

LLMs and Tools

Tool Augmentation

Advanced Large Language Models

ELL8299 · AIL861 · ELL881



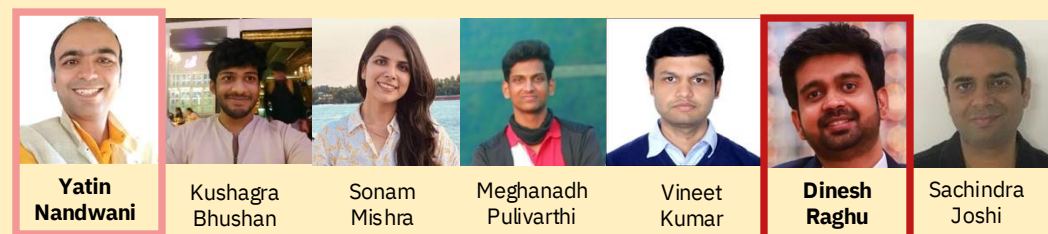
Dinesh Raghu
Senior Researcher & Manager, IBM Research

Conversational AI Group @ IBM Research India

Agentic AI

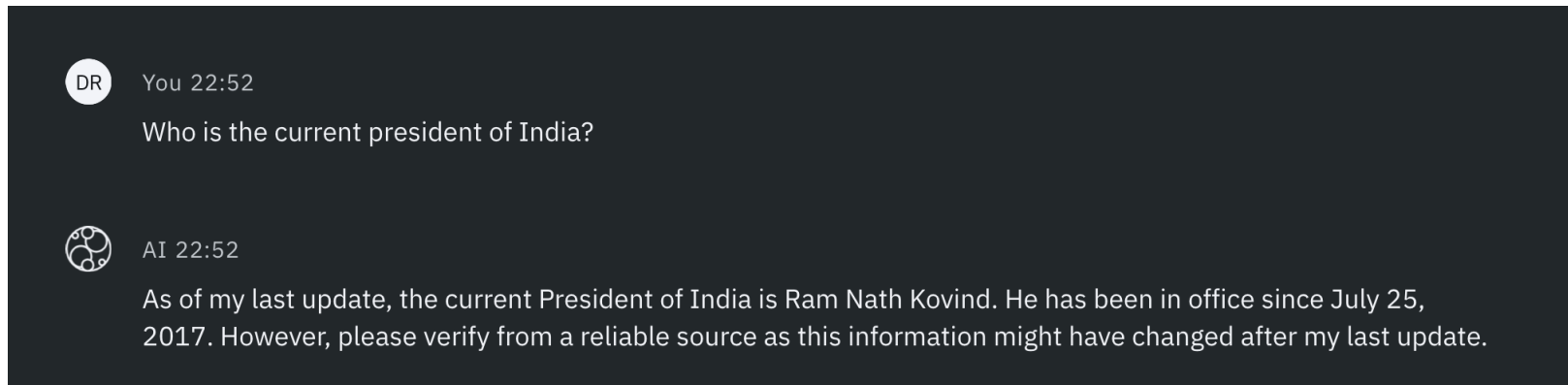


Retrieval Augmented Generation



Limitations of LLMs

1. inability to access up-to-date information on recent events



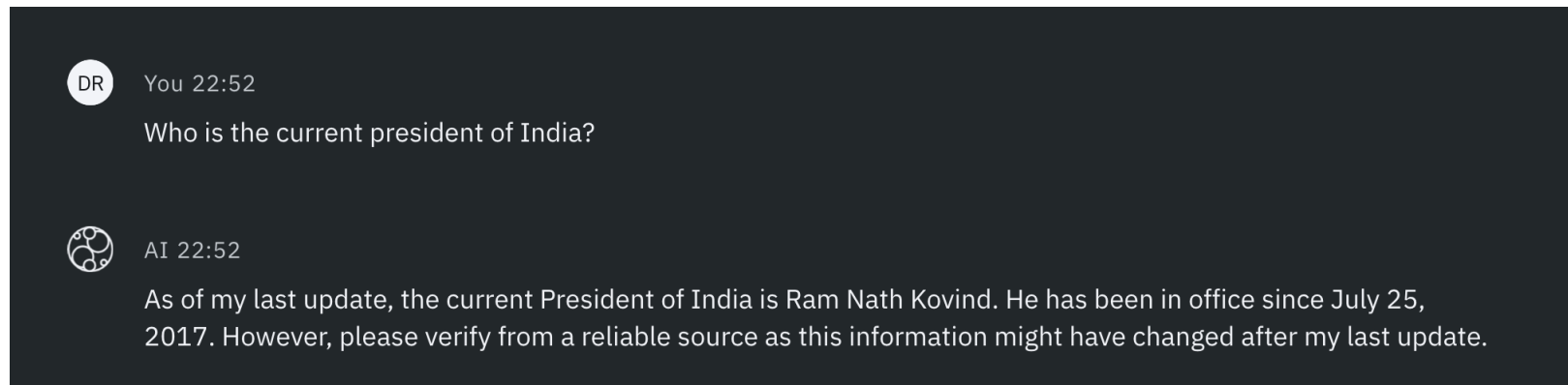
Interaction with **mixtral-8x7b-instruct-v01** in October 2024

Limitations of LLMs



Search Engine

1. inability to access up-to-date information on recent events



Interaction with **mixtral-8x7b-instruct-v01** in October 2024

LLMs with Tools



Search Engine

Who is the current prime minister of India?

Searched 3 sites ^

Q "current prime minister of India October 2024"
bing.com

PM India — Prime Minister of India
pmindia.gov.in

BYJU'S — Cabinet Ministers of India 2024 [Updated] Know the Portfolios of Council of Ministers...
byjus.com

