

# Unlocking reasoning and planning abilities in Large language models

Logesh Kumar Umapathi

Lead Machine learning engineer @ Saama Technologies

Hi 🖐️

I am Logesh Kumar Umapathi

Lead ML Research Engineer @ Saama Technologies

**Twitter:** @logesh\_umapathi

**Linkedin:** [www.linkedin.com/in/logeshkumar/](https://www.linkedin.com/in/logeshkumar/)

**Website:** logeshumapathi.com

**OSS :** <https://github.com/infinitylogesh/mutate>



# Agenda

- What is reasoning?
- How is reasoning measured?
- Eliciting reasoning:
  - Direct prompting
  - Recursive and Iterative Prompting
  - Tool usage

## What is reasoning ?

Reasoning is the ability to make inferences using evidence and logic. Reasoning can be divided into multiple types of skills such as Commonsense ,Mathematical and , Symbolic reasoning etc. Often, reasoning involves deductions from inference chains, called as multi-step reasoning.

# How is reasoning measured ( in the literature) ?

- Mathematical reasoning:
  - [GSM8K](#)
  - [SVAMP](#)
  - [AQuA](#)
  - [MAWPS](#)
- Commonsense Reasoning:
  - [ARC](#)
  - [CSQA](#)
  - [StrategyQA](#)
- Symbolic Reasoning:
  - Last letter concatenation ([Wei et al., 2022b](#))
  - Coin flip ([Wei et al., 2022b](#))





# How is reasoning measured ( in the literature) ?

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
<b>MMLU [49]</b> Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	<b>70.0%</b> 5-shot	<b>70.7%</b> 5-shot U-PaLM [50]	<b>75.2%</b> 5-shot Flan-PaLM [51]
<b>HellaSwag [52]</b> Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	<b>85.5%</b> 10-shot	<b>84.2%</b> LLaMA (validation set) [28]	<b>85.6</b> ALUM [53]
<b>A12 Reasoning Challenge (ARC) [54]</b> Grade-school multiple choice science questions. Challenge-set.	<b>96.3%</b> 25-shot	<b>85.2%</b> 25-shot	<b>85.2%</b> 8-shot PaLM [55]	<b>86.5%</b> ST-MOE [18]
<b>WinoGrande [56]</b> Commonsense reasoning around pronoun resolution	<b>87.5%</b> 5-shot	<b>81.6%</b> 5-shot	<b>85.1%</b> 5-shot PaLM [3]	<b>85.1%</b> 5-shot PaLM [3]
<b>HumanEval [43]</b> Python coding tasks	<b>67.0%</b> 0-shot	<b>48.1%</b> 0-shot	<b>26.2%</b> 0-shot PaLM [3]	<b>65.8%</b> CodeT + GPT-3.5 [57]
<b>DROP [58] (F1 score)</b> Reading comprehension & arithmetic.	<b>80.9</b> 3-shot	<b>64.1</b> 3-shot	<b>70.8</b> 1-shot PaLM [3]	<b>88.4</b> QDGAT [59]
<b>GSM-8K [60]</b> Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	<b>57.1%</b> 5-shot	<b>58.8%</b> 8-shot Minerva [61]	<b>87.3%</b> Chinchilla + SFT+ORM-RL, ORM reranking [62]





**Table 2.** Performance of GPT-4 on academic benchmarks. We compare GPT-4 alongside the best SOTA (with benchmark-specific training) and the best SOTA for an LM evaluated few-shot. GPT-4 outperforms existing LMs on all benchmarks, and beats SOTA with benchmark-specific training on all datasets except DROP. For each task we report GPT-4’s performance along with the few-shot method used to evaluate. For GSM-8K, we included part of the training set in the GPT-4 pre-training mix (see Appendix E), and we use chain-of-thought prompting [11] when evaluating. For multiple-choice questions, we present all answers (ABCD) to the model and ask it to choose the letter of the answer, similarly to how a human would solve such a problem.

## Eliciting reasoning

L Take the last letters of the words in "Augusta Ada King" and concatenate them using a space ?





 The last letters of the words in "Augusta Ada King" are "a a g n g". Concatenating them using a space would give the string "a a g n g".   

L What are the words in "Augusta Ada King" sentence?

 The words in the sentence "Augusta Ada King" are:   



- Augusta
- Ada
- King

L What are the last letters of these words ?

 The last letters of the words in "Augusta Ada King" are:   

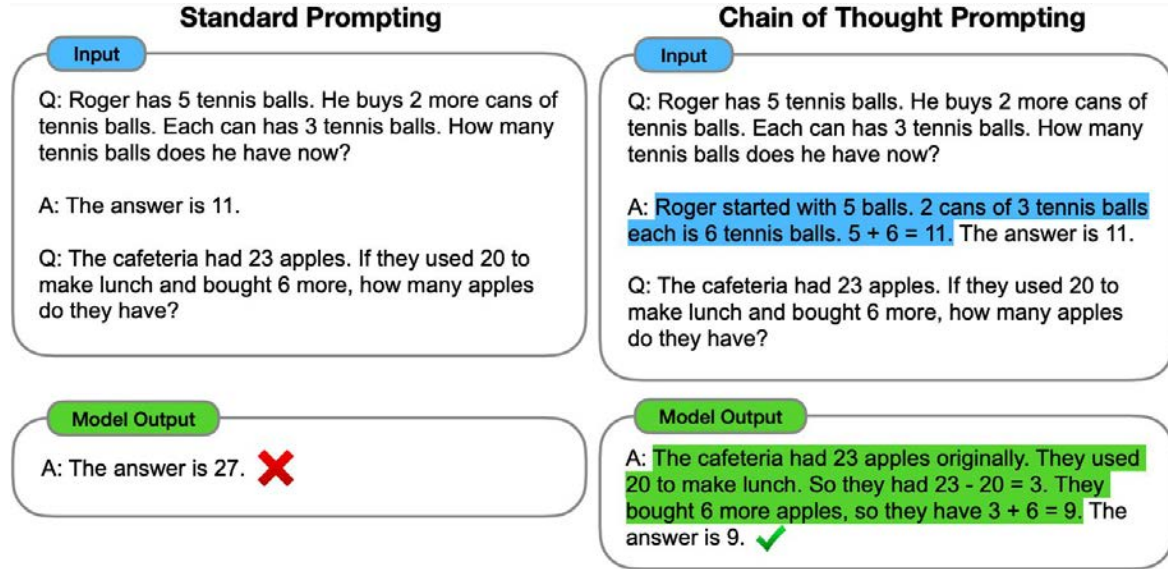
- Augusta ends with "a"
- Ada ends with "a"
- King ends with "g"

L Concatenate these letters with space

 If we concatenate the last letters of the words "Augusta Ada King" with space, we get the string "a a g".   



