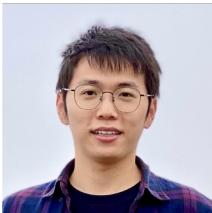




Microsoft

Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models

Pan Lu
2023.05.05



Pan Lu



Baolin Peng



Hao Cheng



Michel Galley



Kai-Wei Chang



Ying Nian Wu



Song-Chun Zhu



Jianfeng Gao

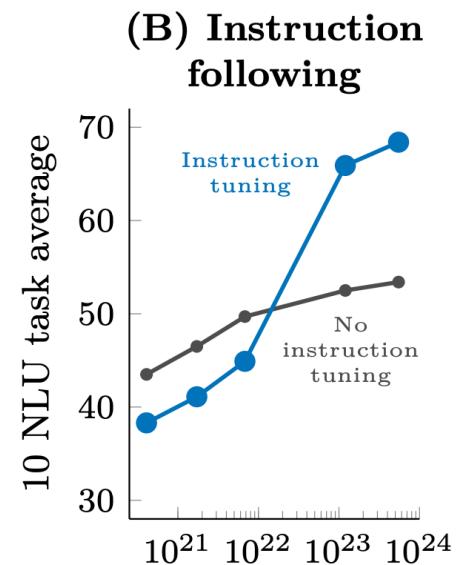
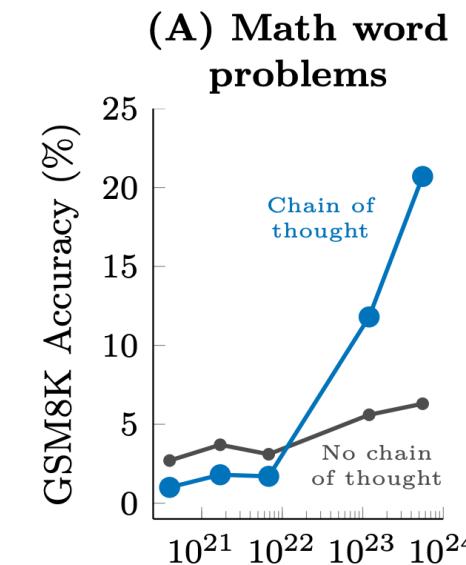
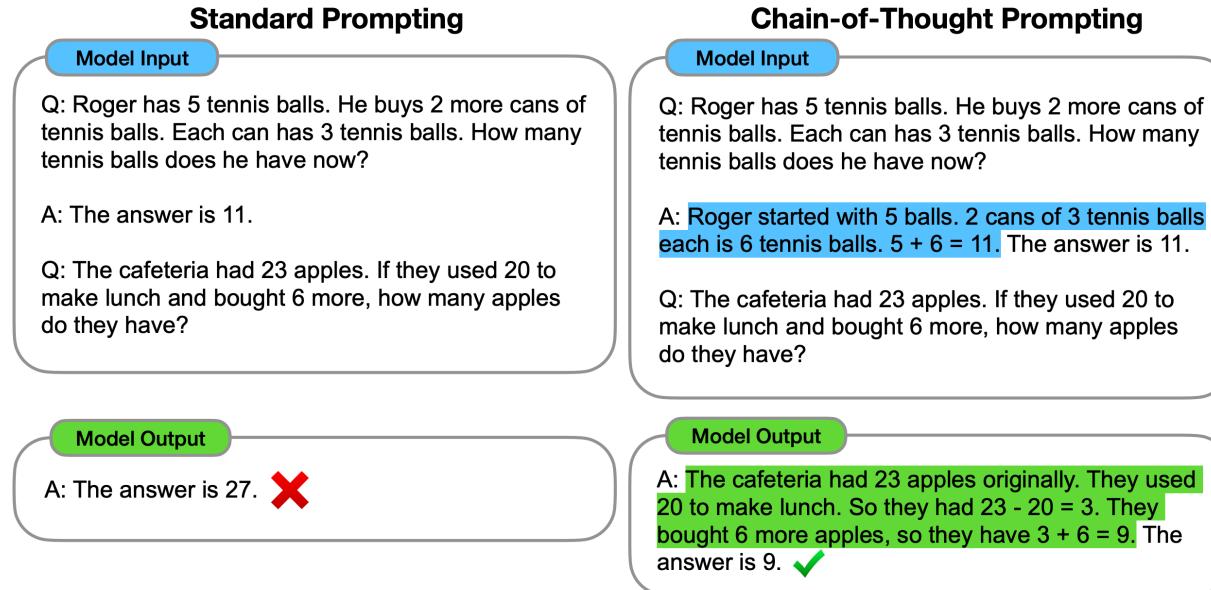
About Me



- 4th-year Ph.D. Candidate at UCLA
- Research interests
 - ❖ **Large language models** for planning, reasoning, and generation
 - ❖ **Mathematical reasoning** in mathematics, sciences, and medical domains
 - ❖ **Trustworthy NLP models**: explainable, reliable, and socially responsible
 - ❖ **Conversational agents**: value-aware and socially intelligent
 - ❖ **Multimodal reasoning** for vision-and-language applications
- Lead organizer for NeurIPS-21 MathAI4ED and NeurIPS-22 MathAI Workshops
- Keynote presenter at IJCAI-23 Tutorial on mathematical reasoning

<https://lupantech.github.io/>

Large Language Models and Emergent Abilities



Jason et al, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022

Jason et al, Transactions on Machine Learning Research, 2022

Chain-of-Thought Prompting

Math Word Problems (free response)

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 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm³, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Chain-of-Thought Prompting (CoT) unlocks and boosts the compositional reasoning ability in the natural language space.

Limitations of LLMs (CoT)

- Lack the abilities for **visual perception**

Question: Which is the main persuasive appeal used in this ad?

Context: A black dog standing in front of a fence.

Options:

- (A) logos (reason)
- (B) ethos (character)
- (C) pathos (emotion)



Answer: The answer is C. BECAUSE: The ad is trying to make you feel a certain way about the product. It is not trying to persuade you with logic or with an appeal to character.

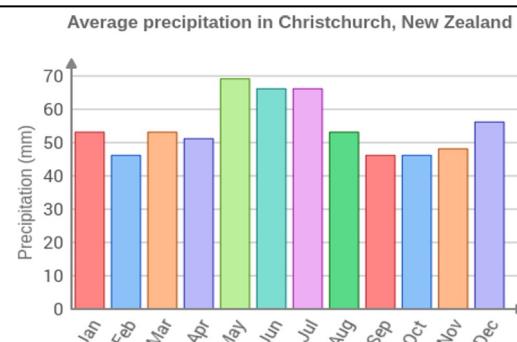
Question: Which month is the wettest on average in Christchurch?

Context: Use the graph to answer the question below.

A series of photographs showing a variety of colorful objects.

Options:

- (A) April
- (B) December
- (C) May



Answer: The answer is B. BECAUSE: The graph shows the average monthly rainfall in Christchurch from 1971 to 2000.

The wettest month on average is December, with an average rainfall of 154.8 mm.

The second wettest month is May, with an average rainfall of 153.1 mm.

The third wettest month is April, with an average rainfall of 151.4 mm.



Hugging Face



Limitations of LLMs (CoT)

- An inability to access **domain-specific knowledge** and **up-to-date** information

Question: Complete the sentence.

A dinosaur bone turning into rock over millions of years is a ().

Options: (A) chemical change (B) physical change

Answer: The answer is B. BECAUSE: A **physical change** is a change to a substance that does not change what the substance is made of.

A **chemical change** is a change to a substance that does change what the substance is made of.\n

In this sentence, the word 'change' is a verb. It is an action word. The word 'change' is also a noun. It is a thing.

A **physical change** is a change to a substance that does not change what the substance is made of. For example, when water freezes, it changes from a liquid to a solid. But it is still made of water molecules. So, freezing is a **physical change**.

A **chemical change** is a change to a substance that does change what the substance is made of. For example, when water evaporates, it changes from a liquid to a gas. But it is still made of water molecules. So, evaporating is a **physical change**.

In this sentence, **the word 'change' is a physical change**. A dinosaur bone turning into rock over millions of years is a physical change.



Google

Bing



Limitations of LLMs

- An inability to understand **domain-specific context information**

Table:

[TITLE] Apples per tree

Stem | Leaf

1 | 1, 3, 6

2 | 2, 3, 3, 6

3 | 0

4 | 3

5 | 2, 6

6 | 0, 0, 2

7 | 2, 8

8 | 4, 5, 5, 6, 7

Apples per tree	
Stem	Leaf
1	1, 3, 6
2	2, 3, 3, 6
3	0
4	3
5	2, 6
6	0, 0, 2
7	2, 8
8	4, 5, 5, 6, 7

Question: A farmer counted the number of apples on each tree in her orchard. How many trees have at least 51 apples but fewer than 83 apples? (unit: trees)

Answer:

(Step 1) To solve this problem, look at the table and find all the trees that have at least 51 apples but fewer than 83 apples.

(Step 2) There are three trees that have at least 51 apples but fewer than 83 apples. **The answer is 3.**

Output: 3

Ground truth: 7

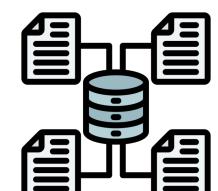


Table:

yellow tiger's eye | \$0.85

piece of green slate | \$0.59

piece of red sandstone | \$0.19

piece of rose quartz | \$0.61

smooth piece of marble | \$0.45

yellow tiger's eye	\$0.85
piece of green slate	\$0.59
piece of red sandstone	\$0.19
piece of rose quartz	\$0.61
smooth piece of marble	\$0.45

Question: How much money does Connor need to buy a piece of rose quartz, a piece of green slate, and a piece of red sandstone? (unit: \$)

Answer:

(Step 1) Connor needs to buy a piece of rose quartz, a piece of green slate, and a piece of red sandstone. To find the total amount of money Connor needs, add the prices of the three items.

(Step 2) $\$0.85 + \$0.59 + \$0.19 = \1.63

(Step 3) Connor needs **$\$1.63$** to buy a piece of rose quartz, a piece of green slate, and a piece of red sandstone.

The answer is 1.63.

Output: 1.63

Ground truth: 1.39

Limitations of LLMs

- Lack the ability to perform **precise mathematical reasoning**

Table:

[TITLE] Blocks from school

Name | Number of blocks

Wendy | 14

Carson | 8

Zach | 17

Lila | 17

Nicole | 3

Shane | 7

Henry | 5

Devin | 17

Blocks from school

Name | **Number of blocks**

Wendy | 14

Carson | 8

Zach | 17

Lila | 17

Nicole | 3

Shane | 7

Henry | 5

Devin | 17

Question: Some students compared how many blocks they live from school. What is the mean of the numbers?

Answer:

(Step 1) The mean is the average of the numbers.

(Step 2) To find the mean, add the numbers and divide by the number of students.

(Step 3) $14 + 8 + 17 + 17 + 3 + 7 + 5 + 17 = 86$

(Step 4) There are 8 students, so divide 86 by 8.