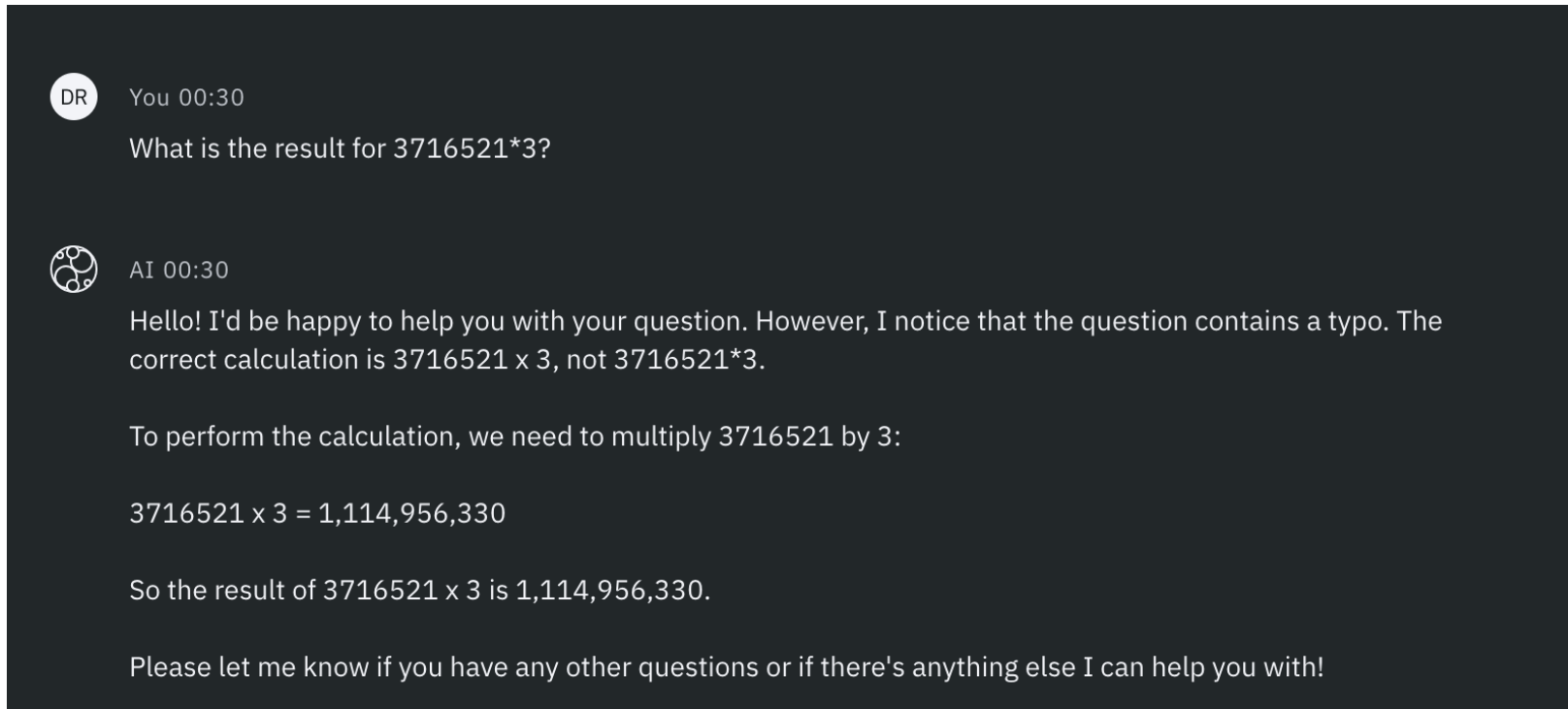
The current Prime Minister of India is **Narendra Modi**. He began his third consecutive term on June 9, 2024, following a decisive victory in the general elections. This term builds upon his previous initiatives focusing on economic development, national security, and welfare programs

PM INDIA BYJU'S

Interaction with **ChatGPT** in October 2024

Limitations of LLMs

2. a lack of mathematical skills to perform precise calculations



The screenshot shows a chat interface with a dark background. On the left, a user icon labeled 'DR' is shown. The user's message is 'What is the result for 3716521*3?'. On the right, an AI icon is shown. The AI's response is: 'Hello! I'd be happy to help you with your question. However, I notice that the question contains a typo. The correct calculation is 3716521 x 3, not 3716521*3. To perform the calculation, we need to multiply 3716521 by 3: 3716521 x 3 = 1,114,956,330. So the result of 3716521 x 3 is 1,114,956,330. Please let me know if you have any other questions or if there's anything else I can help you with!'.

Correct Answer:
11,149,563

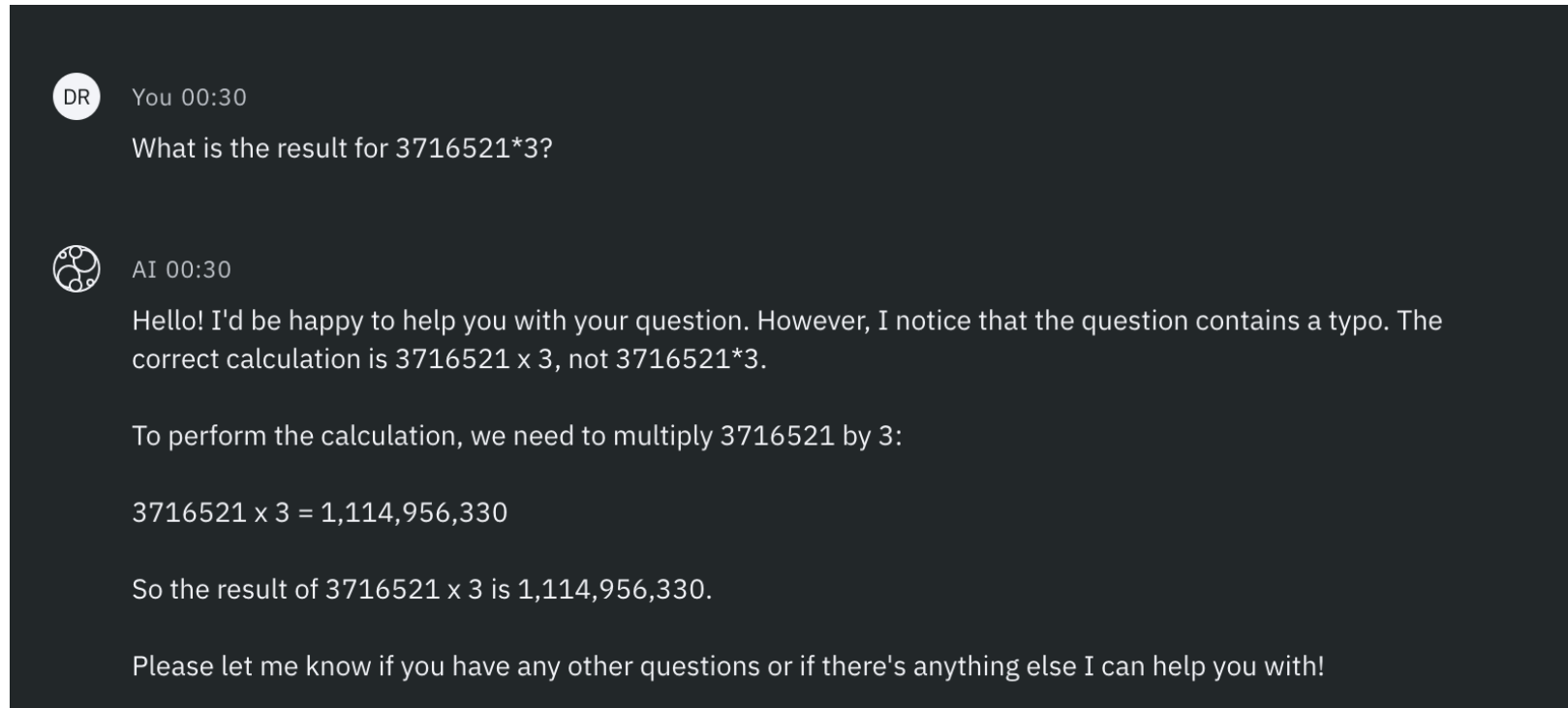
Interaction with **llama-2-13b-chat** in October 2024