# Chain of thought prompting and Self consistency



# Chain of thought prompting

Table 2: Standard prompting versus chain of thought prompting on five arithmetic reasoning benchmarks. Note that chain of thought prompting is an emergent ability of model scale—it does not positively impact performance until used with a model of sufficient scale.

Model		GSM8K		SVAMP		ASDiv		AQuA		MAWPS	
		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	4.1	<b>4.4</b>	10.1	<b>12.5</b>	16.0	<b>16.9</b>	20.5	<b>23.6</b>	16.6	<b>19.1</b>
LaMDA	420M	2.6	0.4	2.5	1.6	3.2	0.8	23.5	8.3	3.2	0.9
	2B	3.6	1.9	3.3	2.4	4.1	3.8	22.9	17.7	3.9	3.1
	8B	3.2	1.6	4.3	3.4	5.9	5.0	22.8	18.6	5.3	4.8
	68B	5.7	<b>8.2</b>	13.6	<b>18.8</b>	21.8	<b>23.1</b>	22.3	20.2	21.6	<b>30.6</b>
	137B	6.5	<b>14.3</b>	29.5	<b>37.5</b>	40.1	<b>46.6</b>	25.5	20.6	43.2	<b>57.9</b>
GPT	350M	2.2	0.5	1.4	0.8	2.1	0.8	18.1	8.7	2.4	1.1
	1.3B	2.4	0.5	1.5	1.7	2.6	1.4	12.6	4.3	3.1	1.7
	6.7B	4.0	2.4	6.1	3.1	8.6	3.6	15.4	13.4	8.8	3.5
	175B	15.6	<b>46.9</b>	65.7	<b>68.9</b>	70.3	<b>71.3</b>	24.8	<b>35.8</b>	72.7	<b>87.1</b>
Codex	-	19.7	<b>63.1</b>	69.9	<b>76.4</b>	74.0	<b>80.4</b>	29.5	<b>45.3</b>	78.7	<b>92.6</b>
PaLM	8B	4.9	4.1	15.1	<b>16.8</b>	23.7	<b>25.2</b>	19.3	<b>21.7</b>	26.2	<b>30.5</b>
	62B	9.6	<b>29.9</b>	48.2	46.7	58.7	<b>61.9</b>	25.6	22.4	61.8	<b>80.3</b>
	540B	17.9	<b>56.9</b>	69.4	<b>79.0</b>	72.1	<b>73.9</b>	25.2	<b>35.8</b>	79.2	<b>93.3</b>

# Program-aided Language Models

## Chain-of-Thought (Wei et al., 2022)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves. They sold 93 in the morning and 39 in the afternoon. So they sold  $93 + 39 = 132$  loaves. The grocery store returned 6 loaves. So they had  $200 - 132 - 6 = 62$  loaves left. The answer is 62.



## Program-aided Language models (this work)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls.  
`tennis_balls = 5`  
 2 cans of 3 tennis balls each is  
`bought_balls = 2 * 3`  
 tennis balls. The answer is  
`answer = tennis_balls + bought_balls`

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves  
`loaves_baked = 200`  
 They sold 93 in the morning and 39 in the afternoon  
`loaves_sold_morning = 93`  
`loaves_sold_afternoon = 39`  
 The grocery store returned 6 loaves.  
`loaves_returned = 6`  
 The answer is  
`answer = loaves_baked - loaves_sold_morning`  
`- loaves_sold_afternoon + loaves_returned`

```
>>> print(answer)
74
```

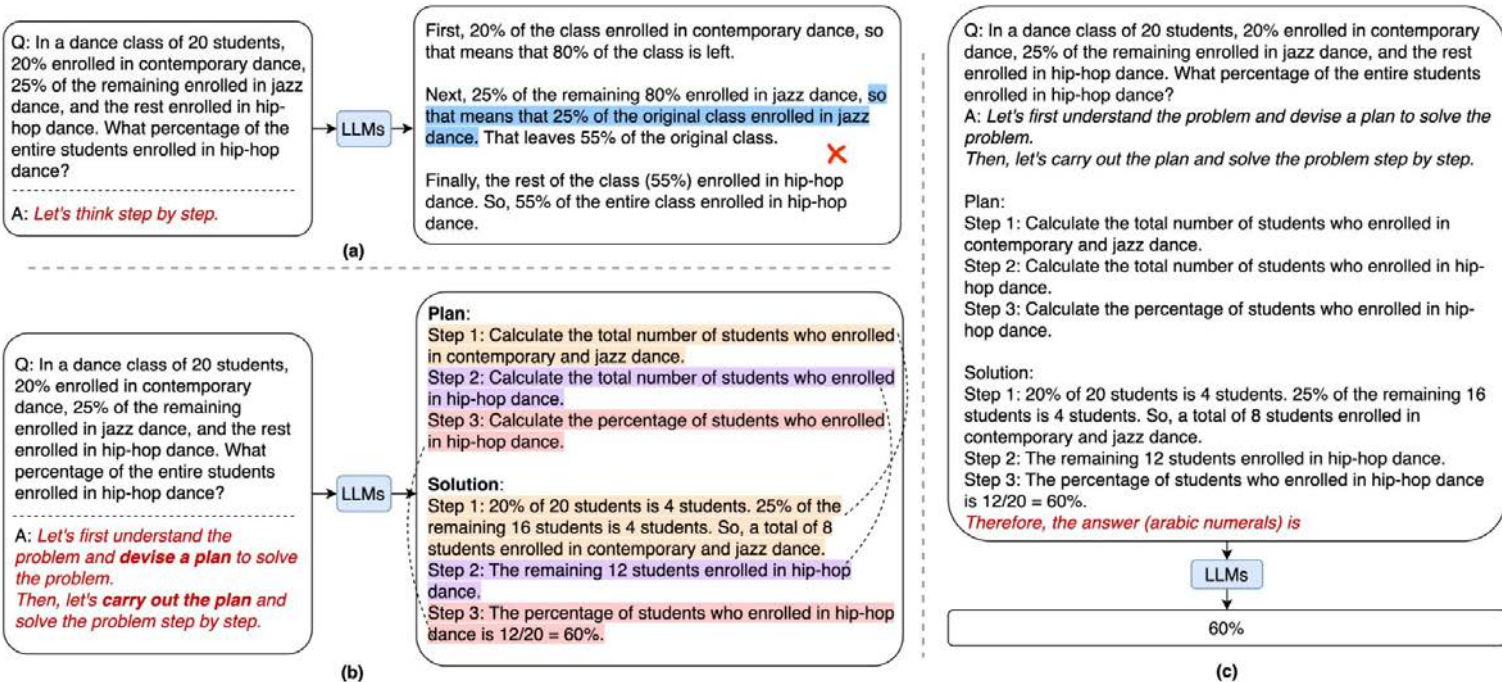


# Program-aided Language Models

	GSM8K	GSM-HARD	SVAMP	ASDIV	SINGLEEQ	SINGLEOP	ADDSUB	MULTIARITH
DIRECT <sub>Codex</sub>	19.7	5.0	69.9	74.0	86.8	93.1	90.9	44.0
COT <sub>UL2-20B</sub>	4.1	-	12.6	16.9	-	-	18.2	10.7
COT <sub>LaMDA-137B</sub>	17.1	-	39.9	49.0	-	-	52.9	51.8
COT <sub>Codex</sub>	65.6	23.1	74.8	76.9	89.1	91.9	86.0	95.9
COT <sub>PaLM-540B</sub>	56.9	-	79.0	73.9	92.3	94.1	91.9	94.7
COT <sub>Minerva 540B</sub>	58.8	-	-	-	-	-	-	-
PAL	<b>72.0</b>	<b>61.2</b>	<b>79.4</b>	<b>79.6</b>	<b>96.1</b>	<b>94.6</b>	<b>92.5</b>	<b>99.2</b>

