%). We also conducted error analysis and found that most of the errors made by the "step back prompt" were attributed to the inherent limitations of LLMs' reasoning ability, while abstract skills were relatively easy to demonstrate to LLMs, indicating the direction of future improvements similar to the "step back prompt" method.

Method

The inspiration for "STEP-BACK PROMPTING" comes from observing that many tasks contain a lot of details, and large language models (LLMs) find it difficult to retrieve relevant facts to process these tasks. As shown in the first example (top) in Figure 2, for a physical problem "What will happen to the pressure P of an ideal gas if its temperature increases by 2 times and its volume increases by 8 times?", LLM may deviate from the first principle of the ideal gas law when reasoning directly about the problem. Similarly, a question "Which school did Estella Leopold go to between August and November 1954?" is very difficult to answer directly due to the detailed time frame constraints. In both cases, asking a "step back" question helps the model effectively solve the problem.

We define the step-back question as a question derived from the original question at a higher level of abstraction. For example, instead of directly asking "which school Estella Leopold went to during a specific time period," a step-back question (as shown at the bottom of Figure 2) will ask about "educational history," which is a high-level concept that covers the original question. In this case, answering the step-back question about "Estella Leopold's educational history" will provide all the necessary information to reason about "which school Estella Leopold went to during a specific time period." The premise is that the step-back question is usually much easier. Reasoning based on facts at such an abstract level helps avoid reasoning errors in intermediate steps, as shown in the Chain-of-Thought example in Figure 2.

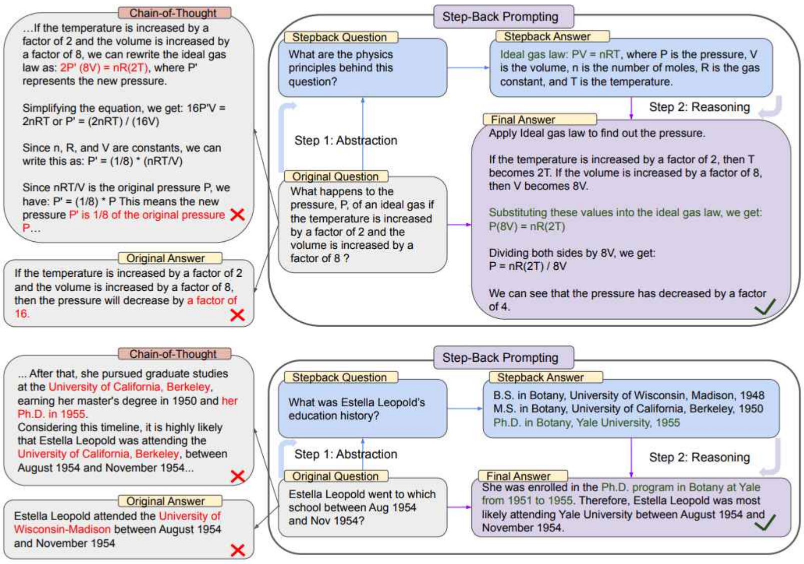

Figure 2: Illustration of STEP-BACK PROMPTING with two steps of Abstraction and Reasoning guided by concepts and principles. *Top*: an example of MMLU high-school physics (Hendrycks et al., 2020) where the first principle of Ideal Gas Law is retrieved via abstraction. *Bottom*: an example from TimeQA (Chen et al., 2021) where the high-level concept of education history is a result of the abstraction. *Left*: PaLM-2L (Anil et al., 2023) fails to answer the original question. Chain-of-Thought prompting (Wei et al., 2022b; Kojima et al., 2022) ran into errors during intermediate reasoning steps (highlighted as red). *Right*: PaLM-2L (Anil et al., 2023) successfully answers the question via STEP-BACK PROMPTING.

In short, the "Step Back Tip" includes two simple steps:

- Abstraction: Instead of dealing with the problem directly, we **first prompt LLM to ask a generic step back question about a higher level concept or principle and retrieve relevant facts about the higher level concept or principle** . The step back question for each task is unique in order to retrieve the most relevant facts.
- Reasoning: **Based on facts about high-level concepts or principles** , LLM can reason out the solution to the original problem. We call this abstract basic reasoning (Abstraction-grounded Reasoning).