Limitations of LLMs



Calculator

2. a lack of mathematical skills to perform precise calculations



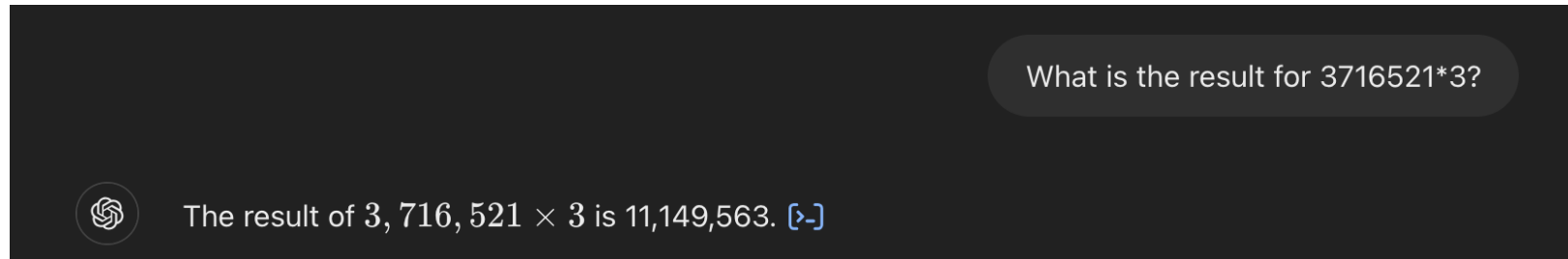
Correct Answer:
11,149,563

Interaction with **llama-2-13b-chat** in October 2024

LLMs with Tools



Calculator

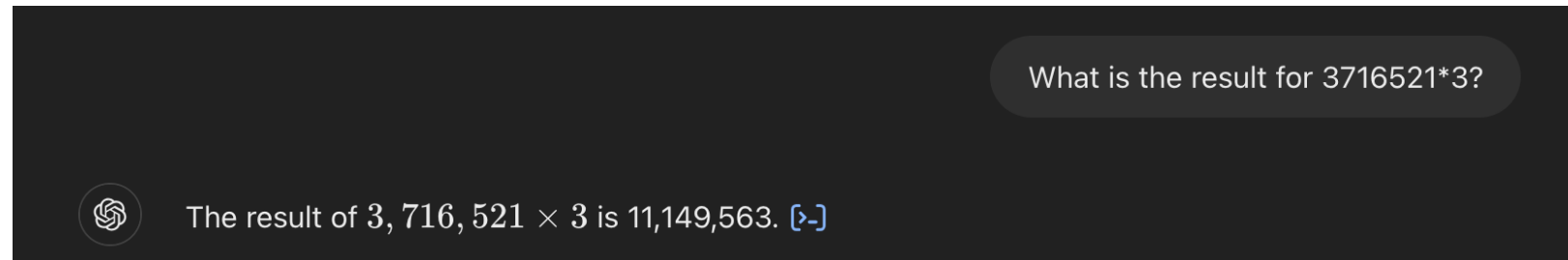


Interaction with **ChatGPT** in October 2024

LLMs with Tools



Calculator



```
python

# Calculating the product
result = 3716521 * 3
result

Result
11149563
```

Interaction with **ChatGPT** in October 2024

Why are Tools important?

1. Overcoming LLM knowledge limits
 - “Who leads the Super 4s table in Asia Cup 2025?” (Web Search Tool)
2. Accuracy in computation & reasoning
 - “What’s the determinant of this 4×4 matrix?” (Python Function)

Why are Tools important?

1. Overcoming LLM knowledge limits
 - “Who leads the Super 4s table in Asia Cup 2025?”(Web Search Tool)
2. Accuracy in computation & reasoning
 - “What’s the determinant of this 4×4 matrix?” (Python Function)
3. Accessing enterprise or private data (Enterprise APIs)
 - “Show me the top 10 customers by revenue this quarter.”
 - “What’s the flight status of AI 176 from Delhi to San Francisco?”
 - An autonomous agent, that can create a PR for high priority issue raised in your GitHub Org.

LLMs and Tools

Part 1: Incorporating Tools during Fine-tuning

Part 2: Teaching LLMs to Use External APIs

Part 3: Automating Complex Tasks

LLMs and Tools

Part 1: Incorporating Tools during Fine-tuning (Tool Augmentation)

Part 2: Teaching LLMs to Use External APIs (Function Calling)

Part 3: Automating Complex Tasks (AI Agents)

Tool Augmentation: Outline

1. GSM8K: Math Word Problems
2. Tool Augmented Language Models (TALM)
3. Program-aided Language Models (PAL)
4. Toolformer
5. WebGPT (if time permits)

Calculator as a Tool: GSM8K

- a dataset of 8.5K high quality linguistically diverse Grade School Math word problems

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Final Answer: 6

- GPT-3 175B performance was ~20%
- GPT-3 6B performance was ~8%

*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

Calculator as a Tool: GSM8K

- a dataset of 8.5K high quality linguistically diverse Grade School Math word problems

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4*2 = \langle\langle 4*2=8 \rangle\rangle$ 8 dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12*8 = \langle\langle 12*8=96 \rangle\rangle$ 96 cookies

She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = \langle\langle 96/16=6 \rangle\rangle$ 6 cookies