Table 1: Problem solve rate (%) on mathematical reasoning datasets. The highest number on each task is in **bold**. The results for DIRECT and PaLM-540B are from [Wei et al. \(2022\)](#), the results for LaMDA and UL2 are from [Wang et al. \(2022b\)](#), and the results for Minerva are from [Lewkowycz et al. \(2022\)](#). We ran PAL on each benchmark 3 times and report the average; the standard deviation is provided in Table 7.

# Plan-and-Solve Prompting



# LEARNING MATH REASONING FROM SELF-SAMPLED CORRECT AND PARTIALLY-CORRECT SOLUTION

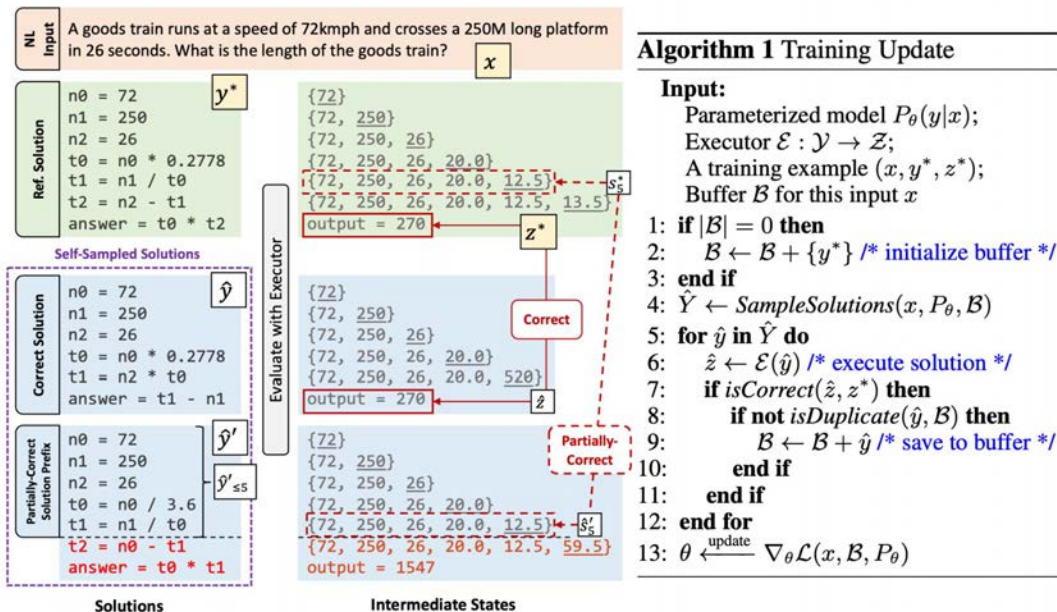
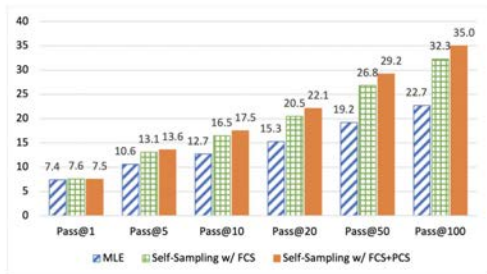
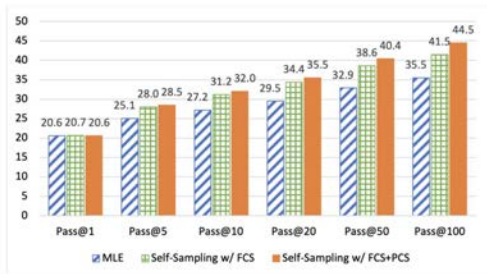


Figure 1: Examples of self-sampled correct and partially-correct solutions from MathQA (more in Appendix D). The steps and intermediate states marked in red are incorrect.

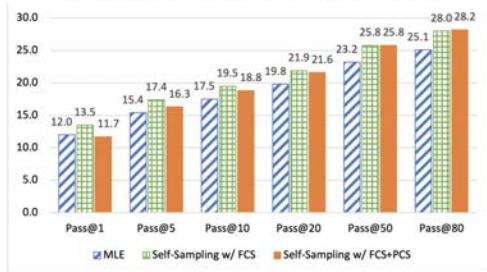
# LEARNING MATH REASONING FROM SELF-SAMPLED CORRECT AND PARTIALLY-CORRECT SOLUTION



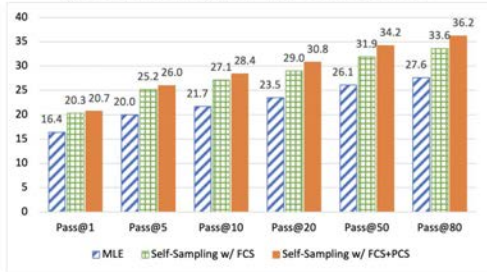
(a) GSM5.5K-Python with GPT-Neo 125M



(b) GSM5.5K-Python with GPT-Neo 2.7B



(c) MathQA-Python-Filtered with GPT-Neo 125M



(d) MathQA-Python-Filtered with GPT-Neo 2.7B

Figure 2: Percentage of the problems solved (PASS@ $k$ ) on the dev set of GSM5.5K-Python and MathQA-Python-Filtered, comparing our self-sampling approach and the common MLE objective. All our methods include partially-correct solutions and use the MLE-Aug loss for learning.