Experiment

Baseline Methods

- **PaLM-2L, PaLM-2L 1-shot**: PaLM-2L is either queried directly with the question or has a single demonstration exemplar of question-answer included in the prompt.
- **PaLM-2L + CoT, PaLM-2L + CoT 1-shot**: PaLM-2L model is queried with zero-shot CoT prompting (Kojima et al., 2022): "*Let's think step by step*" is appended to the question. For 1-shot, One demonstration example of a question and answer pair is provided in the prompt, where the answer is in the style of CoT (Wei et al., 2022b).
- **PaLM-2L + TDB**: Zero-shot prompting with "*Take a deep breath and work on this problem step-by-step.*" (Yang et al., 2023) prepended to the question.
- **PaLM-2L + RAG**: For Sections 5 and 6, we use retrieval-augmented generation (RAG) where the retrieved passage is used as context by the LLM.
- **GPT-4 and Llama2-70B**: we run GPT-4 and Llama2-70B on MMLU tasks for all methods. In addition, we also run GPT-4 on all baselines for all tasks.

The questions in the MMLU benchmark test require deeper reasoning. In addition, they also require understanding and applying formulas, which are usually physical and chemical principles and concepts. In this case, we first demonstrate abstract skills to the model in the form of concepts and basic principles, such as Newton's first law, Doppler effect, Gibbs free energy, etc. The implicit step back question here is: "What physical or chemical principles and concepts are involved in solving this task?" We provide demonstrations to the model and let it repeat the relevant principles of solving the task based on its own Knowledge Base.

For MMLU High School Physics and Chemistry, we first prompt the basic principles behind the model generation problem. Using the basic principles of generation, we further prompt the model to generate the final answer through a small number of examples. The prompts used to generate the basic principles are shown in Table 7 of MMLU High School Physics and Chemistry.

MMLU Physics/Chemistry First-Principle Prompt
You are an expert at Physics/Chemistry. You are given a Physics/Chemistry problem. Your task is to extract the Physics/Chemistry concepts and principles involved in solving the problem. Here are a few examples: Question: <Question Example1> Principles Involved: <Principles Example1> ... Question: <Question Example5> Principles Involved: <Principles Example5> Question: <Question> Principles Involved:

Table 7: Prompt of extracting the underlying principles involved in MMLU physics and chemistry questions.

After extracting the basic principles for solving specific problems, we formulated the prompts in Table 8 to query the model to obtain the final answer.

MMLU Physics/Chemistry Final Answer Prompt
You are an expert at Physics/Chemistry. You are given a Physics/Chemistry problem and a set of principles involved in solving the problem. Solve the problem step by step by following the principles. Here are a few examples: Question: <Question Example1> Principles: <Principles Example1> Answer: <Answer Example1> ... Question: <Question Example5> Principles: <Principles Example5> Answer: <Answer Example5> Question: <Question> Principles: <Principles> Answer:

Table 8: Prompt of querying the model for final answer with first principles behind the question in MMLU high-school Physics and Chemistry.

Table 1 shows the average model performance of three model families: PaLM-2L, GPT-4, and Llama2-70B under different settings.

In terms of physics and chemistry, the performance of PaLM-2L base line is 66.4% and 70.9%, respectively.

We found that CoT (thought chain) and TDB (deep breathing) zero-shot prompts did not significantly improve model performance, which may be due to the difficulty and depth of reasoning inherent in these tasks. The PaLM-2L 1-shot example and PaLM-2L + CoT 1-shot example also did not show much improvement compared to the base line, highlighting the challenge of showing the inference steps to the model. In contrast, "STEP-BACK PROMPTING" significantly improved model performance: + 7% and + 11% compared to PaLM-2L. Similarly, when using the GPT4 and Llama2-70B models, "step back prompts" were very competitive among all the base line methods we tested, indicating that "step back prompts" are model-independent.

Table 1: Strong performance of STEP-BACK PROMPTING on MMLU tasks across three model families. CoT: zero-shot Chain of Thought prompting (Kojima et al., 2022), TDB: Take a Deep Breath prompting (Yang et al., 2023).