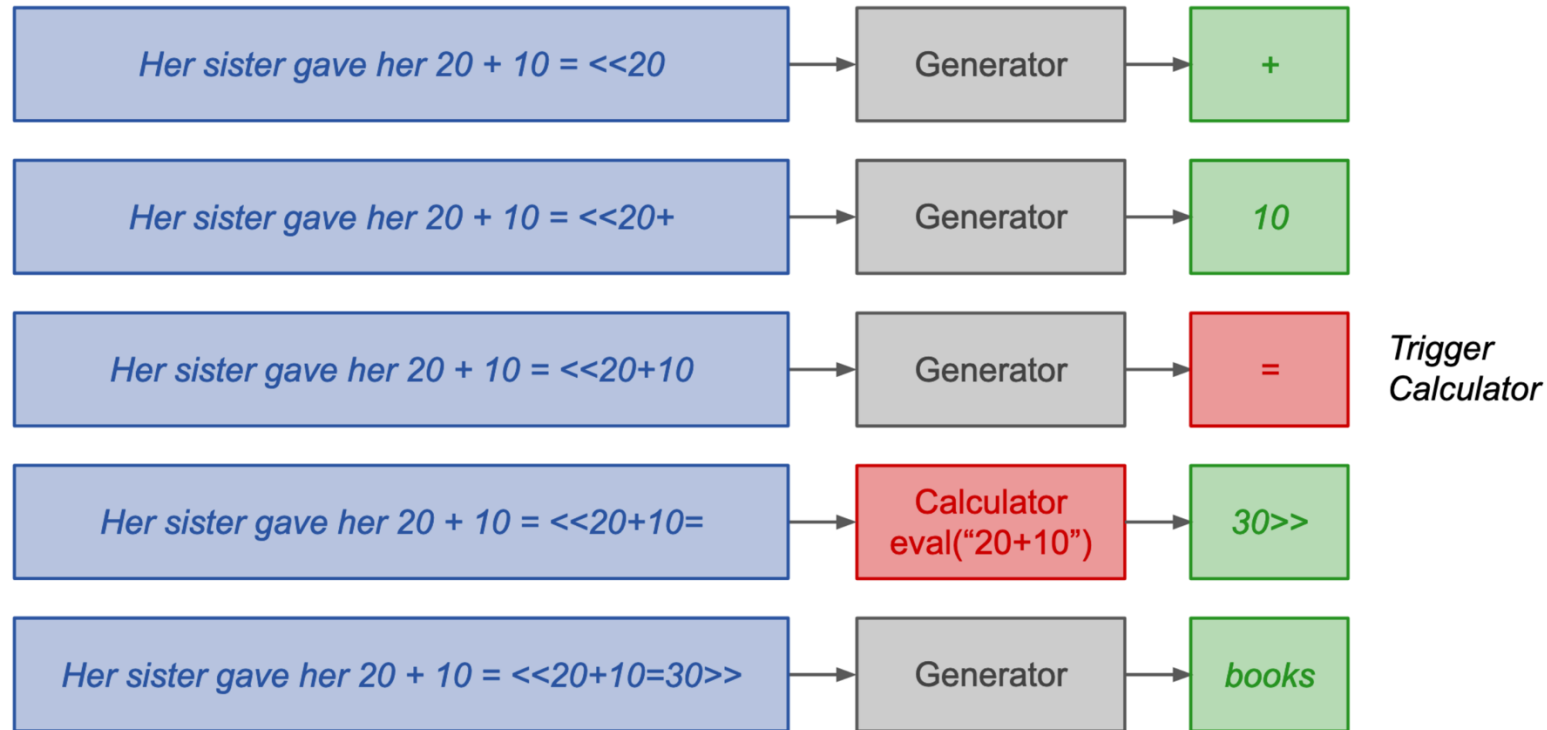
Final Answer: 6

- GPT 3 (6B and 175B) performance was ~10-20%

*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

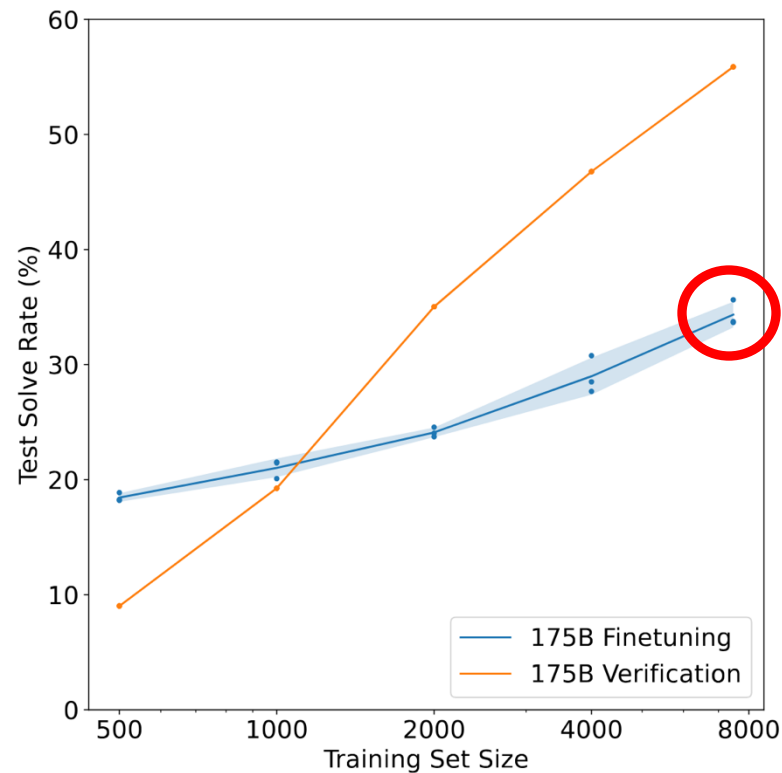
Calculator as a Tool: GSM8K

How to leverage calculator annotations during inference?