(Step 5) $86 \div 8 = 10.75$

(Step 6) The mean is 10.75. The answer is 10.75.

Output: 10.75

Ground truth: 11

Table:

[TITLE] Lawn mowers sold

Day | Number of lawn mowers

Tuesday | 8

Wednesday | 7

Thursday | 5

Friday | 7

Saturday | 9

Sunday | 8

Monday | 8

Lawn mowers sold

Day | **Number of lawn mowers**

Tuesday | 8

Wednesday | 7

Thursday | 5

Friday | 7

Saturday | 9

Sunday | 8

Monday | 8

Question: A garden supply store kept track of how many lawn mowers it sold in the past 7 days. What is the median of the numbers?

Answer:

(Step 1) The median is the "middle" value. To find the median, put the numbers in order from least to greatest and find the number in the middle.

(Step 2) 8, 7, 5, 7, 9, 8, 8

(Step 3) The median is 7. The answer is 7.

Output: 7

Ground truth: 8



Limitations of LLMs

- An inability to combine **different tools** to perform complex reasoning

Table:

[TITLE] Tour boat schedule

Ocean City | 8:15 A.M. | 9:00 A.M. | 9:15 A.M. | 9:30 A.M. | 10:00 A.M.

Whale Watch Harbor | 9:30 A.M. | 10:15 A.M. | 10:30 A.M. | 10:45 A.M. | 11:15 A.M.

Oyster Lighthouse | 10:15 A.M. | 11:00 A.M. | 11:15 A.M. | 11:30 A.M. | 12:00 P.M.

Fisherman's Cove | 11:15 A.M. | 12:00 P.M. | 12:15 P.M. | 12:30 P.M. | 1:00 P.M.

Surfing Beach | 12:00 P.M. | 12:45 P.M. | 1:00 P.M. | 1:15 P.M. | 1:45 P.M.

Tour boat schedule					
Ocean City	8:15 A.M.	9:00 A.M.	9:15 A.M.	9:30 A.M.	10:00 A.M.
Whale Watch Harbor	9:30 A.M.	10:15 A.M.	10:30 A.M.	10:45 A.M.	11:15 A.M.
Oyster Lighthouse	10:15 A.M.	11:00 A.M.	11:15 A.M.	11:30 A.M.	12:00 P.M.
Fisherman's Cove	11:15 A.M.	12:00 P.M.	12:15 P.M.	12:30 P.M.	1:00 P.M.
Surfing Beach	12:00 P.M.	12:45 P.M.	1:00 P.M.	1:15 P.M.	1:45 P.M.

Question: Look at the following schedule. Haley is at Ocean City at 9:45 A.M. How soon can she get to Surfing Beach?

Options: (A) 1:45 P.M. (B) 12:00 P.M. (C) 10:30 A.M. (D) 12:15 P.M.

Answer:

(Step 1) Find 9:45 A.M. on the schedule.

(Step 2) The next stop is Whale Watch Harbor at 10:15 A.M.

(Step 3) The earliest she can get to Surfing Beach is 12:00 P.M. The answer is 12:00 P.M.

Output: (B) 12:00 P.M.

Ground truth: (A) 1:45 P.M.



Tool Use



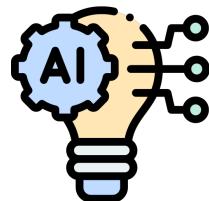
Hugging Face



Google



How to **compose** these numerous tools to tackle complex tasks?



Tool-Augmented LLMs!

Existing Tool-Augmented LLMs

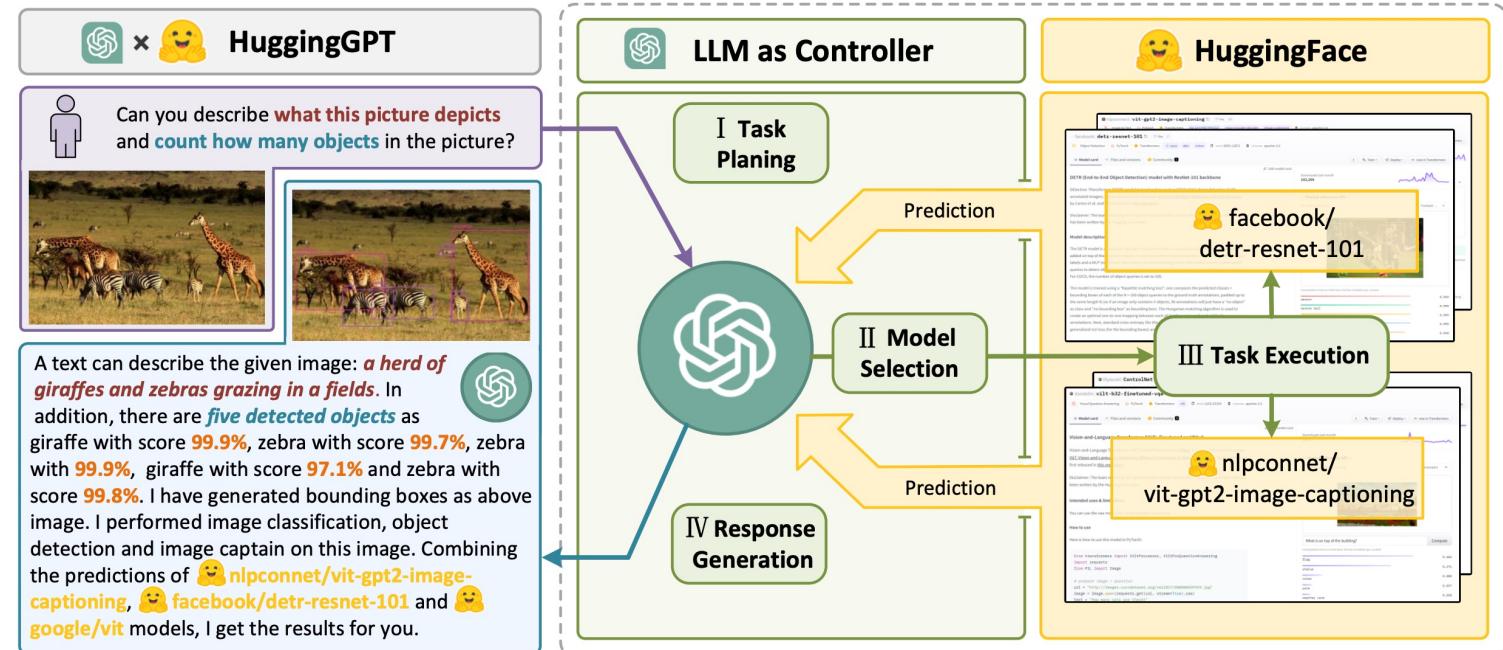
The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Toolformer: Language Models Can Teach Themselves to Use Tools, <https://arxiv.org/abs/2302.04761>



HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, <https://arxiv.org/abs/2303.17580>



Constrained to a small set of tools or task-specific tools

Existing Tool-Augmented LLMs

Query: How many muffins can each kid have for it to be fair?



Generated Code

```
def execute_command(image):
    image_patch = ImagePatch(image)
    muffin_patches = image_patch.find("muffin")
    kid_patches = image_patch.find("kid")
    return str(len(muffin_patches) // len(kid_patches))
```

Query: Drink with zero alcohol



```
def execute_command(image):
    image_patch = ImagePatch(image)
    drink_patches = image_patch.find("drink")
    for drink_patch in drink_patches:
        drink_name = drink_patch.simple_query("What is this?")
        alcoholic = llm_query(f"Does the {drink_name} have alcohol?")
        if alcoholic == "no":
            return drink_patch
    return None
```

Execution

```
muffin_patches =
image_patch.find("muffin")
```



```
kid_patches =
image_patch.find("kid")
```



► len(muffin_patches)=8
► len(kid_patches)=2

► 8//2 = 4

Result: 4

► drink_patches=



► drink_name = 'tullamore dew'
► alcoholic = 'yes'

► drink_name = 'bacardi'
► alcoholic = 'yes'

► drink_name = 'gin'
► alcoholic = 'yes'

► drink_name = 'dr pepper'
► alcoholic = 'no'

Result:



ViperGPT: Visual Inference via Python Execution, <https://arxiv.org/abs/2303.08128>

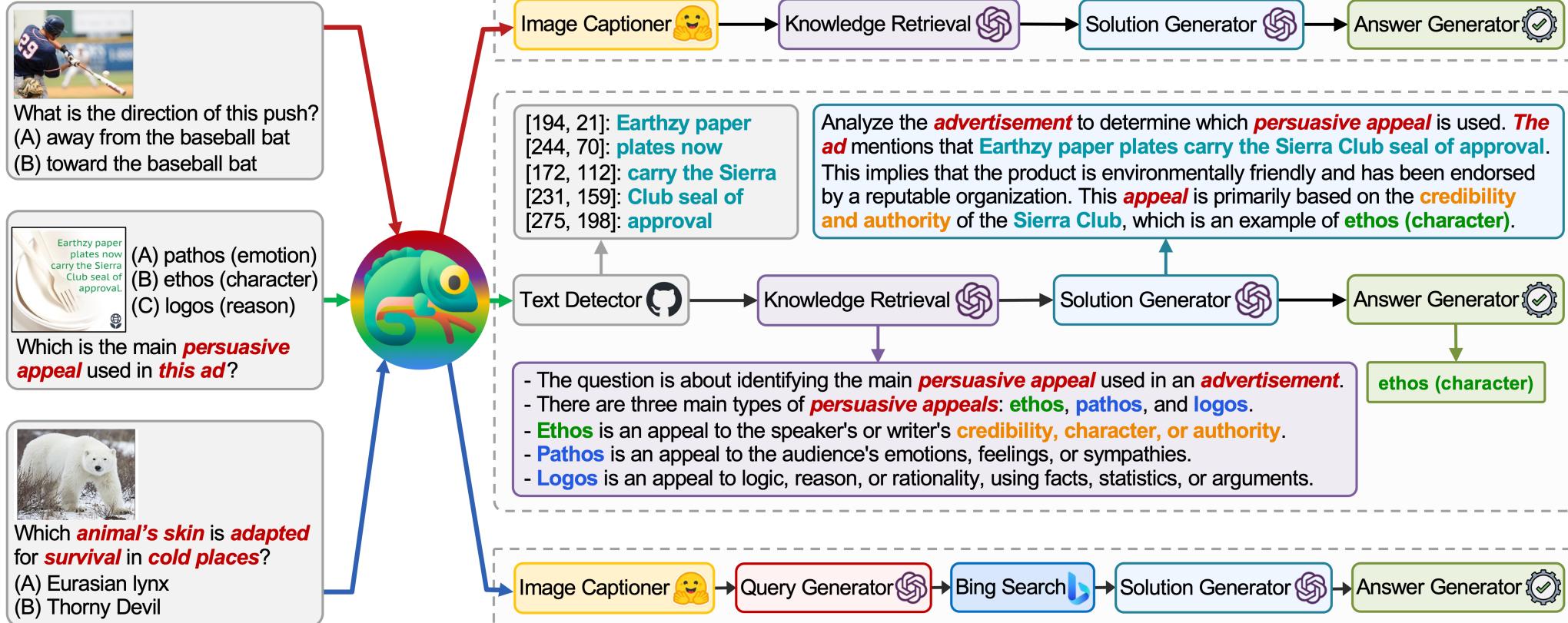


Constrained to domain-specific commands or programs

Comparisons of Existing Work

Model	Tool Use						Skill Dimension					Inference & Extension		
	Size	🕸	👩	⌚	🧞	🐍	Image	Web	Know.	Math	Table	Composition	Planning	Plug-n-Play
CoT [49]	1	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Lila [33]	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
PoT [6]	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
Code4Struct [47]	1	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
PAL [9]	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
MathPrompter [15]	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
ART [37]	4	✓	✗	✗	✓	✓	✗	✓	✗	✓	✗	✓	✗	✓
Toolformer [43]	5	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	natural lang.	✗
WebGPT [34]	10	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	program	✗
MM-ReAct [52]	>10	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	word match	✓
Visual ChatGPT [51]	>10	✓	-	-	✗	✗	✓	✗	✗	✗	✗	✓	natural lang.	✓
ViperGPT [45]	>10	✓	-	-	✗	✗	✓	✗	✓	✓	✗	✓	program	✓
VisProg [11]	>10	✓	-	-	✗	✓	✓	✗	✗	✗	✗	✓	program	✓
HuggingGPT [44]	>10	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	natural lang.	✓
Chameleon (ours)	>10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	natural lang.	✓

