TAKE A STEP BACK: EVOKING REASONING VIA ABSTRACTION IN LARGE LANGUAGE MODELS

Huaixiu Steven Zheng* **Swaroop Mishra*** **Xinyun Chen** **Heng-Tze Cheng**
Ed H. Chi **Quoc V Le** **Denny Zhou**

Google DeepMind

Present by: Junqi Qu

Reasoning remains hard for LLMs: motivation

- Large Language Models (LLMs) have shown emergent abilities, including multi-step reasoning, through scaling.
- Techniques like Chain-of-Thought (CoT) prompting aim to improve this by generating intermediate steps.
- **The Core Problem:** Despite progress, even state-of-the-art LLMs struggle with complex reasoning tasks.
- Models can get lost in low-level details, leading to errors in the reasoning path.

Abstraction-grounded reasoning

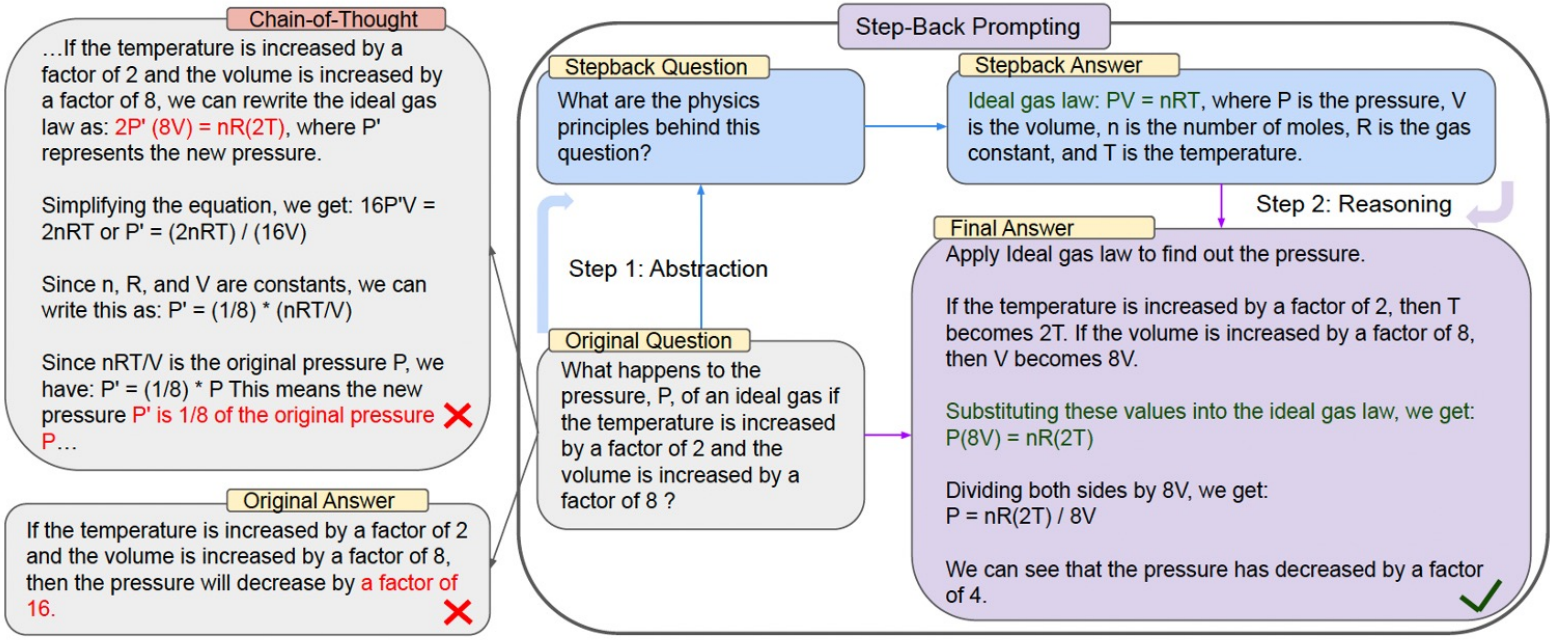
- **Inspiration:** Human problem-solving. We often "step back" from a difficult problem to consider the high-level principles involved.
- **The Proposal: Step-Back Prompting.**
  - A technique that enables LLMs to perform
  - **abstraction**—deriving high-level concepts and first principles from specific instances.
  - The model then uses these abstractions to guide its reasoning process, reducing the chance of error.