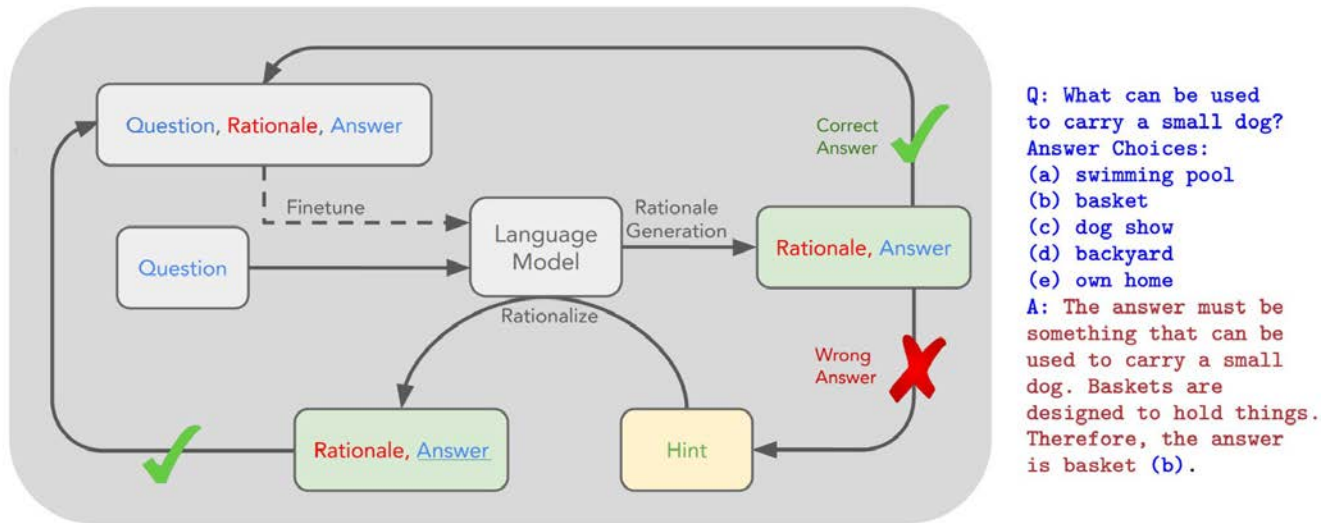


Figure 1: An overview of STaR and a STaR-generated rationale on CommonsenseQA. We indicate the fine-tuning outer loop with a dashed line. The questions and ground truth answers are expected to be present in the dataset, while the rationales are generated using STaR.

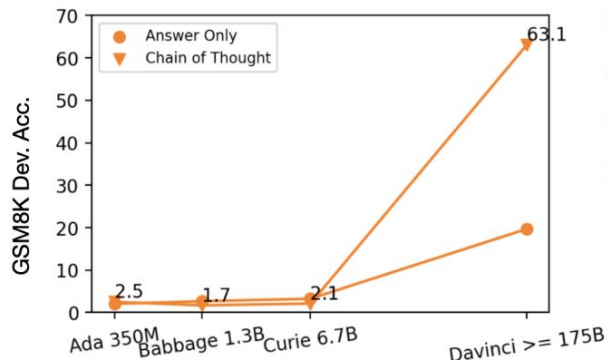


Table 2: We find that STaR substantially improves GSM8K performance over the baselines, despite training on only 25.0% of the data for the model without rationalization, and 28.7% of the dataset (with 0.5% from rationalization) for the model with rationalization.

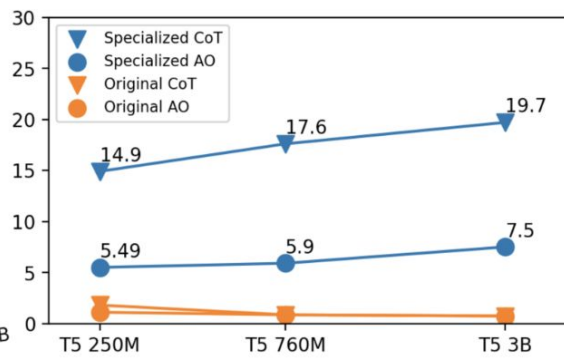
	GSM8K Test Accuracy (%)	Train Data Used (%)
Few-shot Direct GPT-J	3.0	~0
Few-shot CoT GPT-J	3.1	~0
GPT-J Direct Finetuned	5.8	100
STaR without rationalization	10.1	25.0
STaR with rationalization	<b>10.7</b>	28.7

# Specializing Smaller Language Models towards Multi-Step Reasoning

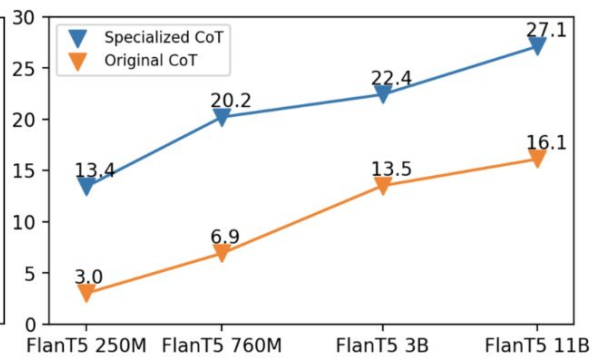
## Reasoning



A. GPT phase change curve, almost flat in small scale

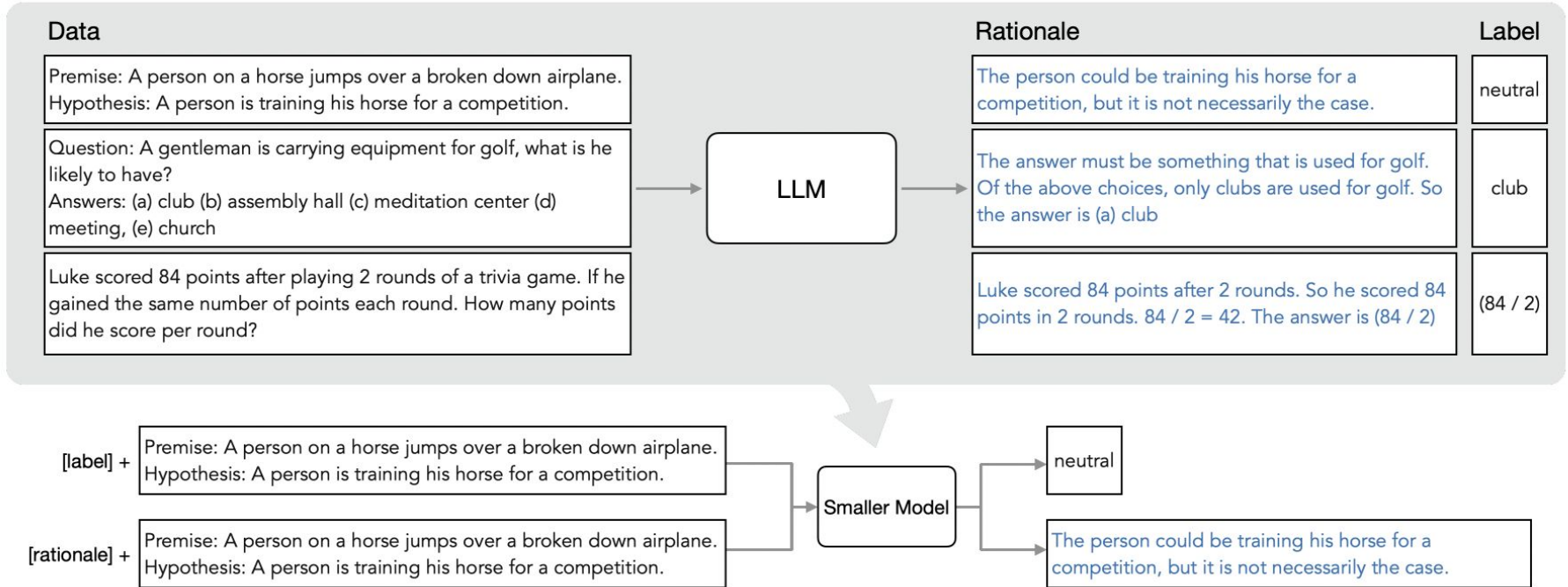


B. Specialized T5 exhibit log-linear scaling curve



C. Specialization lifts up FlanT5 log-linear curve

# Distilling Step-by-Step



# Distilling Step-by-Step

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \lambda \mathcal{L}_{\text{rationale}}, \quad (3)$$