Chameleon: Examples



Chameleon: Examples

Employee	Pay period
Bronson Le	March 16-31
Total earnings	\$1,200.00
Federal income tax	\$133.44
State income tax	\$62.10
Other taxes	\$91.00
Pay after taxes	?

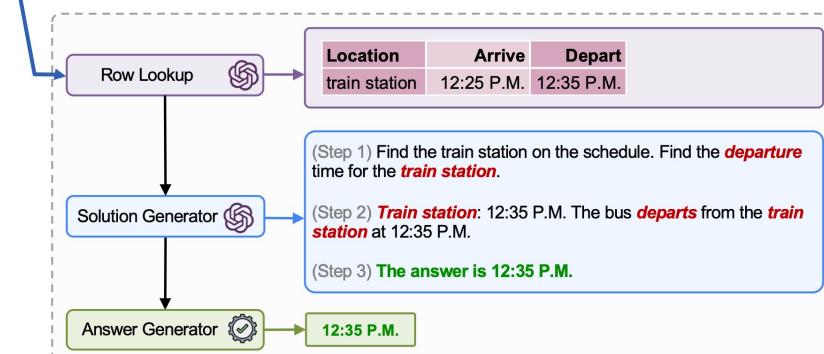
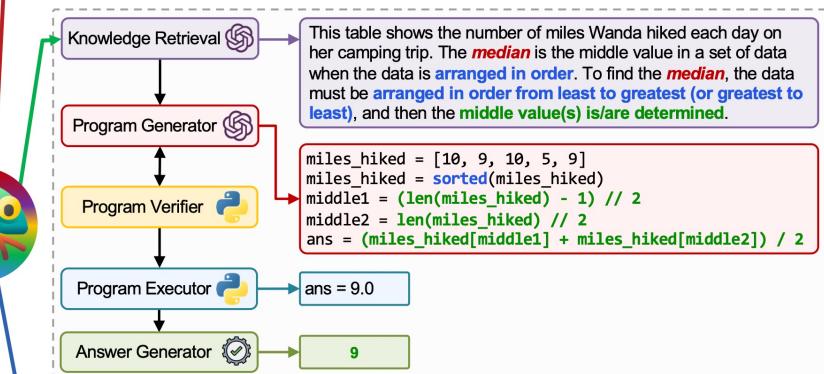
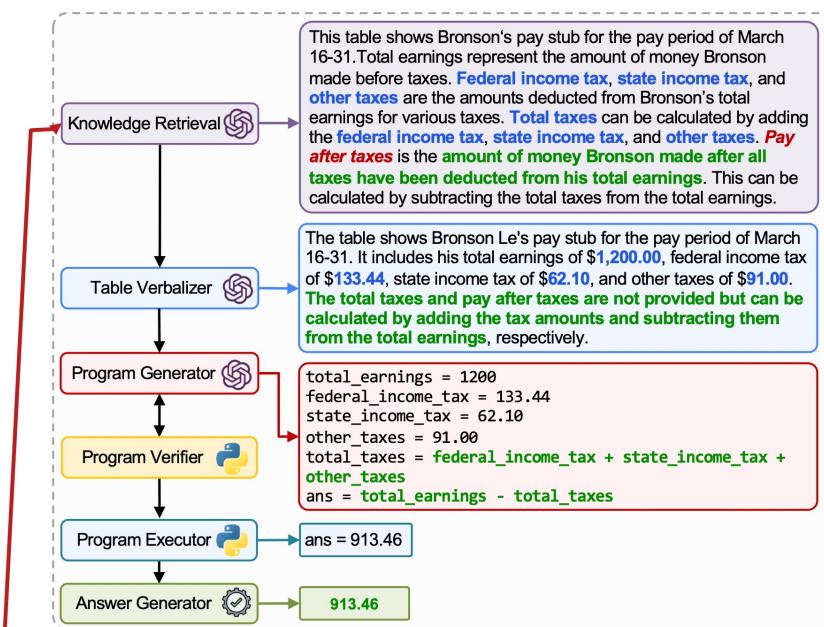
Look at Bronson's pay stub. Bronson lives in a state that has state income tax. **How much did Bronson make after taxes?**

Miles hiked	
Day	Miles
Sunday	10
Monday	9
Tuesday	10
Wednesday	5
Thursday	9

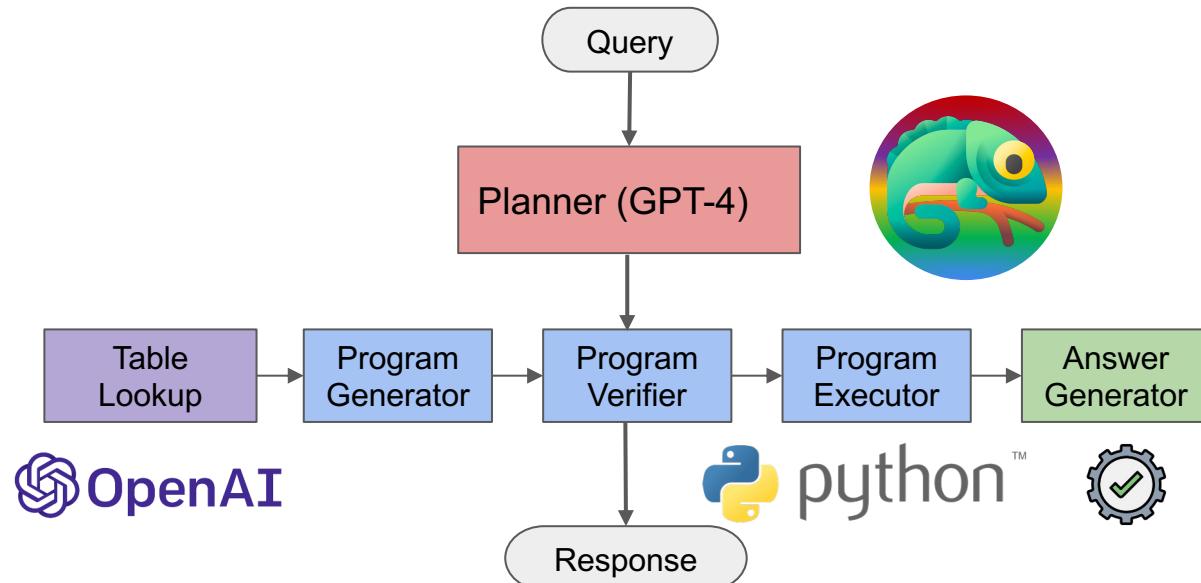
Wanda went on a camping trip and logged the number of miles she hiked each day. What is the **median** of the numbers?

Bus schedule		
Location	Arrive	Depart
stadium	10:20 A.M.	10:25 A.M.
park	10:35 A.M.	10:45 A.M.
hotel	11:10 A.M.	11:15 A.M.
airport	12:05 P.M.	12:10 P.M.
train station	12:25 P.M.	12:35 P.M.
bus station	1:10 P.M.	1:10 P.M.

Look at the following schedule. When does the bus **depart** from the **train station**?
 (A) 12:35 P.M. (B) 1:10 P.M.
 (C) 1:10 P.M. (D) 10:45 A.M.



Chameleon: LLM Planner



Chameleon: LLM Planner



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 to solve the query.

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.

Chameleon: LLM Planner



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

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.

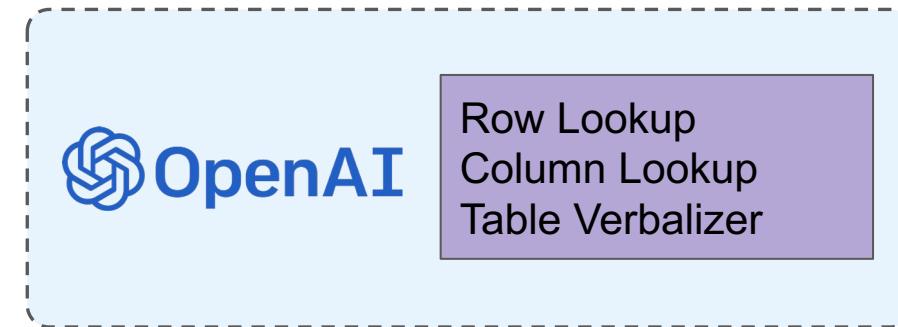
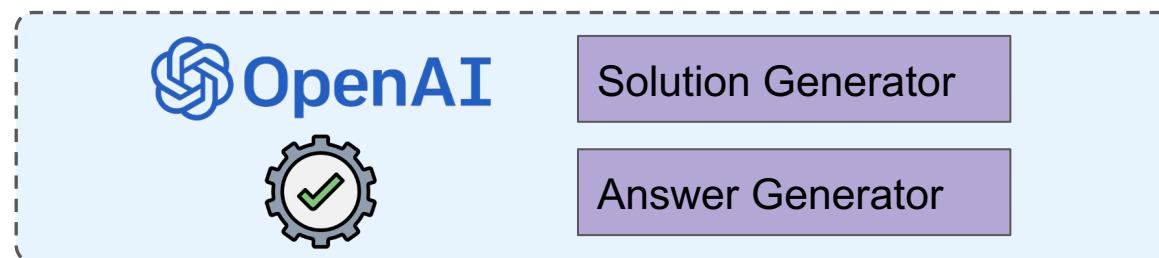
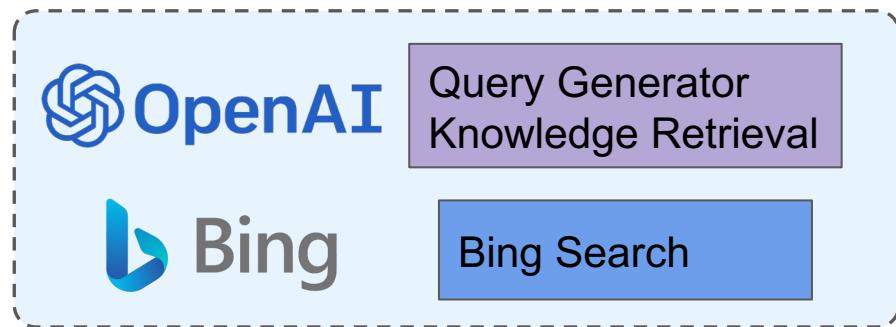
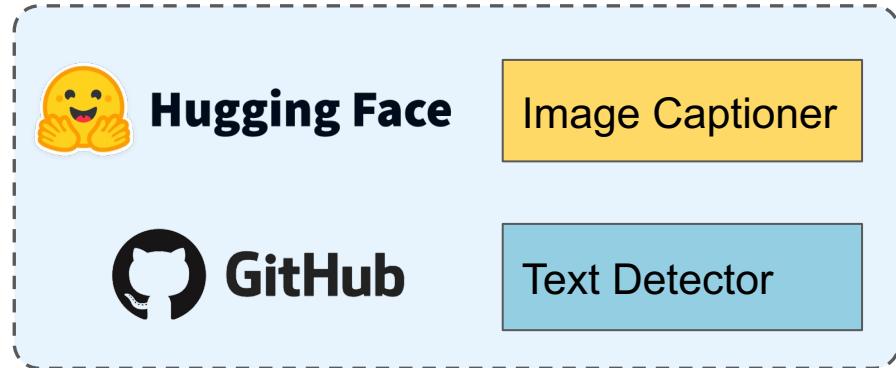
Options: (A) neither; the samples have the same temperature (B) sample A (C) sample B