How it works: A Two-Step process

- Step-Back Prompting consists of two steps, demonstrated to the LLM via few-shot examples
- **Step 1: Abstraction**
  - The LLM is prompted to generate a more generic, "step-back question" about a higher-level concept or principle.
  - It then answers this abstract question to retrieve relevant facts.
- Step 2: Reasoning
  - Grounded on the high-level facts retrieved in Step 1, the LLM reasons through the steps to solve the original, specific question.
- **Abstraction-grounded Reasoning**.

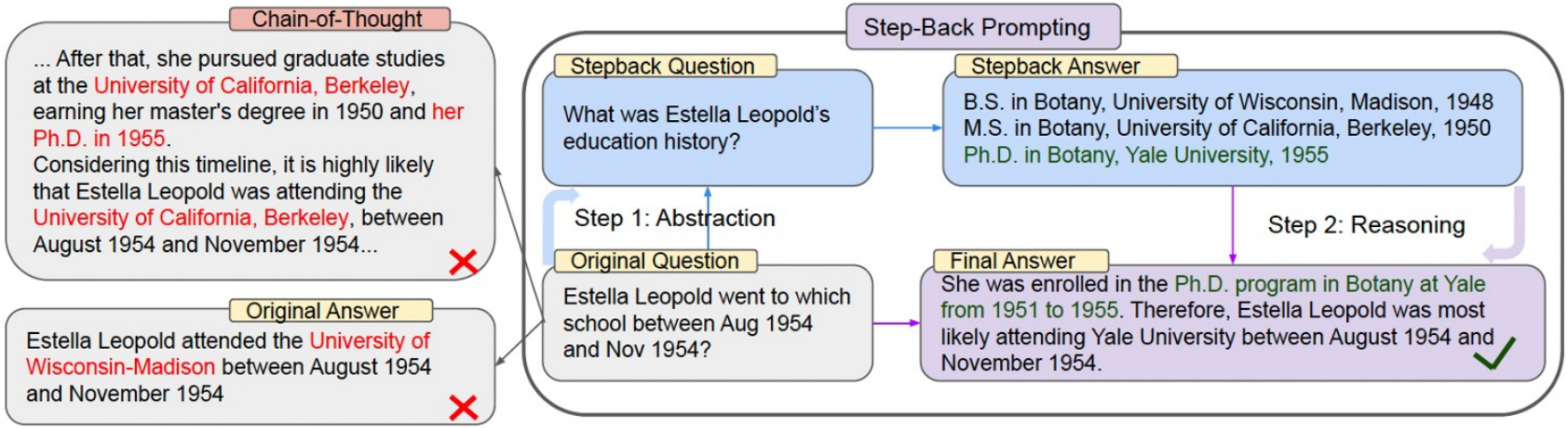
Example

- MMLU Physics



Another example

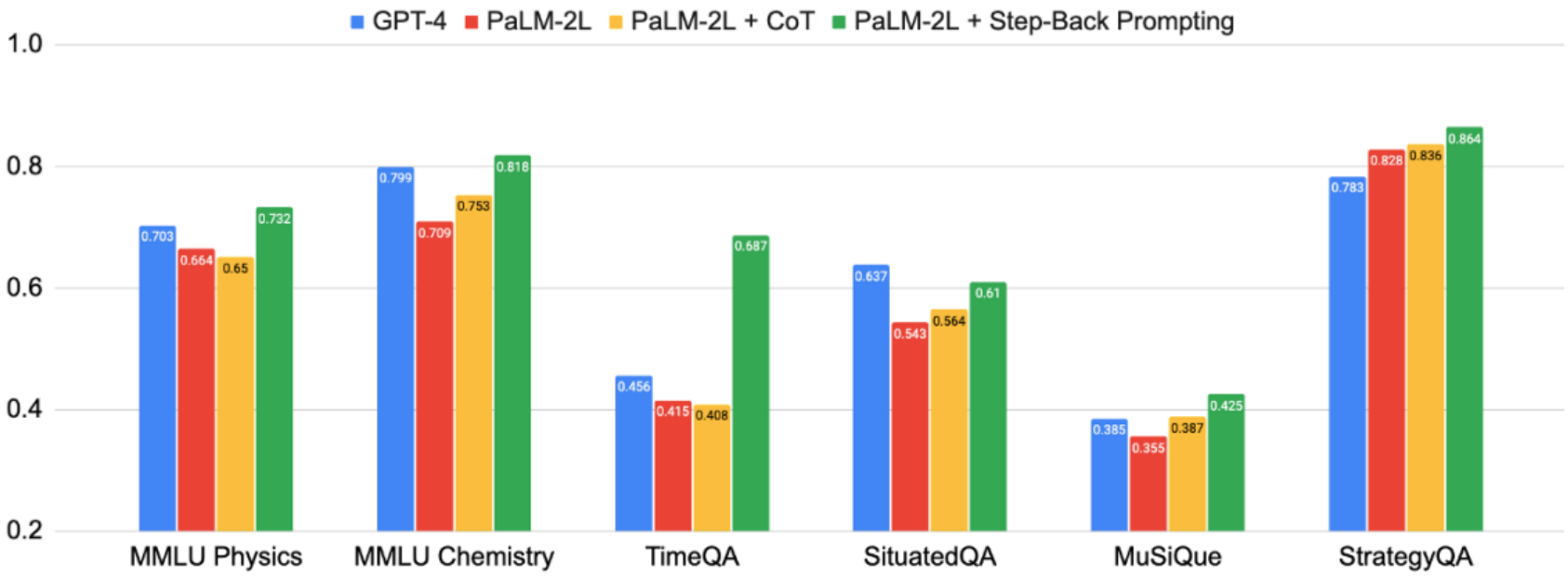
- Knowledge QA



Experimental setup

- **Models:** PaLM-2L, GPT-4, Llama2-70B.
- **Tasks:** A diverse set of challenging reasoning benchmarks.
  - **STEM:** MMLU (Physics, Chemistry), GSM8K.
  - **Knowledge QA:** TimeQA, SituatedQA.
  - **Multi-Hop Reasoning:** MuSiQue, StrategyQA.
- **Baselines:**
  - Standard zero-shot and one-shot prompting.
  - Chain-of-Thought (CoT).
  - "Take a Deep Breath" (TDB).
  - Retrieval-Augmented Generation (RAG) for knowledge-intensive tasks.

Results


GPT-4
PaLM-2L
PaLM-2L + CoT
PaLM-2L + Step-Back Prompting
1.0
0.8
0.6
0.4
0.2
0.703
0.664
0.65
0.732
0.799
0.709
0.753
0.818
0.456
0.415
0.408
0.687
0.637
0.543
0.564
0.61
0.385
0.355
0.387
0.425
0.783
0.828
0.836
0.864
MMLU Physics
MMLU Chemistry
TimeQA
SituatedQA
MuSiQue
StrategyQA