*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

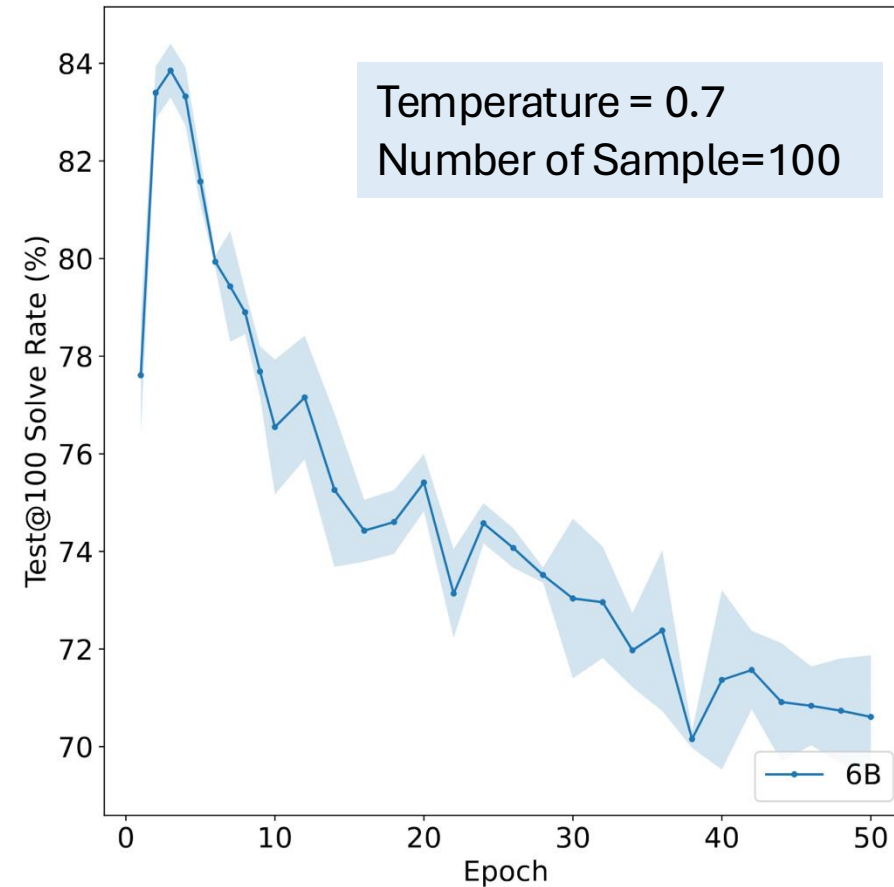
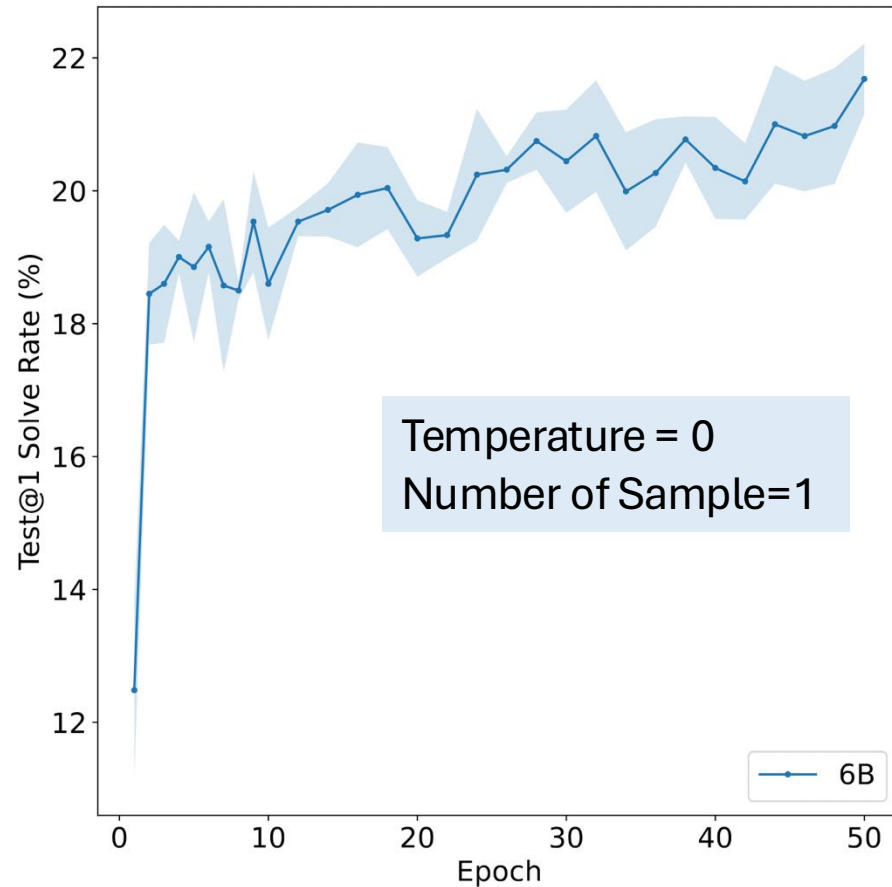
Calculator as a Tool: GSM8K



A comparison between finetuning and verification using 175B model. Verification considers 100 solutions per problem