Method	MMLU Physics	MMLU Chemistry
PaLM-2L	66.4% (0.8%)	70.9% (0.9%)
PaLM-2L 1-shot	64% (1.6%)	75.6% (0.4%)
PaLM-2L + CoT	65% (2%)	75.3% (1.5%)
PaLM-2L + CoT 1-shot	61.5% (1.8%)	76.6% (1%)
PaLM-2L + TDB	65.7% (0.7%)	73.8% (1.1%)
PaLM-2L + Step-Back (ours)	**73.2%** (1.9%)	**81.8%** (1.4%)
GPT-4	69.4% (2.0%)	80.9% (0.7%)
GPT-4 1-shot	78.4% (2.4%)	80.5% (1.6%)
GPT-4 + CoT	82.9% (0.5%)	85.3% (1.0%)
GPT-4 + CoT 1-shot	79.3% (1.0%)	82.8% (0.5%)
GPT-4 + TDB	74.4% (4.0%)	81.5% (1.3%)
GPT-4 + Step-Back (ours)	**84.5%** (1.2%)	**85.6%** (1.4%)
Llama2-70B	51.9% (3.6%)	55.7% (2.1%)
Llama2-70B 1-shot	57.3% (1.6%)	58.5% (2.5%)
Llama2-70B + CoT	59.3% (2.0%)	64.1% (1.2%)
Llama2-70B + CoT 1-shot	59.6% (2.0%)	**68.1%** (1.4%)
Llama2-70B + TDB	60.4% (2.1%)	63.6% (1.9%)
Llama2-70B + Step-Back (ours)	**64.8%** (1.5%)	66.7% (1.6%)

Error analysis

Comparing the prediction results of the "Step Back Prompt" and the base line PaLM-2L model in the MMLU high school physics test, we found that the "Step Back Prompt" corrected 20.5% of the errors in the base line and introduced 11.9% of the errors.

In order to further understand the sources of errors in the "step back prompt", we annotated all the erroneous predictions of the "step back prompt" in the test set and classified them into 5 categories (see Appendix E.1 for examples of each category).

- Principle error: The error occurs in the abstract step, and the basic principles of model generation are incorrect or incomplete.

- Factual error: When the model restates its own factual knowledge, there is at least one factual error.
- Mathematical Error: In intermediate steps involving mathematical calculations, there is at least one mathematical error.
- Context loss: The model response loses the context in the problem and deviates from the answer to the original problem.
- Inference error: We define inference error as the model making at least one error in the intermediate inference step before reaching the final answer.

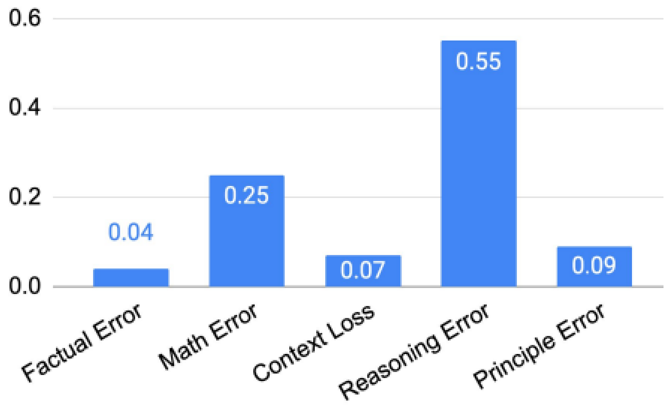

Figure 4: Error Analysis of STEP-BACK PROMPTING on MMLU high-school Physics: five classes of errors Step-Back makes with Reasoning being the dominating class.

In addition to principle errors pointing to the failure of abstract steps, all five types of errors occur in the reasoning step. As shown in Figure 4, principle errors account for only a small part of the errors made by the model: over 90% of errors occur in the reasoning step. Among the four types of errors in the reasoning process, reasoning errors and mathematical errors are the main categories of errors.

This is consistent with the findings of the aforementioned ablation study that there are very few examples required to demonstrate abstract skills to large language models (LLMs). The inference step remains a bottleneck in how well "STEP-BACK PROMPTING" performs on complex inference tasks such as MMLU. For MMLU physics in particular, inference and mathematical skills are essential for successful problem-solving: **Even if the basic principles are correctly retrieved, deep inference and mathematical operations through typical multi-step inference processes are required to arrive at the correct final answer** .