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"]



Chameleon: Module Inventory





Chameleon: Knowledge Retrieval



Read the following question, and generate the background knowledge as the context information that could be helpful for answering the question.

Question: Which property do these three objects have in common?

Options: (A) hard (B) soft (C) yellow

Metadata: 'pid': 43, 'has_image': True, 'grade': 4, 'subject': 'natural science', 'topic': 'physics', 'category': 'Materials', 'skill': 'Compare properties of objects'

Detected text in the image: ['handkerchief', 'slippers', 'leisure suit']

Knowledge:

- This question is about comparing the properties of three objects: a handkerchief, slippers, and a leisure suit.
- The objects are related to the topic of physics and the skill of comparing properties of objects.
- Properties of objects can include physical characteristics such as color, texture, shape, size, weight, and material.



Chameleon: Query Generator

Read the following question and metadata, and generate the query for browser search as the context information that could be helpful for answering the question.

Question: Which property do these two objects have in common?

Options: (A) hard (B) bendable

Metadata: 'pid': 329, 'has_image': True, 'grade': 2, 'subject': 'natural science', 'topic': 'physics', 'category': 'Materials', 'skill': 'Compare properties of objects'

Detected text in the image: ([[41, 183], [131, 183], [131, 199], [41, 199]], 'rubber gloves'), ([[245, 183], [313, 183], [313, 197], [245, 197]], 'rain boots')

Search Query: Common material properties of jump rope and rubber gloves





Chameleon: Row Lookup



Read the following question and table. Each row is separated by a newline ('\n') and each column is separated by a vertical bar ('|'). Return the simplified table that only remains the rows that are relevant to the question. If all rows are relevant, or the number of rows is fewer than three, return the original table.

Question: In preparation for graduation, some teachers and students volunteered for the various graduation committees. How many people are on the music committee?

Table:

Committee	Students	Teachers
Program	5	17
Ticket	20	5
Music	20	15
Schedule	15	20
Food	18	2

Simplified Table:

Committee	Students	Teachers
Music	20	15



Chameleon: Table Verbalizer



Read the following question and table. Write a textual description of the table. The description should keep the critical information in the table for answering the question. The description should not answer the question.

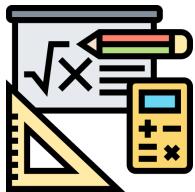
Table:

Committee	Students	Teachers
Program	5	17
Ticket	20	5
Music	20	15
Schedule	15	20
Food	18	2

Table description: The table shows the number of students and teachers on each of the four graduation committees: Program, Ticket, Music, and Schedule. The Music committee has 20 students and 15 teachers.



Chameleon: Program Generator



Read the following table and then write Python code to answer a question.

Table:

Price	Quantity demanded	Quantity supplied
\$895	21,000	3,400
\$945	17,200	7,400
\$995	13,400	11,400
\$1,045	9,600	15,400
\$1,095	5,800	19,400

Questions: Look at the table. Then answer the question. At a price of \$995, is there a shortage or a surplus? Please select from the following options: ['shortage', 'surplus'].

Code:

```
# Python Code, return 'ans'. Make sure that 'ans' is a string selected
# from the options in the question
quantity_demanded_at_price_955 = 13400
quantity_supplied_at_price_955 = 11400
if quantity_demanded_at_price_955 > quantity_supplied_at_price_955:
    ans = 'shortage'
else:
    ans = 'surplus'
```



Chameleon: Solution Generator



 OpenAI

Given the question (and the context), select the answer from the options ["A", "B", "C", "D", "E"]. You should give concise and step-by-step solutions. Finally, conclude the answer in the format of "the answer is [ANSWER]", where [ANSWER] is one from the options ["A", "B", "C", "D", "E"].

Question: Which property do these two objects have in common?

Context: Select the better answer.

Options: (A) hard (B) bendable

Metadata: 'pid': 6493, 'has_image': True, 'grade': 2, 'subject': 'natural science', 'topic': 'physics', 'category': 'Materials', 'skill': 'Compare properties of objects'
Image caption: A pair of scissors next to a pair of scissors.

Detected text with coordinates in the image: [([53, 185], 'jump rope'), ([233, 183],), 'rubber gloves')]

Retrieved knowledge:

- Properties of objects can include physical characteristics such as color, texture, shape, size, weight, and material. In this case, the two objects have the property of being bendable in common.

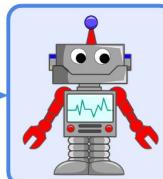
Solution: An object has different properties. ... The property that both objects have in common is bendable. Therefore, the answer is B.

ScienceQA: Science Question Answering

Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



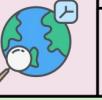
Answer: The answer is A.

BECAUSE:

Lecture: A force is a **push** or a **pull** that one object applies to a second object. The direction of a **push** is **away from** the object that is pushing. The direction of a **pull** is **toward** the object that is pulling.

Explanation: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a **pull**.

Domain Diversity in ScienceQA

Nature Science	Social Science	Language Science		
Biology Genes to traits  Classification Adaptations Traits and heredity Ecosystems Classification Scientific names Heredity Ecological interactions Cells Plants Animals Plant reproduction	Physics Materials  Magnets Velocity and forces Force and motion Particle motion and energy Heat and thermal energy States of matter Kinetic and potential energy Mixture	Geography State capitals  Geography Maps Oceania: geography Physical Geography The Americas: geography Oceans and continents Cities States	History Colonial America English colonies in North America The American Revolution World History Greece Ancient Mesopotamia World religions American history  Medieval Asia	Civics Social skills  Government The Constitution
Chemistry Solutions  Physical and chemical change Atoms and molecules Chemical reactions	Writing Strategies Supporting arguments Sentences, fragments, and run-ons Word usage and nuance Creative techniques Audience, purpose, and tone Pronouns and antecedents Persuasive strategies Editing and revising Visual elements Opinion writing	Vocabulary Categories  Shades of meaning Comprehension strategies Context clues	Verbs Verb tense 	
Earth Science Weather and climate Rocks and minerals Astronomy Fossils Earth events  Plate tectonics	Engineering Designing experiments  Engineering practices	Grammar Sentences and fragments Phrases and clauses	Capitalization Formatting	
		Figurative Language Literary devices 	Punctuation Fragments	
			Phonology Rhyming	
			Reference Research skills	

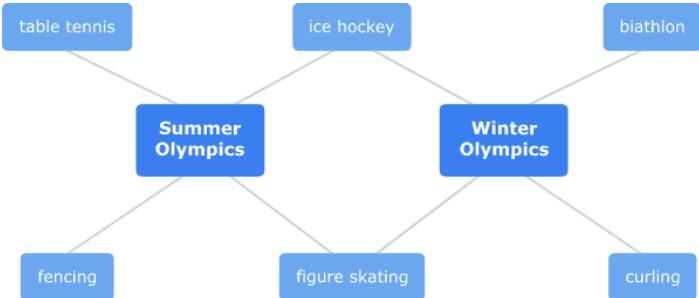
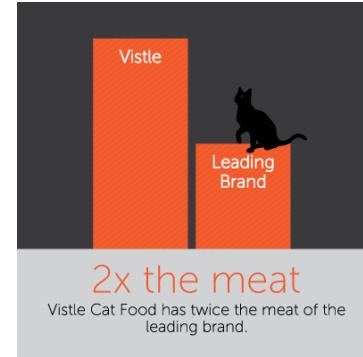
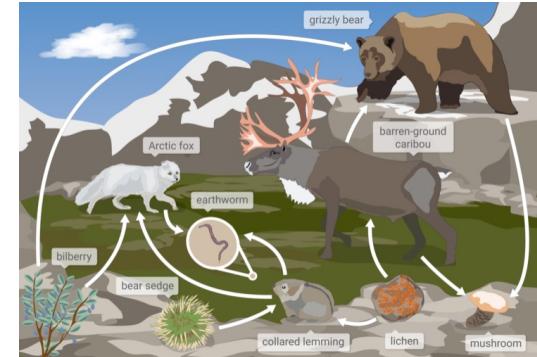
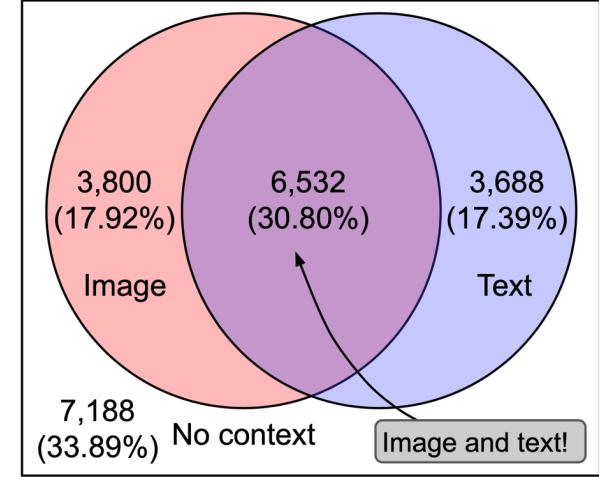
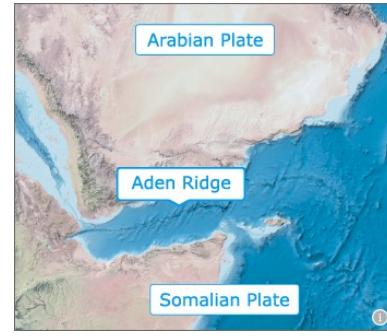
3 subjects