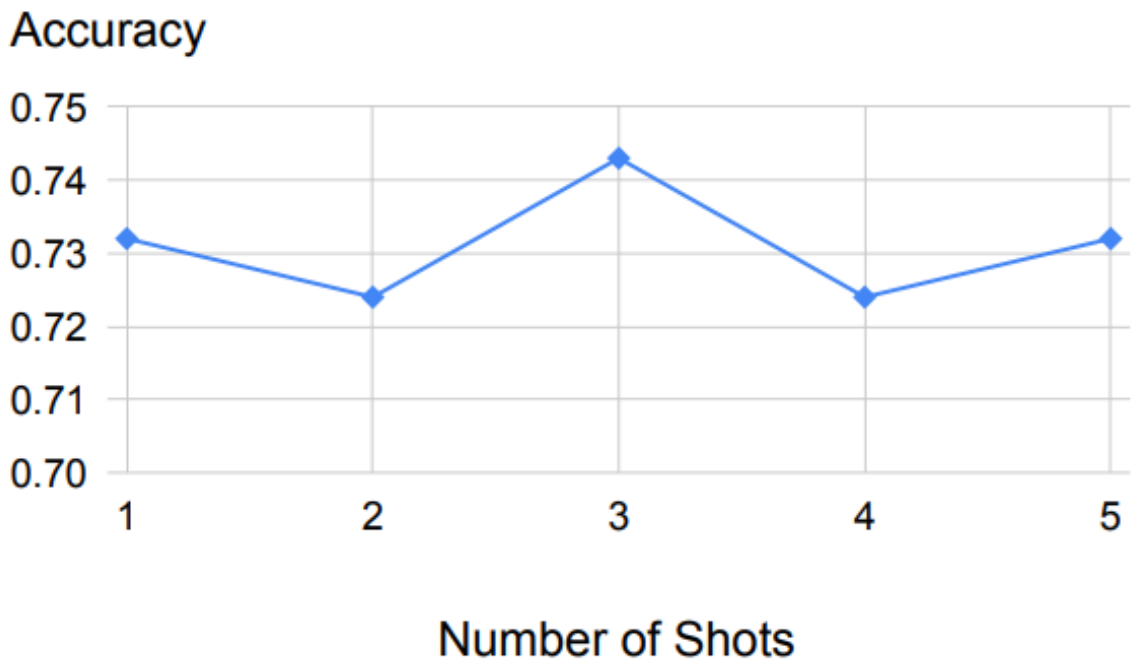
STEM tasks

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%)

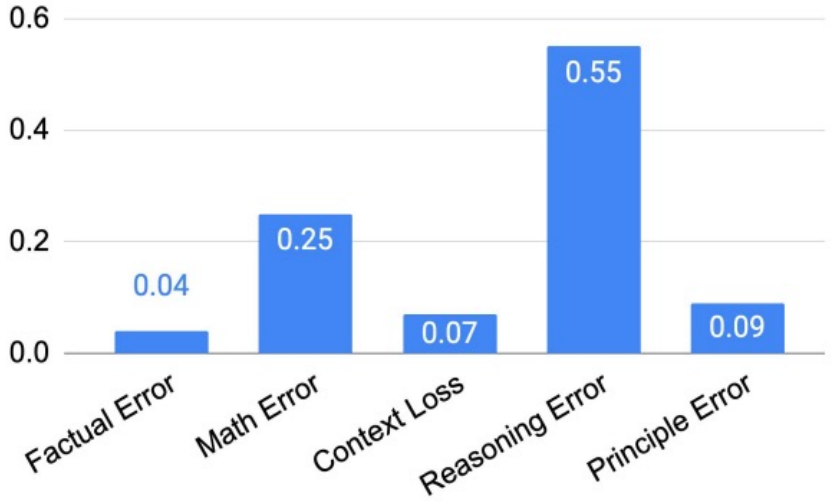
Ablation studies

Figure 3: Ablation study of STEP-BACK PROMPTING accuracy using PaLM-2L on MMLU high-school Physics against the number of few shot exemplars: robust performance with respect to a varying number of shots.



where do errors come from?

- **Principle Error**: The error happens at the step of Abstraction, where the first principles generated by models are wrong or incomplete.
- **Factual Error**: There is at least one factual error when the model recites its own factual knowledge
- **Math Error**: There is at least one math error in the intermediate steps when math calculations are involved in deriving the final answer.
- **Context Loss**: There is at least one error where the model response loses context from the question, and deviates from addressing the original question
- **Reasoning Error**: We define Reasoning Error as when the model makes at least one error in the intermediate Reasoning steps before arriving at the final answer.



Step-Back vs. Decomposition

- Step-Back Prompting is fundamentally different from problem decomposition methods.
- **Decomposition:** Breaks a problem into smaller, lower-level sub-problems. It's a **one-to-many** mapping, going deeper.
  - *Example:* "Who did Steve Jobs work for in 1990?" -> "Was he employed in 1990? If so, by whom?"
- **Abstraction (Step-Back):** Generalizes a specific question to a single, higher-level concept. It's a **many-to-one** mapping, going broader.
  - *Example:* "Who did Steve Jobs work for in 1990?" -> "What is Steve Jobs' employment history?"