where  $\mathcal{L}_{\text{label}}$  is the label prediction loss in Eq. 1 and  $\mathcal{L}_{\text{rationale}}$  is the *rationale generation loss*:

$$\mathcal{L}_{\text{rationale}} = \frac{1}{N} \sum_{i=1}^N \ell(f(x_i), \hat{r}_i). \quad (4)$$

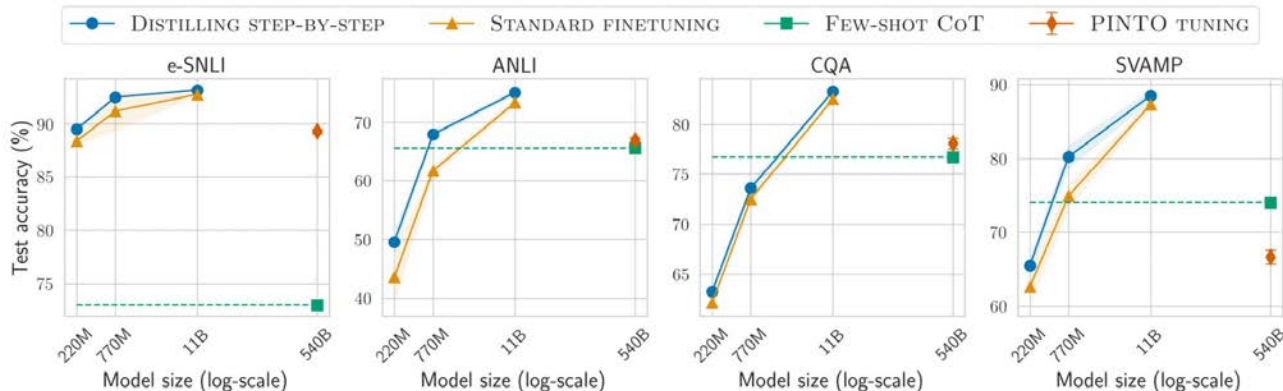
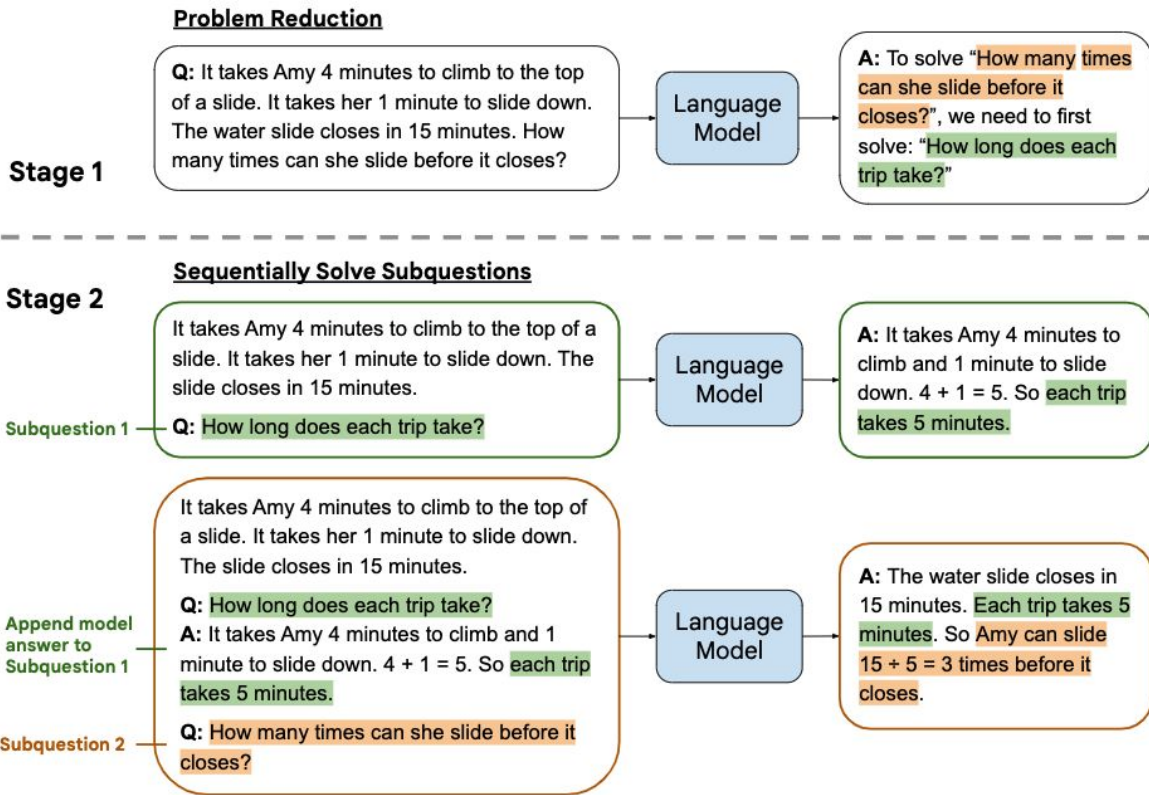


Figure 6: We perform Distilling step-by-step and Standard finetuning, using the full human-labeled datasets, on varying sizes of T5 models and compare their performance to LLM baselines, i.e., Few-shot CoT and PINTO Tuning. Distilling step-by-step is able to outperform LLM baselines by using much smaller models, e.g., over 700× smaller model on ANLI. Standard finetuning fails to match LLM’s performance using the same model size.

## Recursive and Iterative prompting

# LEAST-TO-MOST PROMPTING



# LEAST-TO-MOST PROMPTING - Prompts

## 9.2.4 LEAST-TO-MOST PROMPTING I: PROBLEM DECOMPOSITION (5 EXAMPLES)

Q: The median age in the city was 22.1 years. 10.1% of residents were under the age of 18; 56.2% were between the ages of 18 and 24; 16.1% were from 25 to 44; 10.5% were from 45 to 64; and 7% were 65 years of age or older. Which age group is larger: under the age of 18 or 18 and 24?