*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

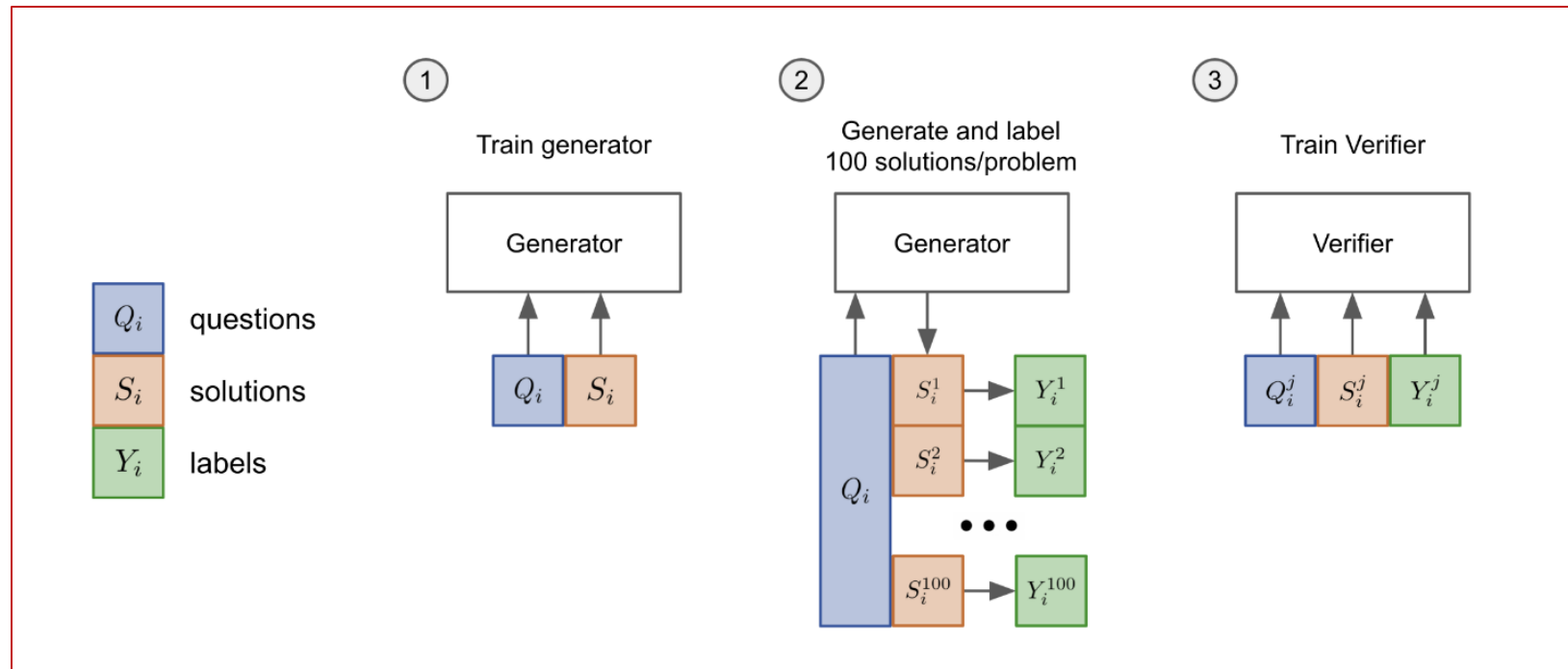
Calculator as a Tool: GSM8K



*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

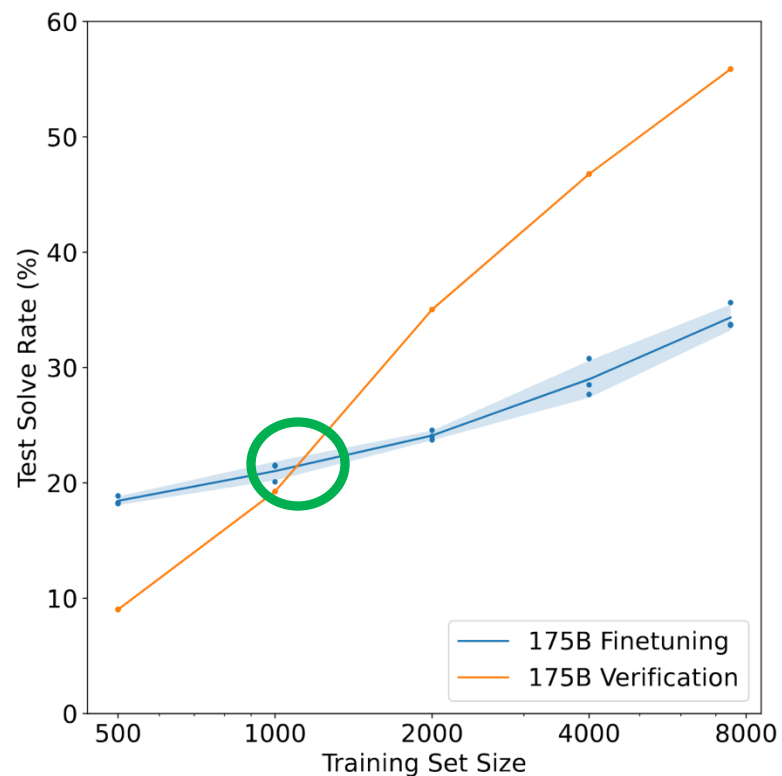
Calculator as a Tool: GSM8K

Verification training pipeline



*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

Calculator as a Tool: GSM8K



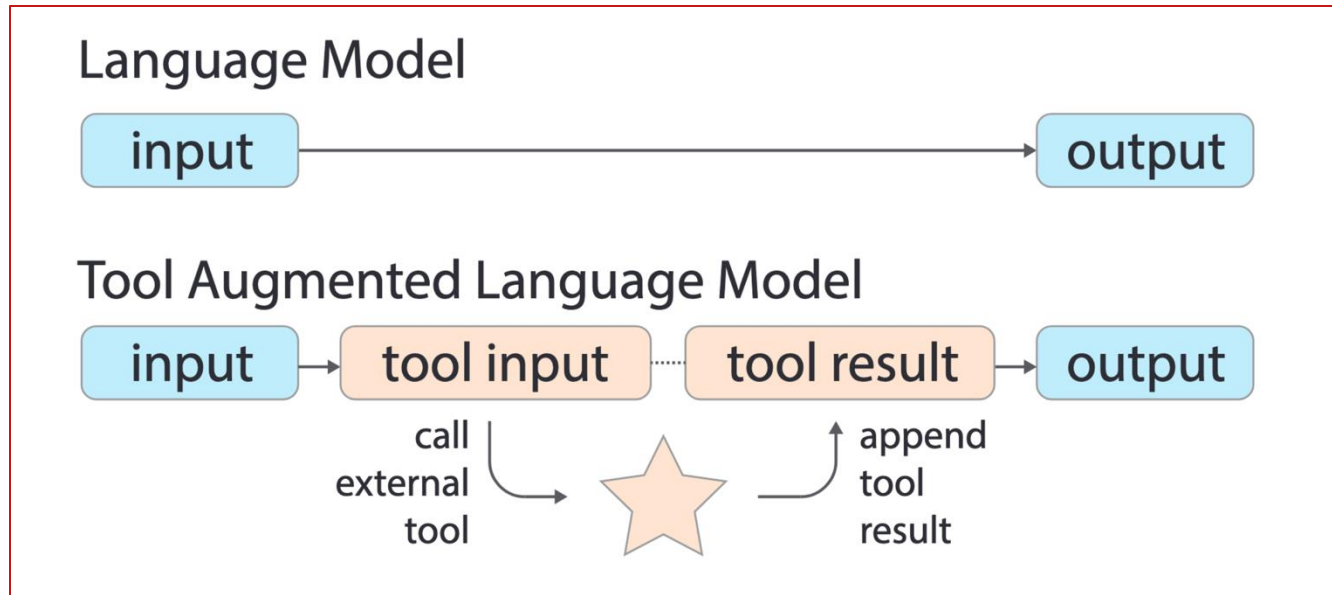
A comparison between finetuning and verification using 175B model. Verification considers 100 solutions per problem

*Training Verifiers to Solve Math Word Problems, Cobbe et al., Nov 2021

Summary

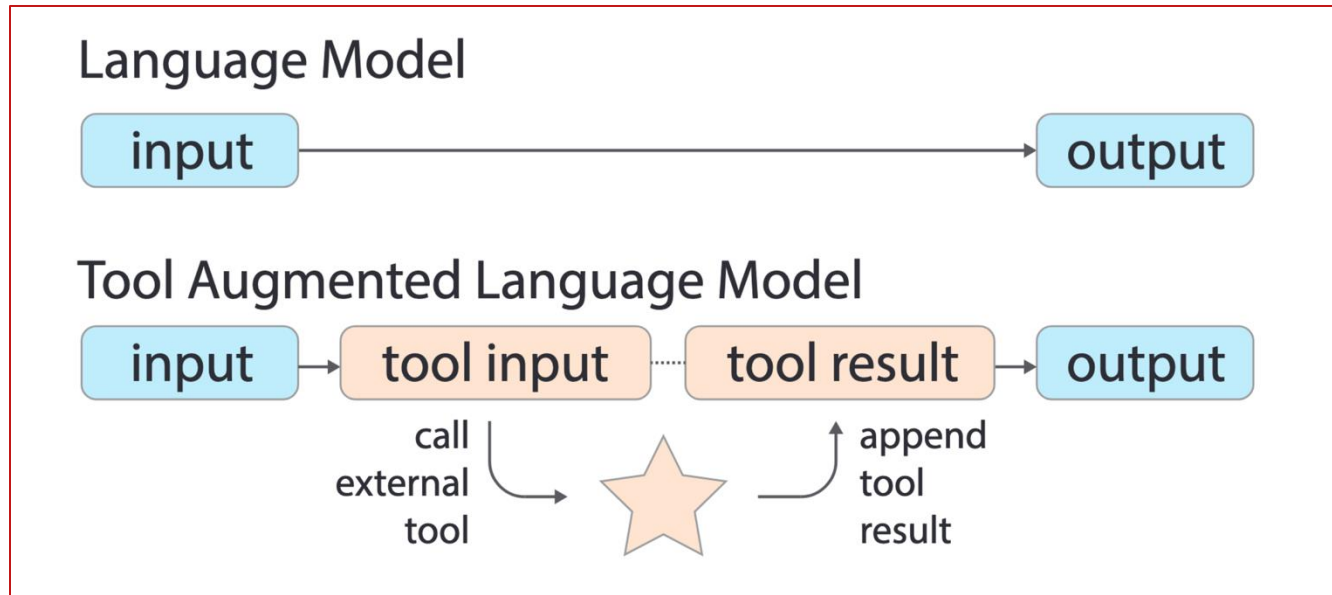
1. Training Verifiers to Solve Math Word Problems
 - Introduced tool augmentation in LLMs during finetuning
 - Use verifiers when the data is less