26 topics

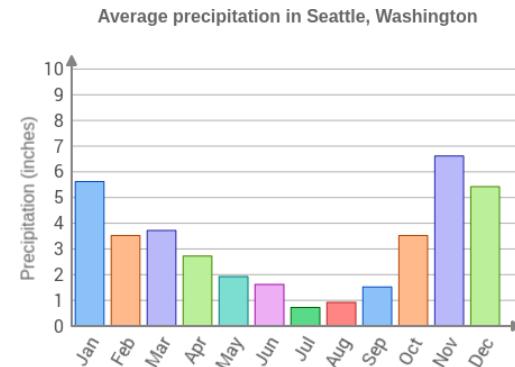
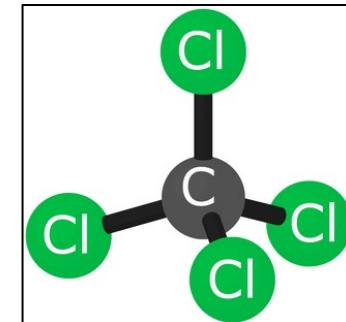
127 categories

379 skills

Context Diversity in ScienceQA



Planet	Volume (billions of km ³)	Primary composition
Mercury	60	rock
Venus	930	rock
Earth	1,090	rock
Mars	160	rock
Jupiter	1,431,280	gas
Saturn	827,130	gas
Uranus	68,330	ice
Neptune	62,530	ice





Module Inventory for ScienceQA



Hugging Face

Image Captioner



GitHub

Text Detector



OpenAI

Knowledge Retrieval
Query Generator
Solution Generator



Bing

Bing Search



Answer Generator

Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



Answer: The answer is A.

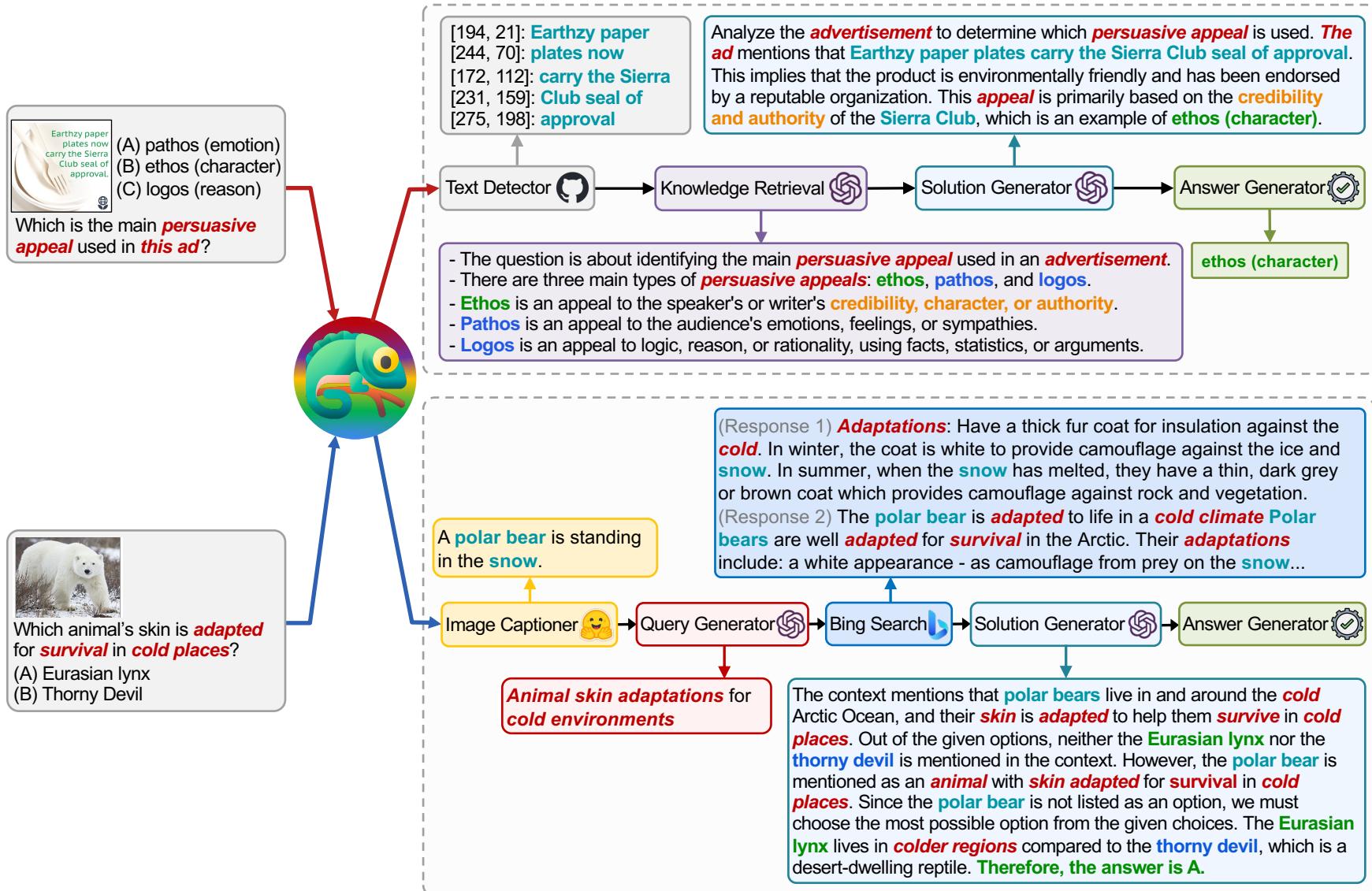
BECAUSE:

Lecture: A force is a **push** or a **pull** that one object applies to a second object. The direction of a **push** is **away from** the object that is pushing. The direction of a **pull** is **toward** the object that is pulling.

Explanation: The **baby's hand** applies a **force** to the **cabinet door**. This force causes the **door** to **open**. The direction of this force is **toward** the **baby's hand**. This force is a **pull**.



Chameleon for ScienceQA



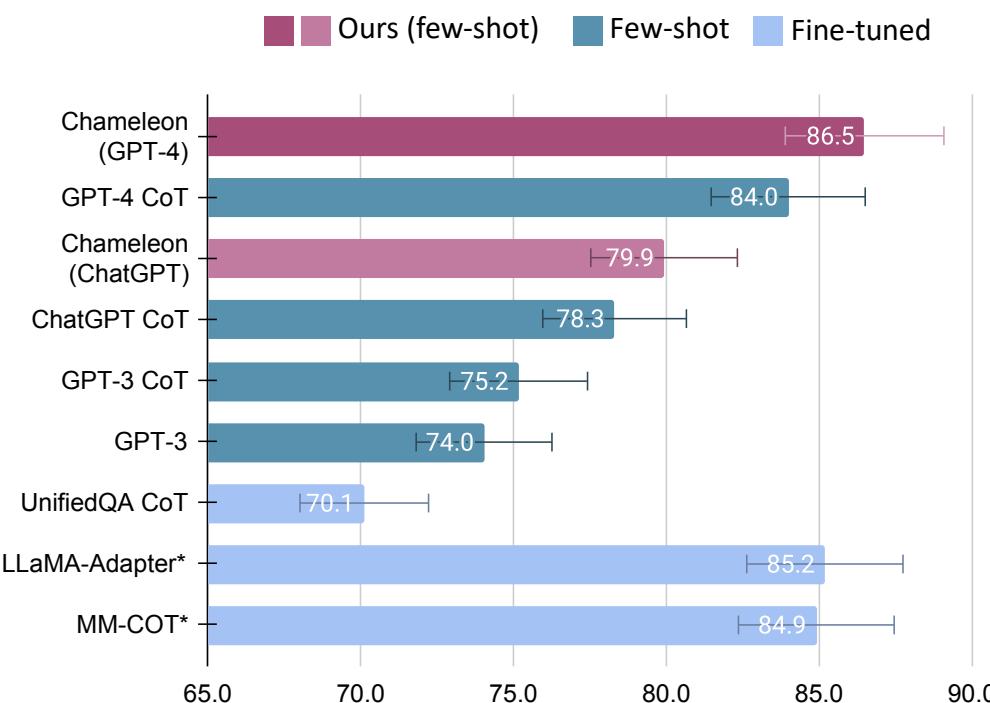


Results on ScienceQA

Model	#Tuned Params	ALL	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
<i>Heuristic baselines</i>										
Random Choice [27]	-	39.83	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67
Human [27]	-	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
<i>Fine-tuned models</i>										
MCAN [55]	95M	54.54	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72
Top-Down [1]	70M	59.02	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16
BAN [18]	112M	59.37	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94
DFAF [10]	74M	60.72	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17
ViLT [19]	113M	61.14	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90
Patch-TRM [29]	90M	61.42	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50
VisualBERT [22, 23]	111M	61.87	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92
UnifiedQA [17]	223M	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
UnifiedQA CoT [27]	223M	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
MM-COT _T [57]	223M	70.53	71.09	70.75	69.18	71.16	65.84	71.57	71.00	69.68
MM-COT [57]	223M	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37
MM-COT _{Large} [57]	738M	91.68	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31
LLaMA-Adapter _T [56]	1.2M	78.31	79.00	73.79	80.55	78.30	70.35	83.14	79.77	75.68
LLaMA-Adapter [56]	1.8M	85.19	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05
<i>Few-shot GPT-3</i>										
GPT-3 [4]	0M	74.04	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87
GPT-3 CoT [27]	0M	75.17	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68

Published results (Above) ▲

Few-shot ChatGPT										
ChatGPT CoT	0M	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
Chameleon (ChatGPT)	0M	79.93	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53
Few-shot GPT-4										
GPT-4 CoT	0M	83.99	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04
Chameleon (GPT-4)	0M	86.54	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72



TabMWP: Tabular Math Word Problems

square beads	\$2.97 per kilogram
oval beads	\$3.41 per kilogram
flower-shaped beads	\$2.18 per kilogram
star-shaped beads	\$1.95 per kilogram
heart-shaped beads	\$1.52 per kilogram
spherical beads	\$3.42 per kilogram
rectangular beads	\$1.97 per kilogram

Question: If Tracy buys 5 kilograms of spherical beads, 4 kilograms of star-shaped beads, and 3 kilograms of flower-shaped beads, how much will she spend? (unit: \$)

Answer: **31.44**

Solution:

Find the cost of the spherical beads. Multiply: $\$3.42 \times 5 = \17.10 .

Find the cost of the star-shaped beads. Multiply: $\$1.95 \times 4 = \7.80 .

Find the cost of the flower-shaped beads. Multiply: $\$2.18 \times 3 = \6.54 .

Now find the total cost by adding: $\$17.10 + \$7.80 + \$6.54 = \31.44 .

She will spend **31.44**.