A: To answer the question “Which age group is larger: under the age of 18 or 18 and 24?”, we need to know: “How many percent were under the age of 18?”, “How many percent were between the ages of 18 and 24?”.

Q: Old age pensions were raised by 300 francs per month to 1,700 francs for a single person and to 3,700 francs for a couple, while health insurance benefits were made more widely available to unemployed persons and part-time employees. How many francs were the old age pensions for a single person before they were raised?

A: To answer the question “How many francs were the old age pensions for a single person before they were raised?”, we need to know: “How many francs were the old age pensions for a single person?”, “How many francs were old age pensions raised for a single person?”.

Q: In April 2011, the ECB raised interest rates for the first time since 2008 from 1% to 1.25%, with a further increase to 1.50% in July 2011. However, in 2012-2013 the ECB lowered interest rates to encourage economic growth, reaching the historically low 0.25% in November 2013. Soon after the rates were cut to 0.15%, then on 4 September 2014 the central bank reduced the rates from 0.15% to 0.05%, the lowest rates on record. How many percentage points did interest rates drop between April 2011 and September 2014?

A: To answer the question “How many percentage points did interest rates drop between April 2011 and September 2014?”, we need to know: “What was the interest rate in April 2011?”, “What was the interest rate in September 2014?”.

Q: Non-nationals make up more than half of the population of Bahrain. According to government statistics dated between 2005-2009 roughly 290,000 Indians, 125,000 Bangladeshis, 45,000 Pakistanis, 45,000 Filipinos, and 8,000 Indonesians. How many Pakistanis and Indonesians are in Bahrain?

A: To answer the question “How many Pakistanis and Indonesians are in Bahrain?”, we need to know: “How many Pakistanis are in Bahrain?”, “How many Indonesians are in Bahrain?”.

Q: The Census Bureaus 2006-2010 American Community Survey showed that (in 2010 inflation adjustment dollars) median household income was \$52,056 and the median family income was \$58,942. How many years did the Census Bureaus American Community Survey last?

A: To answer the question “How many years did the Census Bureaus American Community Survey last?”, we need to know: “When did the Census Bureaus American Community Survey start?”, “When did the Census Bureaus American Community Survey end?”.

## 9.2.5 LEAST-TO-MOST PROMPTING II: PROBLEM SOLVING (3 EXAMPLES)

The city of Bangkok has a population of 8,280,925. The census showed that it is home to 81,570 Japanese and 55,893 Chinese nationals, as well as 117,071 expatriates from other Asian countries, 48,341 from Europe, 23,418 from the Americas.

Q: How many Chinese nationals are there?

A: “The census showed that it is home to 81,570 Japanese and 55,893 Chinese nationals”. So the answer is 55,893 Chinese nationals.

Q: How many European nationals are there?

A: “48,341 from Europe”. So the answer is 48,341.

Q: How many more Chinese nationals are there than European nationals?

A: We know that there are 55,893 Chinese nationals. We also know that there are 48,341 European nationals. So there are  $55,893 - 48,341 = 7,552$  more Chinese nationals than European nationals. So the answer is 7552.

The gender distribution of the population was 50.2% male and 49.8% female. Of the adult population. 29 people or 14.6% of the population are between 20 and 29 years old. 28 people or 14.1% are 30 to 39, 36 people or 18.2% are 40 to 49, and 31 people or 15.7% are 50 to 59.

Q: What is the percentage of people between 40 to 49?

A: “36 people or 18.2% are 40 to 49”. So the answer is 18.2.

Q: How many percent of people were not 40 to 49?

A: We know that 18.2% are 40 to 49. So  $100\% - 18.2\% = 81.8\%$  are not 40 to 49. So the answer is 81.8%.

During the 1970s the oil shocks (1973 oil crisis and 1979 energy crisis) hit Iceland hard. Inflation rose to 43% in 1974 and 59% in 1980, falling to 15% in 1987 but rising to 30% in 1988. Iceland experienced moderately strong GDP growth (3% on average) from 1995 to 2004. Growth slowed between 2000 and 2002, but the economy expanded by 4.3% in 2003 and grew by 6.2% in 2004. Growth in 2005 exceeded 6%.

Q: When did the oil crisis occur?

A: “1973 oil crisis and 1979 energy crisis”. So the answer is 1973.

Q: When did the energy crisis occur?

A: “1979 energy crisis”. So the answer is 1979.

Q: How many years passed between the oil and energy crises?

A: We know that the oil crisis occurred in 1973. We also know that the energy crisis occurred in 1979. So  $1979 - 1973 = 6$  years passed between the oil and energy crises. So the answer is 6.

*Least-to-Most Prompting Enables Complex Reasoning in Large Language Models (Zhou et al., 2022).*

## LEAST-TO-MOST PROMPTING - Results

Method	Non-football (DROP)	Football (DROP)	GSM8K
Zero-Shot	43.86	51.77	16.38
Standard prompting	58.78	62.73	17.06
Chain-of-Thought	74.77	59.56	60.87
Least-to-Most	<b>82.45</b>	<b>73.42</b>	<b>62.39</b>

Table 11: Accuracies (%) of different prompting methods on GSM8K and DROP (only the subset containing numerical problems). The base language model is `code-davinci-002`.