TALM: Tool Augmented Language Models*



*TALM: Tool Augmented Language Models, Parisi et al., May 2022

TALM: Tool Augmented Language Models*



An abstract task:

task input text | *tool-call* tool input text | *result* tool output text | *output* task output text

*TALM: Tool Augmented Language Models, Parisi et al., May 2022

TALM: Tool Augmented Language Models*

Natural Questions (NQ)

[Kwiatkowski et al., 2019]

|question when are hops added in brewing process?
|search brewing process |result The boiling process is
where chemical reactions take place...including |output
The boiling process.

|question If Lily's test scores are 85 , 88 and 95 out
of 100 in 3 different subjects , what will be her aver-
age score? |formula $\text{Divide}(\text{Add}(85, \text{Add}(88, 95)), 3)$
|result 89.333333333333 |output 89.33

MathQA

[Amini et al., 2019]

*TALM: Tool Augmented Language Models, Parisi et al., May 2022

TALM: Tool Augmented Language Models*

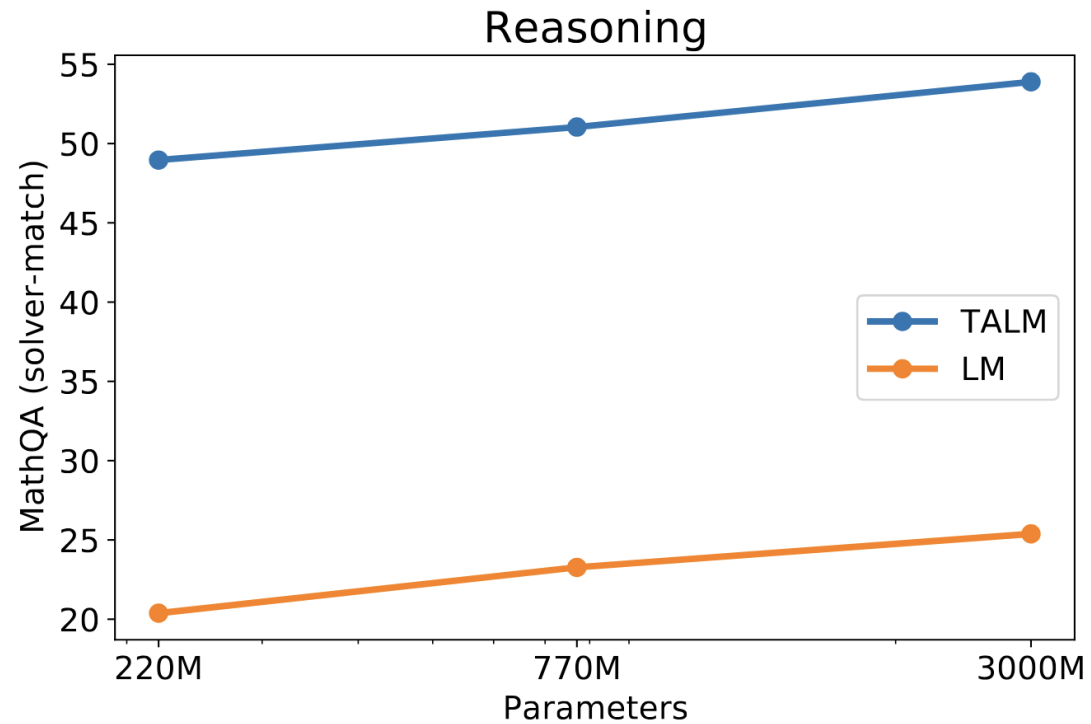
Algorithm 1 Iterative Self-Play Algorithm.

x : task input, y : task output, t : tool input, r : tool output

```
1:  $T = \{x_i, y_i\}_T$  # task set
2:  $D = \{x_j, t_j, r_j, y_j\}_D$  # tool-use set
3:  $P_\theta \leftarrow \text{pretrained LM}$ 
4: for  $t \in [0, 1, \dots, R]$  do # self-play rounds
5: # finetune LM
6:  $\theta \leftarrow \underset{\theta}{\operatorname{argmax}} \prod_D P_\theta(y_j|x_j, t_j, r_j) P_\theta(t_j|x_j)$ 
7: for  $x_i, y_i \in T$  do # iterate task set
8:   for  $n \in [0, 1, \dots, N]$  do
9:      $t_n \leftarrow P_\theta(t|x_i)$  # sample tool query
10:     $r_n \leftarrow \text{Tool}(t_n)$  # call tool API
11:     $y_n \leftarrow P_\theta(y|x_i, t_n, r_n)$  # get task output
12:    if  $|y_n - y_i| < th$  then # filter wrong output
13:       $D \leftarrow D \cup \{x_i, t_n, r_n, y_n\}_1$ 
14: # update tool-use set
```

*TALM: Tool Augmented Language Models, Parisi et al., May 2022

TALM: Tool Augmented Language Models*



*TALM: Tool Augmented Language Models, Parisi et al., May 2022

TALM: Tool Augmented Language Models*

Algorithm 1 Iterative Self-Play Algorithm.

x : task input, y : task output, t : tool input, r : tool output

```
1:  $T = \{x_i, y_i\}_T$  # task set
2:  $D = \{x_j, t_j, r_j, y_j\}_D$  # tool-use set
3:  $P_\theta \leftarrow \text{pretrained LM}$ 
4: for  $t \in [0, 1, \dots, R]$  do # self-play rounds
5: # finetune LM
6:  $\theta \leftarrow \underset{\theta}{\operatorname{argmax}} \prod_D P_\theta(y_j|x_j, t_j, r_j) P_\theta(t_j|x_j)$ 
7: for  $x_i, y_i \in T$  do # iterate task set
8:   for  $n \in [0, 1, \dots, N]$  do
9:      $t_n \leftarrow P_\theta(t|x_i)$  # sample tool query
10:     $r_n \leftarrow \text{Tool}(t_n)$  # call tool API
11:     $y_n \leftarrow P_\theta(y|x_i, t_n, r_n)$  # get task output
12:    if  $|y_n - y_i| < th$  then # filter wrong output
13:       $D \leftarrow D \cup \{x_i, t_n, r_n, y_n\}_1$ 
14: # update tool-use set
```

Limitation: Requires demonstrations

For NQ, 150 demonstrations were used

For MathQA, ~2K demonstrations were used

*TALM: Tool Augmented Language Models, Parisi et al., May 2022

Summary

1. Training Verifiers to Solve Math Word Problems
 - Introduced tool augmentation in LLMs during finetuning
 - Use verifiers when the data is less
2. TALM
 - Introduced a general framework
 - Cannot generalize and requires human annotations

PAL: 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?