STEP-BACK PROMPTING Example

Question	A spherical conductor carries a net charge. How is this charge distributed on the sphere? 0) The charge is evenly distributed on the surface. 1) The charge resides on the surface only; the distribution of charge on the surface depends on what other charged objects are near the sphere. 2) The charge moves continually within the sphere. 3) The charge is distributed uniformly throughout the sphere.
Principles	Coulomb's Law: the force between two charged particles is proportional to the product of their charges and inversely proportional to the square of the distance between them, F = k * q1 * q2 / r, where F is the electric force, k is a constant, q1 and q2 are the charges the particles carry, and r is the distance between them.
Answer	Using the Principles of Coulomb's Law, we can solve the problem as following: Step 1: Apply Coulomb's Law to find out how charges are distributed on the surface. In the case of a spherical conductor, the charges on the surface will repel each other. The further apart the charges are, the less force they will exert on each other. Therefore, the charges will distribute themselves evenly on the surface of the sphere, as this is the configuration that minimizes the repulsive force between them. Step 2: Apply Coulomb's Law to find out what happens if there are other charges present. The distribution of charge on the surface may also be affected by the presence of other charged objects near the sphere. For example, if a negatively charged object is brought near a positively charged sphere, the negative charges on the sphere will be repelled and will move to the opposite side of the sphere. This will result in a non-uniform distribution of charge on the surface of the sphere. Therefore, the correct answer is 1) The charge resides on the surface only; the distribution of charge on the surface depends on what other charged objects are near the sphere.

Table 9: Few-shot demonstration exemplar for MMLU high-school Physics.

Conclusion

We introduced "STEP-BACK PROMPTING" as a simple and universal method to guide deep reasoning by abstracting in large language models. Experiments on large language models (LLMs) in fact-finding, common sense reasoning, and domain-specific reasoning benchmark tests have shown that "STEP-BACK PROMPTING" significantly improves model performance. We assume that abstraction helps models reduce illusions and reason better, which may reflect the true nature of the model, which is often hidden when responding to the original problem without abstraction. We hope that our work can inspire more human-inspired methods to explore the potential capabilities of large language models.

Further thinking

1. The abstract stage can be achieved with a few shots and proves that the model is capable of handling this task
2. Abstract + RAG, better selection of knowledge content
3. Mainly hope to further solve the problems in the reasoning stage, that is, how to strengthen the ability in the reasoning stage.

Current graduation project plan

Create a problem.

The main task of this stage is to enhance the reasoning ability of the model, so the format of the problem is:

Title + Reference (abstract question answer) + ReAlign answer (imitating the ReAlign format in the figure below)

Reformatted Alignment

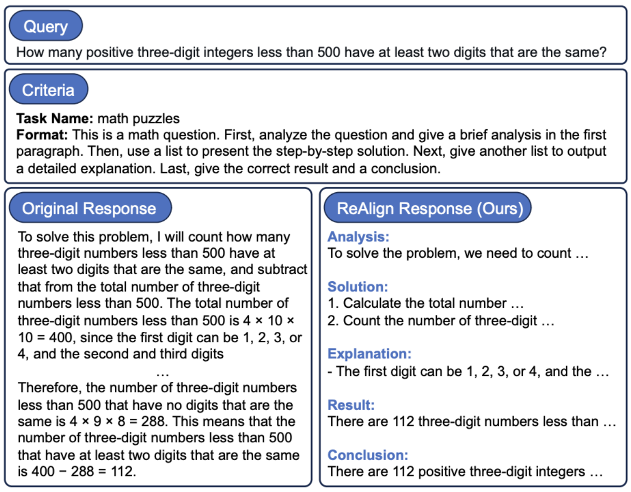

Figure 2: REALIGN realigns the original response with the pre-defined criteria to be a better format. The original response is from the `Open-Platypus` (Lee et al., 2023) dataset. The complete version is shown in Tab. 13.

Plan simultaneous construction, physics experimental problem + mmlu + gpqa problem

Fine-tune

Fine-tune the open source large language model on the constructed dataset, and divide the train and test sets. Evaluate the fine-tuned model on public datasets such as mmlu to test the model's capacity enhancement

Knowledge Base Construction

It is necessary to include knowledge of physics experiments, detailed physics formulas from high school and university, for easy retrieval and use

RAG implementation

Realize Retrieval Enhancement Generation on Constructed Knowledge Base

Prompt Engineering Framework Exploration

On the above-mentioned model after fine-tuning and RAG implementation, we further explore the thinking framework similar to STEP-BACK Prompting through prompt engineering, and improve the reasoning stage ability of LLM