# Plan, Eliminate, and Track

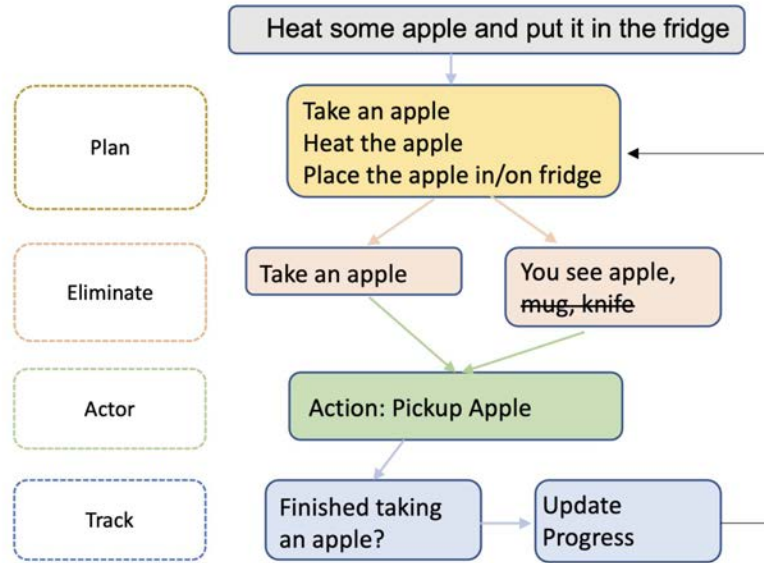
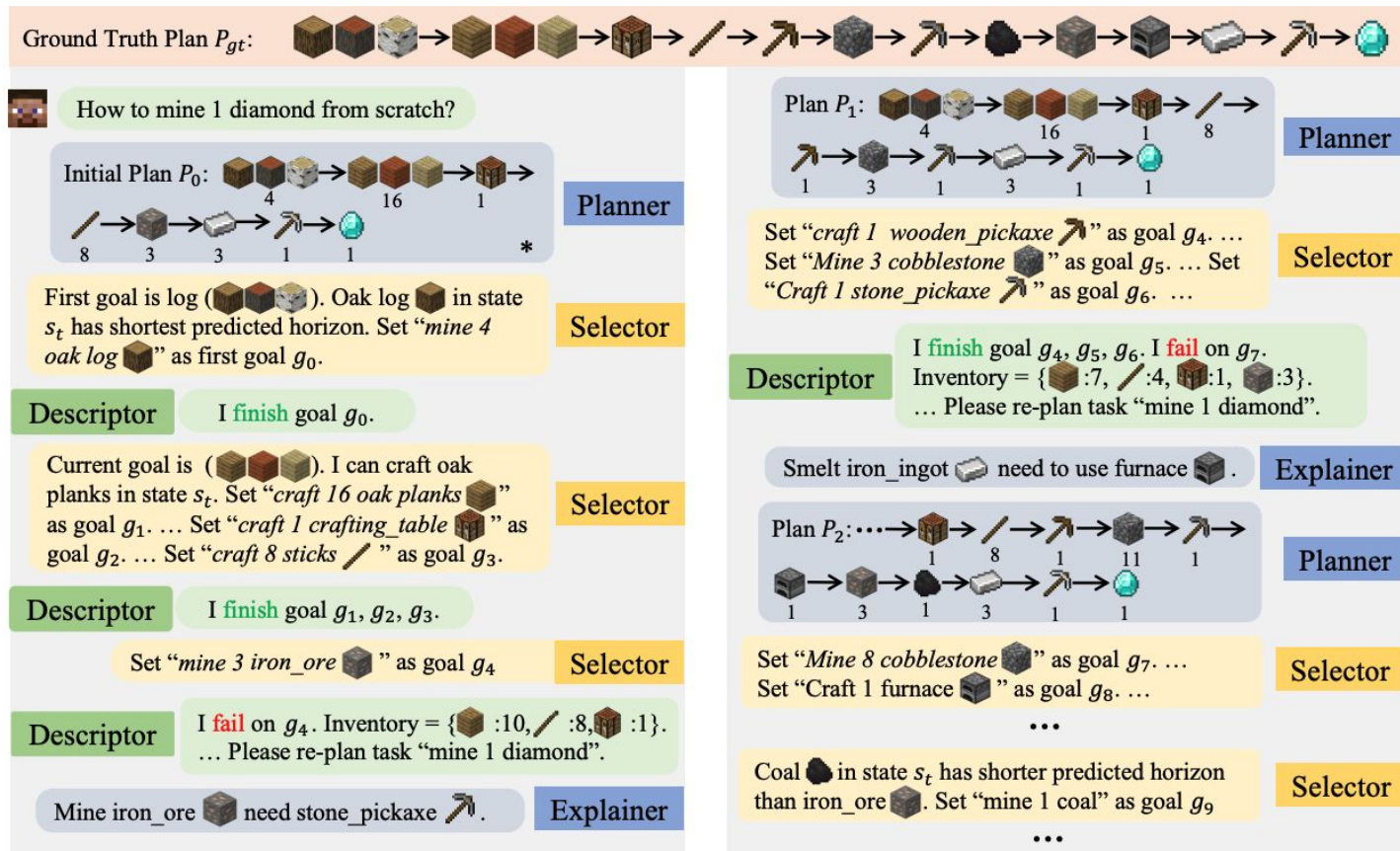


Figure 1. PET framework. Plan module uses LLM to generate a high-level plan. Eliminate Module uses a QA model to mask irrelevant objects in observation. Track module uses a QA model to track the completion of sub-tasks.

# Describe , Explain , Plan and Select



## Tool Usage

# React: Reason and Act

(1) Hotspot QA

**Question:** Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

**Answer:** iPod ✗

(1b) CoT (Reason Only)

**Thought:** Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

**Answer:** iPhone, iPad, iPod Touch ✗

(1c) Act-Only

**Act 1:** Search[Apple Remote]  
**Obs 1:** The Apple Remote is a remote control ...

**Act 2:** Search[Front Row]  
**Obs 2:** Could not find [Front Row]. Similar: ...

**Act 3:** Search[Front Row (software)]  
**Obs 3:** Front Row is a discontinued media center software ...

**Act 4:** Finish[yes] ✗

(1d) ReAct (Reason + Act)

**Thought 1:** I need to search Apple Remote and find the program it was originally designed to interact with.  
**Act 1:** Search[Apple Remote]  
**Obs 1:** The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

**Thought 2:** Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.  
**Act 2:** Search[Front Row]  
**Obs 2:** Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

**Thought 3:** Front Row is not found. I need to search Front Row (software).  
**Act 3:** Search[Front Row (software)]  
**Obs 3:** Front Row is a discontinued media center software ...

**Thought 4:** Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.  
**Act 4:** Finish[keyboard function keys] ✓

---

(2) AlfWorld

You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.  
**Your task is to:** Put some pepper shaker on a drawer.

(2a) Act-Only

**Act 1:** Go to drawer 1  
**Obs 1:** The drawer 1 is closed.

**Act 2:** Open drawer 1  
**Obs 2:** You open the drawer 1. The drawer 1 is open. In it, you see a dish sponge 2, and a spoon 1.

**Act 3:** Go to sinkbasin 1  
**Obs 3:** On the sinkbasin 1, you see a dish sponge 3, a spatula 1, and a spoon 2.

**Act 4:** Take peppershaker 1 from sinkbasin 1  
**Obs 4:** Nothing happens.

**Act 5:** Take peppershaker 1 from sinkbasin 1  
**Obs 5:** Nothing happens. ✗

(2b) ReAct (Reason + Act)

**Act 1:** Think[First I need to find a pepper shaker...more likely to appear in cabinets (1-6), countertops (1-3), ... ]  
**Act 2:** Go to cabinet 1  
**Obs 2:** On the cabinet 1, you see a vase 2.  
(Here the agent go to cabinet 1, then cabinet 2, 3, then countertop 1 and 2)  
**Act 6:** Go to countertop 3  
**Obs 6:** On the countertop 3, you see a apple 1, a bread 1, ..... a pepper shaker 1, and a vase 3.