Sandwich sales		
Shop	Tuna	Egg salad
City Cafe	6	5
Sandwich City	3	12
Express Sandwiches	7	17
Sam's Sandwich Shop	1	6
Kelly's Subs	3	4

Question: As part of a project for health class, Cara surveyed local delis about the kinds of sandwiches sold. Which shop sold fewer sandwiches, Sandwich City or Express Sandwiches?

Options: (A) Sandwich City (B) Express Sandwiches

Answer: **(A) Sandwich City**

Solution:

Add the numbers in the Sandwich City row. Then, add the numbers in the Express Sandwiches row.

Sandwich City: $3 + 12 = 15$. Express Sandwiches: $7 + 17 = 24$.

15 is less than 24. **Sandwich City** sold fewer sandwiches.

2 Tasks

38,431 Problems

35,442 Solutions

37,644 Tables

12.9/54 Avg/Max cells

Context Diversity in TabMWP

Bus schedule					
the school	8:00 A.M.	9:15 A.M.	9:30 A.M.	10:00 A.M.	11:00 A.M.
the zoo	9:00 A.M.	10:15 A.M.	10:30 A.M.	11:00 A.M.	12:00 P.M.
the mall	9:15 A.M.	10:30 A.M.	10:45 A.M.	11:15 A.M.	12:15 P.M.
the grocery store	9:30 A.M.	10:45 A.M.	11:00 A.M.	11:30 A.M.	12:30 P.M.
the science museum	10:30 A.M.	11:45 A.M.	12:00 P.M.	12:30 P.M.	1:30 P.M.
the library	11:15 A.M.	12:30 P.M.	12:45 P.M.	1:15 P.M.	2:15 P.M.
the kickball field	11:45 A.M.	1:00 P.M.	1:15 P.M.	1:45 P.M.	2:45 P.M.
the playground	12:45 P.M.	2:00 P.M.	2:15 P.M.	2:45 P.M.	3:45 P.M.
the doctor's office	1:15 P.M.	2:30 P.M.	2:45 P.M.	3:15 P.M.	4:15 P.M.

topaz	\$18.55 per lb
amethyst	\$19.88 per lb
tiger's eye	\$10.29 per lb
fool's gold	\$16.00 per lb
quartz	\$14.63 per lb
calcite	\$15.39 per lb
granite	\$19.23 per lb

Employee	Pay period
Dhruba Khanal	December 9-15
Total earnings	\$620.00
Federal income tax	\$71.14
State income tax	\$48.90
Other taxes	\$47.00
Total taxes	?
Pay after taxes	?

heart-shaped beads	\$3/kilogram
rectangular beads	\$2/kilogram
spherical beads	\$2/kilogram
oval beads	\$2/kilogram

Option	Change in phone price
Add an upgrade	\$60
Buy a used phone	-\$75

Watermelons harvested	
Day	Number of watermelons
Wednesday	59
Thursday	51
Friday	53
Saturday	52

Birthday party		
Activity	Parents	Children
Singing	14	20
Eating cake	5	10
Jumping rope	16	20
Swimming	16	19
Playing tag	4	9

Park	Number of basketball hoops
Heron Park	2
Kelly Park	7
Westfield Park	4
Pinehurst Park	4
Linden Park	3
Mooreland Park	7
Crestview Park	2
Riverfront Park	4

Stem	Leaf
1	1, 3, 6
2	2, 3, 3, 6
3	0
4	3
5	2, 6
6	0, 0, 2
7	2, 8
8	4, 5, 5, 6, 7



Module Inventory for ScienceQA



Row Lookup
Column Lookup
Table Verbalizer
Knowledge Retrieval
Solution Generator
Program Generator

Program Verifier
Program Executor

Answer Generator

square beads	\$2.97 per kilogram
oval beads	\$3.41 per kilogram
flower-shaped beads	\$2.18 per kilogram
star-shaped beads	\$1.95 per kilogram
heart-shaped beads	\$1.52 per kilogram
spherical beads	\$3.42 per kilogram
rectangular beads	\$1.97 per kilogram

Sandwich sales		
Shop	Tuna	Egg salad
City Cafe	6	5
Sandwich City	3	12
Express Sandwiches	7	17
Sam's Sandwich Shop	1	6
Kelly's Subs	3	4

Question: If Tracy buys 5 kilograms of spherical beads, 4 kilograms of star-shaped beads, and 3 kilograms of flower-shaped beads, how much will she spend? (unit: \$)

Answer: **31.44**

Solution:

Find the cost of the spherical beads. Multiply: $\$3.42 \times 5 = \17.10 .

Find the cost of the star-shaped beads. Multiply: $\$1.95 \times 4 = \7.80 .

Find the cost of the flower-shaped beads. Multiply: $\$2.18 \times 3 = \6.54 .

Now find the total cost by adding: $\$17.10 + \$7.80 + \$6.54 = \31.44 .

She will spend **\\$31.44**.

Question: As part of a project for health class, Cara surveyed local delis about the kinds of sandwiches sold. Which shop sold fewer sandwiches, Sandwich City or Express Sandwiches?

Options: (A) Sandwich City (B) Express Sandwiches

Answer: **(A) Sandwich City**

Solution:

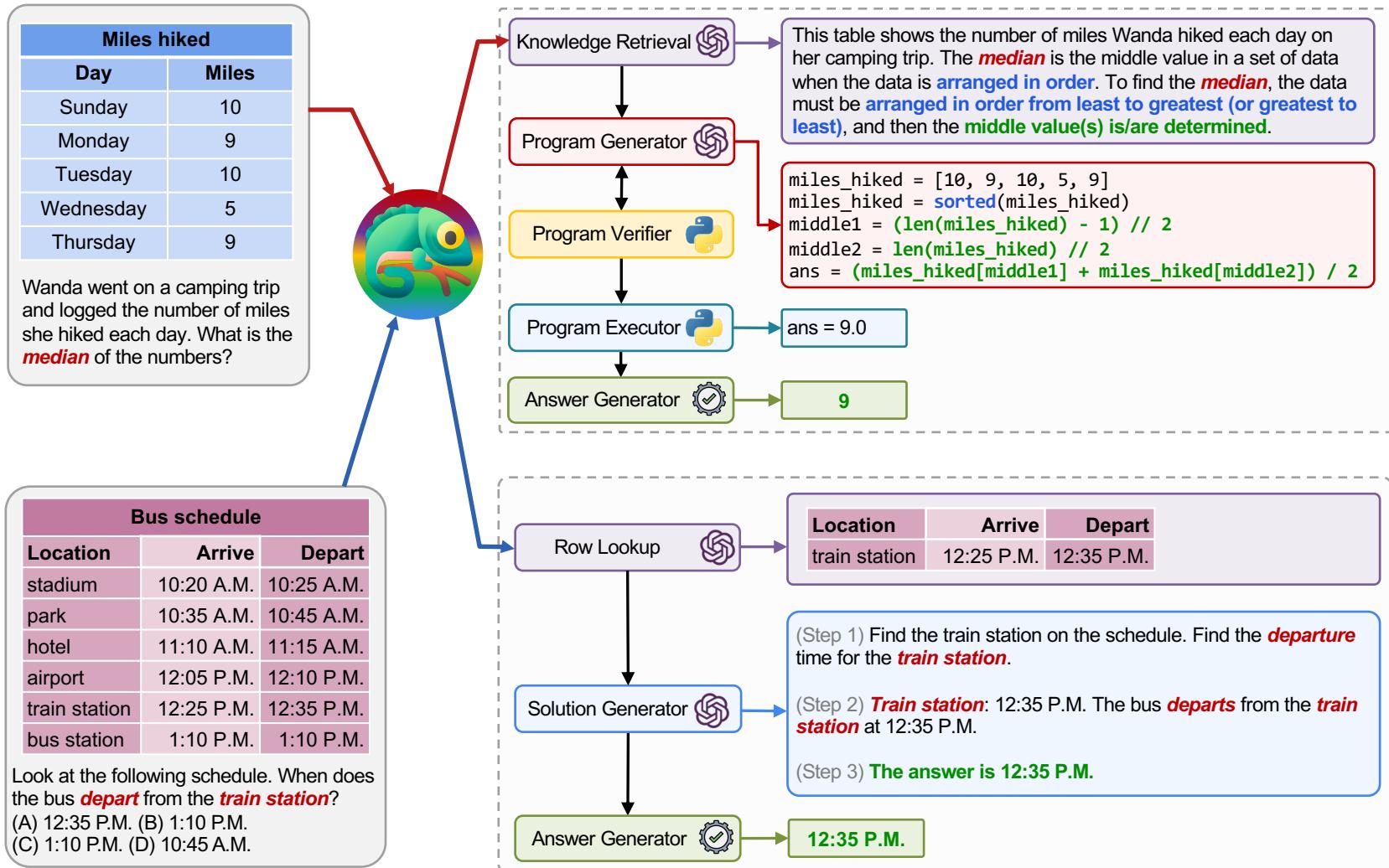
Add the numbers in the Sandwich City row. Then, add the numbers in the Express Sandwiches row.

Sandwich City: $3 + 12 = 15$. Express Sandwiches: $7 + 17 = 24$.

15 is less than 24. **Sandwich City** sold fewer sandwiches.