**Act 7:** Take pepper shaker 1 from countertop 3  
**Obs 7:** You pick up the pepper shaker 1 from the countertop 3.

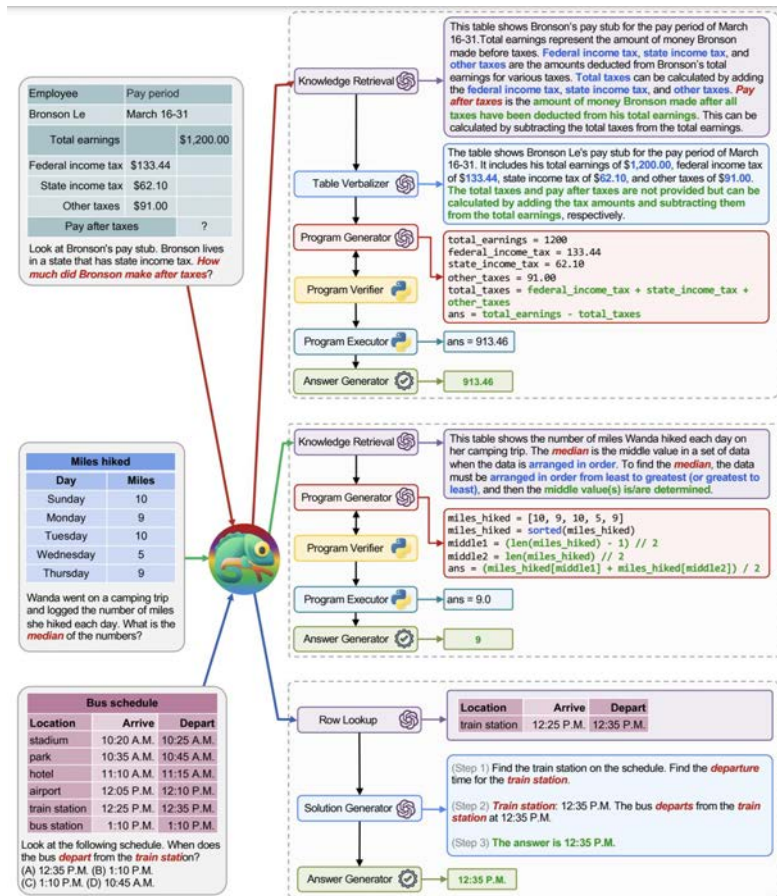
**Act 8:** Think[Now I find a pepper shaker 1. Next, I need to put it in/on drawer 1. ]  
**Act 9:** Go to drawer 1  
**Obs 9:** Drawer 1 is closed.

**Act 10:** Open drawer 1  
**Obs 10:** You open Drawer 1 ...

**Act 11:** Put pepper shaker 1 in/on drawer 1  
**Obs 11:** You put pepper shaker 1 in/on the drawer 1. ✓

REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS (Yao et al., 2023).

# Chameleon



# Chameleon

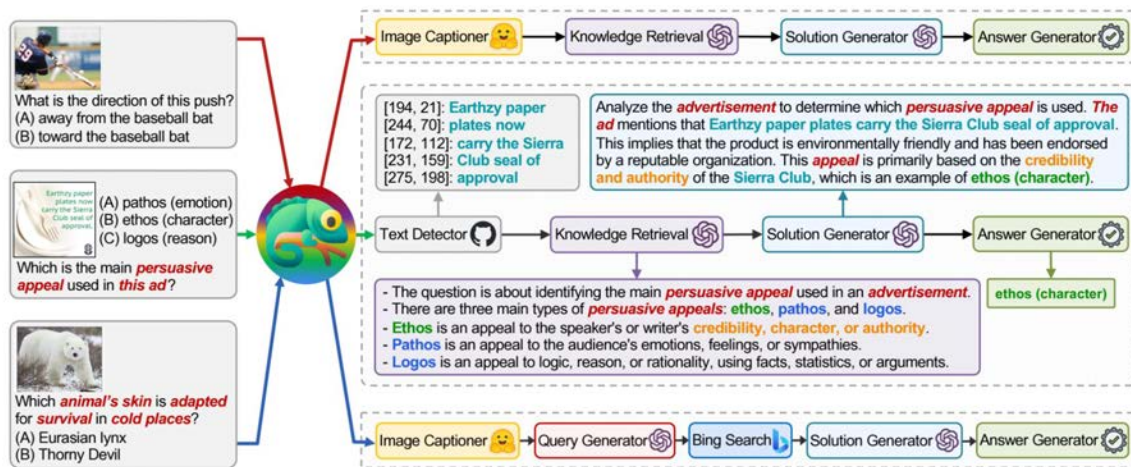


Figure 1: Examples from our **Chameleon** with GPT-4 on ScienceQA [27], a multi-modal question answering benchmark in scientific domains. **Chameleon** is adaptive to different queries by synthesizing programs to compose various tools and executing them sequentially to get final answers.

# Chameleon

▷ *Instruction for the planner model*

You need to act as a policy model, that given a question and a modular set, determines the sequence of modules that can be executed sequentially can solve the question.

The modules are defined as follows:

**Query\_Generator:** This module generates a search engine query for the given question. Normally, we consider using "Query\_Generator" when the question involves domain-specific knowledge.

**Bing\_Search:** This module searches the web for relevant information to the question. Normally, we consider using "Bing\_Search" when the question involves domain-specific knowledge.

**Image\_Captioner:** This module generates a caption for the given image. Normally, we consider using "Image\_Captioner" when the question involves the semantic understanding of the image, and the "has\_image" field in the metadata is True.

**Text\_Detector:** This module detects the text in the given image. Normally, we consider using "Text\_Detector" when the question involves the unfolding of the text in the image, e.g., diagram, chart, table, map, etc., and the "has\_image" field in the metadata is True.

**Knowledge\_Retrieval:** This module retrieves background knowledge as the hint for the given question. Normally, we consider using "Knowledge\_Retrieval" when the background knowledge is helpful to guide the solution.

**Solution\_Generator:** This module generates a detailed solution to the question based on the information provided. Normally, "Solution\_Generator" will incorporate the information from "Query\_Generator", "Bing\_Search", "Image\_Captioner", "Text\_Detector", and "Knowledge\_Retrieval".

**Answer\_Generator:** This module extracts the final answer in a short form from the solution or execution result. This module normally is the last module in the prediction pipeline.

Below are some examples that map the problem to the modules.

▷ *In-context example(s)*

**Question:** Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?

**Context:** The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles.

**Options:** (A) neither; the samples have the same temperature (B) sample A (C) sample B

**Metadata:** 'pid': 19, 'has\_image': True, 'grade': 8, 'subject': 'natural science', 'topic': 'physics', 'category': 'Particle motion and energy', 'skill': 'Identify how particle motion affects temperature and pressure'

**Modules:** ["Text\_Detector", "Knowledge\_Retrieval", "Solution\_Generator", "Answer\_Generator"]

## Acknowledgement & Further reading

- Augmented Language Models: a Survey ( Mialon et al, 2023)
- Towards Reasoning in Large Language Models: A Survey (Huang et al, 2022)
- [LLM Chronicles #2: How To Leverage Emergent Abilities Of LLMs](#)
- [Do Large Language Models \(LLMs\) reason?](#)
- Large Language Models Still Can't Plan ( Valmeekam et al, 2023)
- Papers cited in each slides