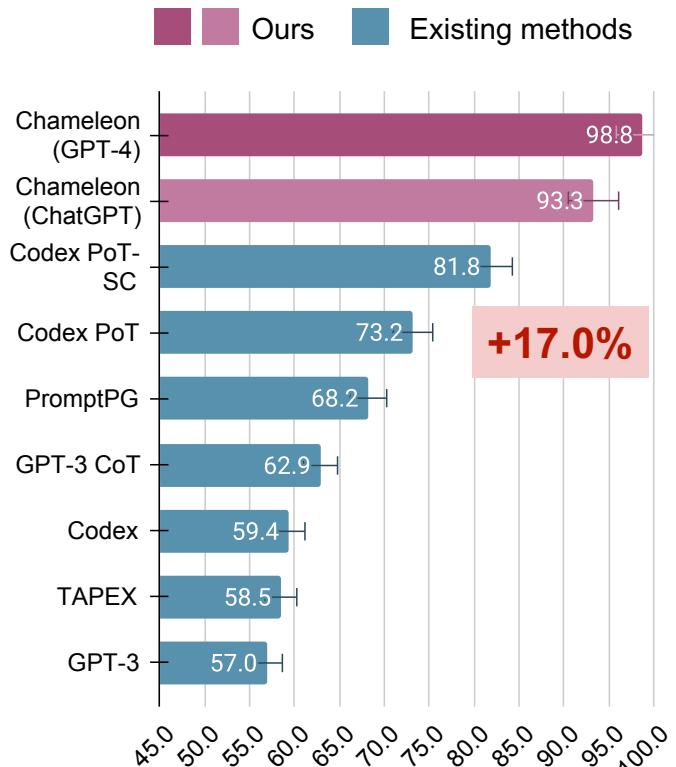
Chameleon for TabMWP





Results on TabMWP

Model	#Tuned Params	ALL	FREE	MC	INT	DEC	EXTR	BOOL	OTH	G1-6	G7-8
<i>Heuristic baselines</i>											
Heuristic guess	-	15.29	6.71	39.81	8.37	0.26	30.80	51.22	26.67	17.55	12.27
Human performance	-	90.22	84.61	93.32	84.95	83.29	97.18	88.69	96.20	94.27	81.28
<i>Fine-tuned models</i>											
UnifiedQA _{SMALL} [17]	41M	29.79	22.27	51.31	27.27	2.83	52.28	48.11	69.52	35.85	21.71
UnifiedQA _{BASE} [17]	223M	43.52	34.02	70.68	40.74	7.90	84.09	55.67	73.33	53.31	30.46
UnifiedQA _{LARGE} [17]	738M	57.35	48.67	82.18	55.97	20.26	94.63	68.89	79.05	65.92	45.92
TAPEX _{BASE} [25]	139M	48.27	39.59	73.09	46.85	11.33	84.19	61.33	69.52	56.70	37.02
TAPEX _{LARGE} [25]	406M	58.52	51.00	80.02	59.92	16.31	95.34	64.00	73.33	67.11	47.07
<i>Zero-shot GPT-3</i>											
GPT-3 [4]	0M	56.96	53.57	66.67	55.55	45.84	78.22	55.44	54.29	63.37	48.41
GPT-3 CoT [49]	0M	57.61	54.36	66.92	55.82	48.67	78.82	55.67	51.43	63.62	49.59
<i>Few-shot GPT-3</i>											
GPT-3 [4]	0M	57.13	54.69	64.11	58.36	40.40	75.95	52.41	53.02	63.10	49.16
GPT-3 CoT [49]	0M	62.92	60.76	69.09	60.04	63.58	76.49	61.19	67.30	68.62	55.31
GPT-3 CoT-PromptPG [28]	0M	68.23	66.17	74.11	64.12	74.16	76.19	72.81	65.71	71.20	64.27
Codex* [5]	0M	59.4	-	-	-	-	-	-	-	-	-
Codex PoT* [6]	0M	73.2	-	-	-	-	-	-	-	-	-
Codex PoT-SC* [6]	0M	81.8	-	-	-	-	-	-	-	-	-
Published results (Above) ▲											
<i>Few-shot ChatGPT</i>											
ChatGPT CoT	0M	82.03	78.43	92.32	75.38	90.30	92.30	92.89	87.62	83.06	80.66
ChatGPT PoT	0M	89.49	90.24	87.35	89.31	93.82	92.10	85.89	55.24	90.60	88.00
Chameleon (ChatGPT)	0M	93.28	93.13	93.72	92.71	94.76	91.29	98.11	78.85	93.37	93.17
<i>Few-shot GPT-4</i>											
GPT-4 CoT	0M	90.81	88.48	97.49	86.16	97.51	96.86	99.11	89.52	92.40	88.70
GPT-4 PoT	0M	96.93	97.40	95.58	98.48	93.22	96.25	98.00	68.57	96.97	96.87
Chameleon (GPT-4)	0M	98.78	98.95	98.29	99.34	97.42	98.58	98.56	93.33	98.95	98.54



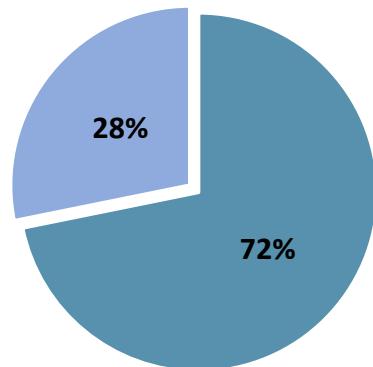


Tool Use for ScienceQA

Called
Not called

Chameleon
(ChatGPT)

Knowledge Retrieval



Text Detector

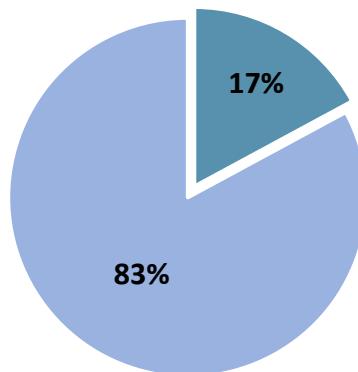
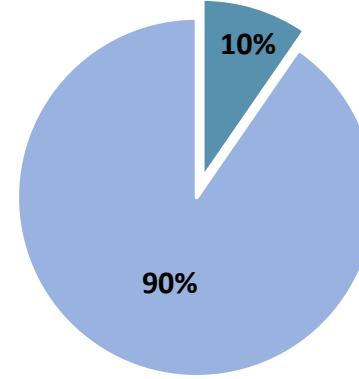
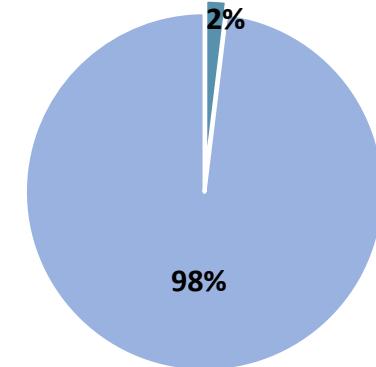


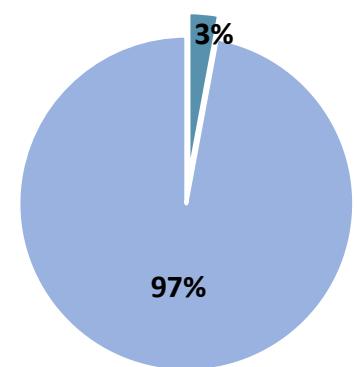
Image Captioner



Query Generator

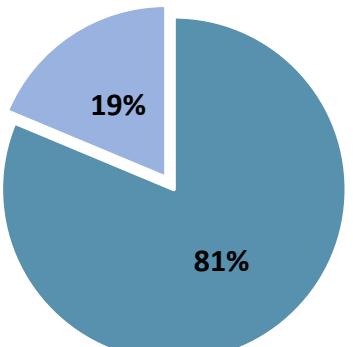


Bing Search



Chameleon
(GPT-4)

Knowledge Retrieval



Text Detector

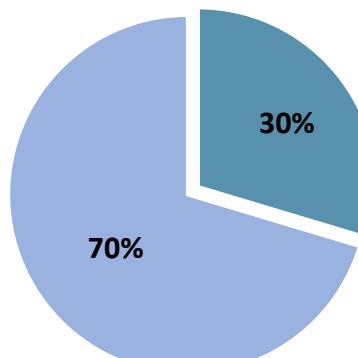
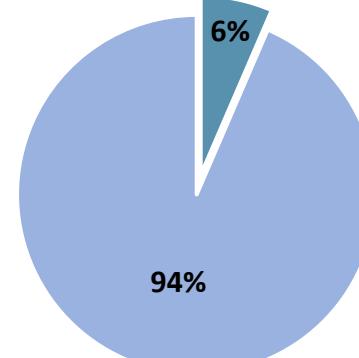
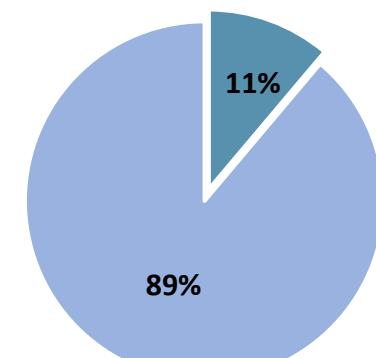


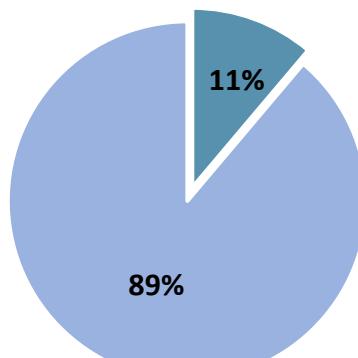
Image Captioner



Query Generator



Bing Search



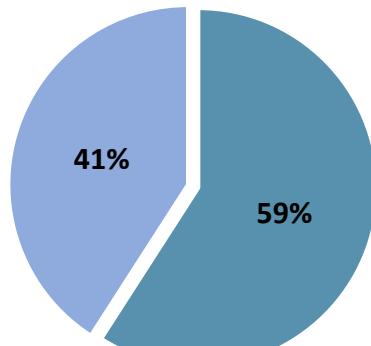


Tool Use for TabMWP

Called
Not called

Chameleon
(ChatGPT)

Row Lookup



Column Lookup

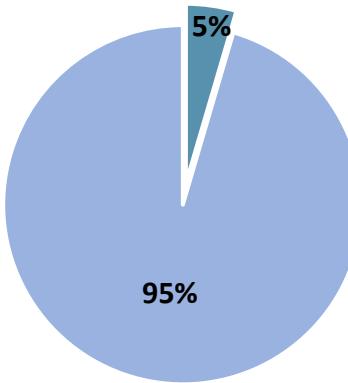
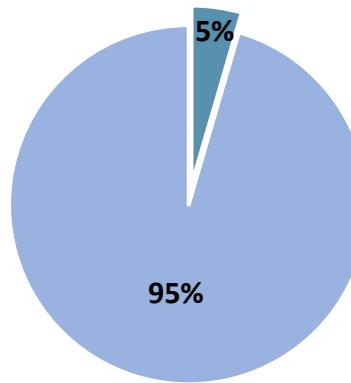
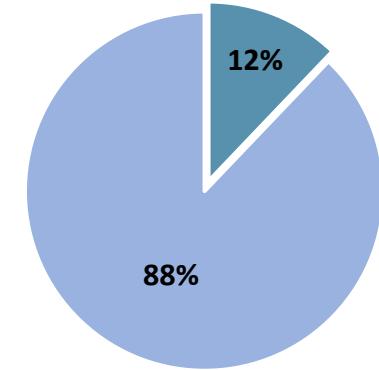


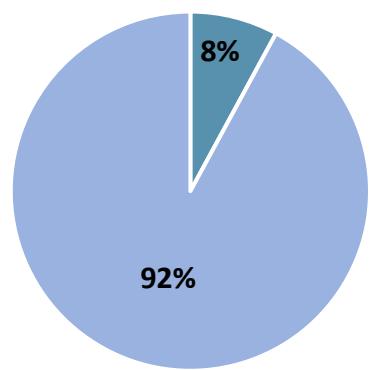
Table Verbalizer



Knowledge Retrieval

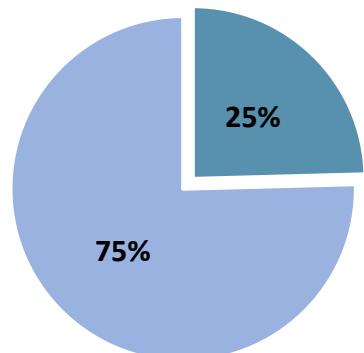


Solution Generator



Chameleon
(GPT-4)

Row Lookup



Column Lookup

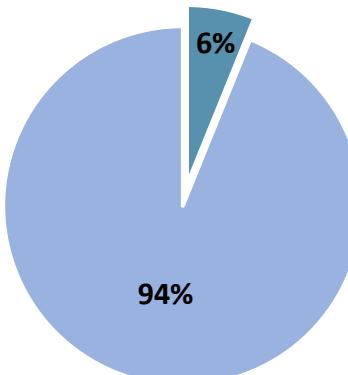
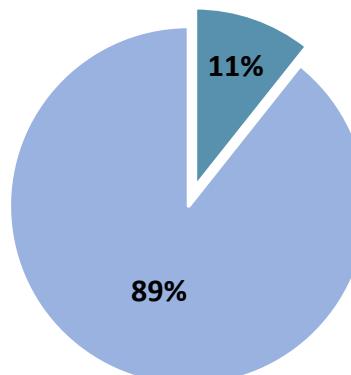
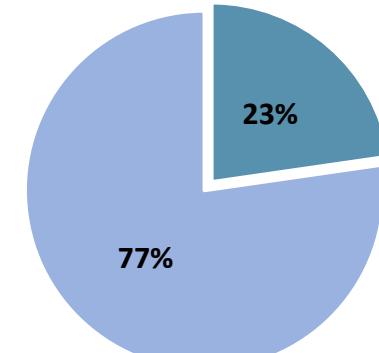


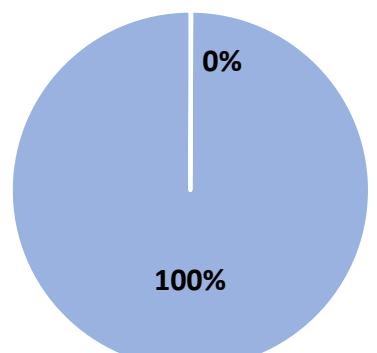
Table Verbalizer

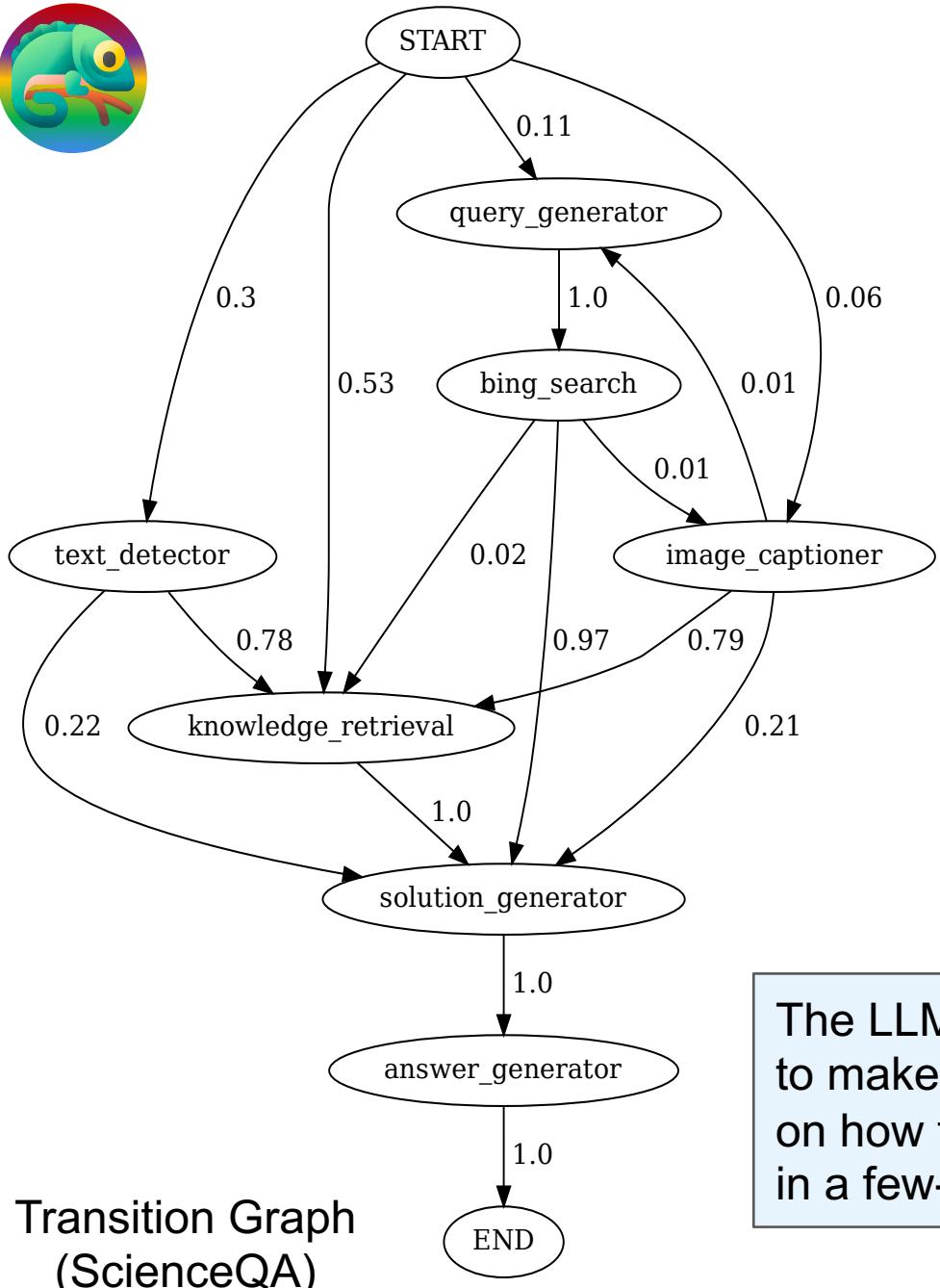


Knowledge Retrieval

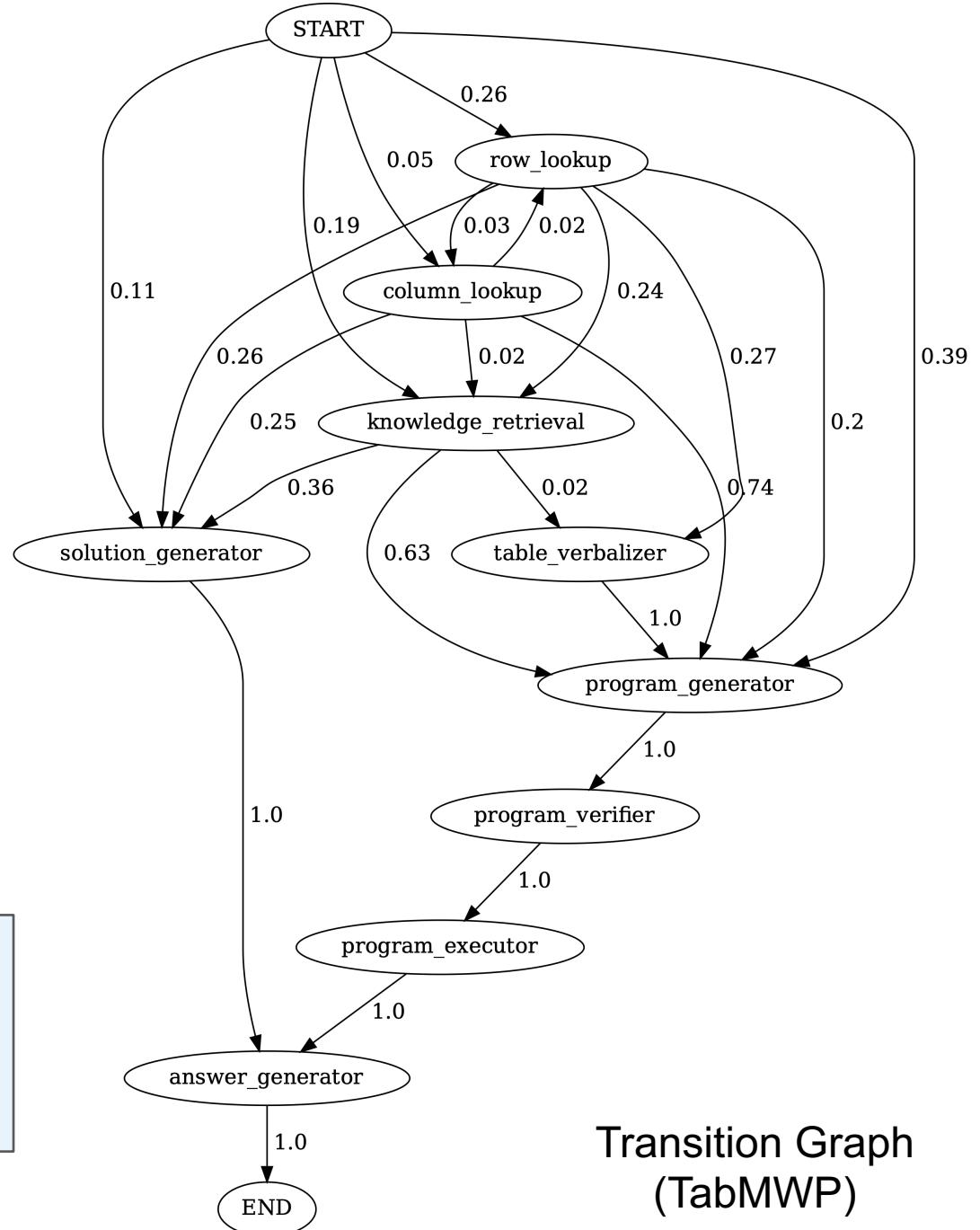


Solution Generator





The LLM planner is able to make good decisions on how to sequence tools in a few-shot setup.



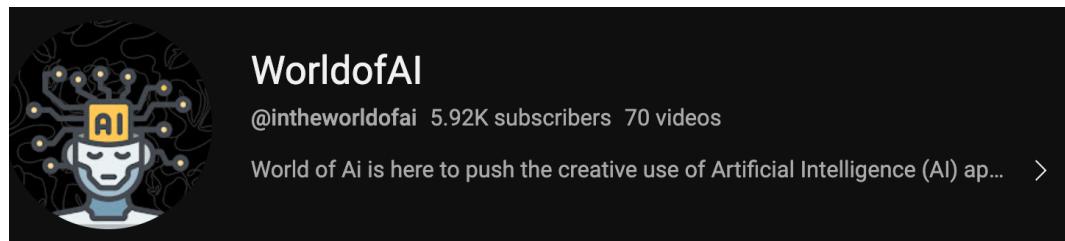


Take-Away

- ❖ **Chameleon**: a plug-and-play compositional reasoning framework that augments LLMs with various tools
- ❖ **Module inventory**: LLM models, off-the-shelf vision models, web search engines, Python functions, and rule-based modules
- ❖ **LLM planner**: generate a program to compose tools by natural language instructions
 - ❖ Efficiently extendable to using new modules
 - ❖ Natural-language-like programs are less error-prone, easy to debug, and user-friendly
 - ❖ Flexible to replace the underlying LLM for the planner as well as each module

More Details

- Paper: <https://arxiv.org/abs/2304.09842>
- Project: <https://chameleon-llm.github.io/>
- Code: <https://github.com/lupantech/chameleon-llm>



<https://www.youtube.com/watch?v=EWFixlk4vjs>