Recap of Chain-of-Thought Prompting

*PAL: Program-aided Language Models, Gao et al., 2022

PAL: Program-aided Language Models*

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?

*PAL: Program-aided Language Models, Gao et al., 2022

PAL: Program-aided Language Models*

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
```



*PAL: Program-aided Language Models, Gao et al., 2022

PAL: Program-aided Language Models*

```
# Q: I have a chair, two potatoes, a cauliflower, a lettuce head, two tables, a  
cabbage, two onions, and three fridges. How many vegetables do I have?
```

```
# note: I'm not counting the chair, tables, or fridges
```

```
vegetables_to_count = {  
    'potato': 2,  
    'cauliflower': 1,  
    'lettuce head': 1,  
    'cabbage': 1,  
    'onion': 2  
}  
print(sum(vegetables_to_count.values()))
```

```
# Q: I have a drum, a flute, a clarinet, a violin, four accordions, a piano, a  
trombone, and a trumpet. How many musical instruments do I have?
```

```
musical_instruments_to_count = {  
    'drum': 1,  
    'flute': 1,  
    'clarinet': 1,  
    'violin': 1,  
    'accordion': 4,  
    'piano': 1,  
    'trombone': 1,  
    'trumpet': 1  
}  
print(sum(musical_instruments_to_count.values()))
```

*PAL: Program-aided Language Models, Gao et al., 2022

PAL: Program-aided Language Models*

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

Problem solve rate (%) on mathematical reasoning datasets.

*PAL: Program-aided Language Models, Gao et al., 2022

PAL: Program-aided Language Models*

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

Problem solve rate (%) on mathematical reasoning datasets.

*PAL: Program-aided Language Models, Gao et al., 2022

Summary

1. Training Verifiers to Solve Math Word Problems
 - Introduced tool augmentation in LLMs during finetuning
 - Use verifiers when the data is less
2. TALM
 - Introduced a general framework
 - Cannot generalize and requires human annotations
3. PAL
 - Requires fewer human annotations
 - Cannot generalize to unseen tools during test

Toolformer*

Image credits: <https://neurips.cc/media/neurips-2023/Slides/71288.pdf>

What other name
is Pittsburgh
known by?



The Steel City

War memorial
Flodden



[...] was created in
memory of the
Battle of Flodden.

3435 *
235 / 9



89691.67

Ø



Thursday,
March 10,
2019

Os Melhores
Escolas em
Jersey



The Best
Schools in
Jersey

*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

Toolformer*

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.

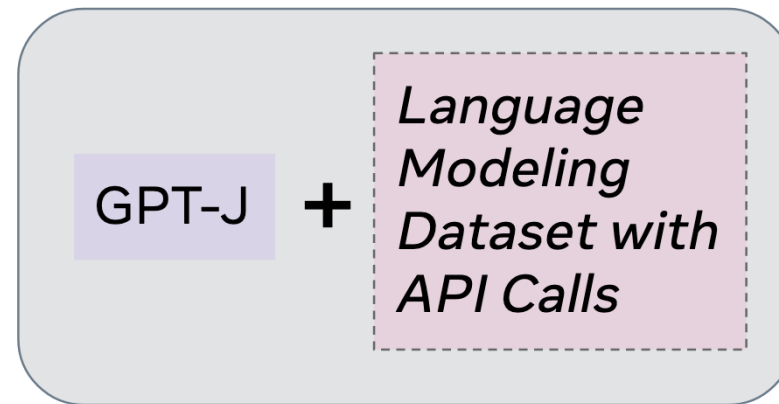
Toolformer autonomously decides

- which APIs to call
- when to call them
- what arguments to pass
- how to best incorporate the results into future token prediction

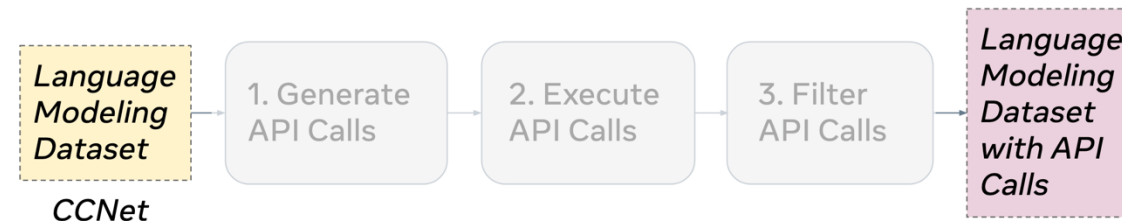
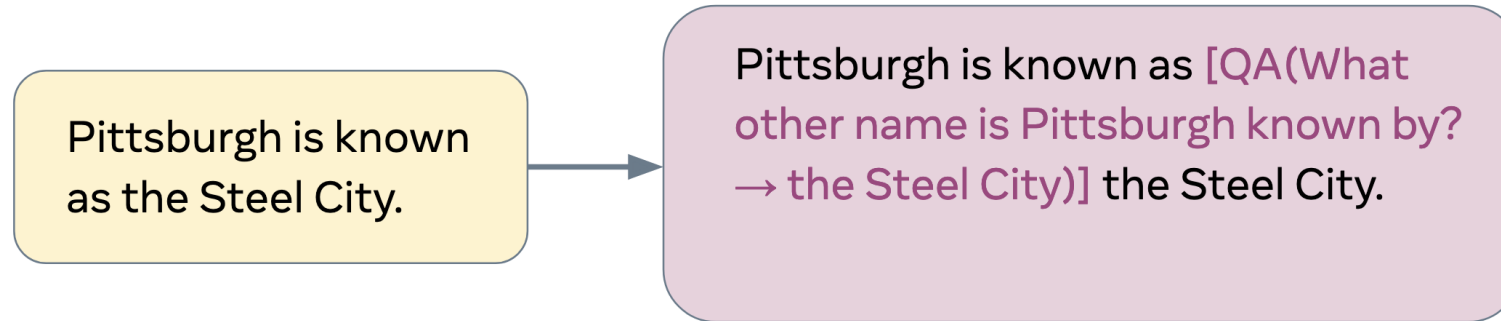
*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

1. Creating a new training dataset augmented with API calls
2. Finetune GPT-J using this new dataset

Toolformer



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

Toolformer*

Image credits: <https://neurips.cc/media/neurips-2023/Slides/71288.pdf>

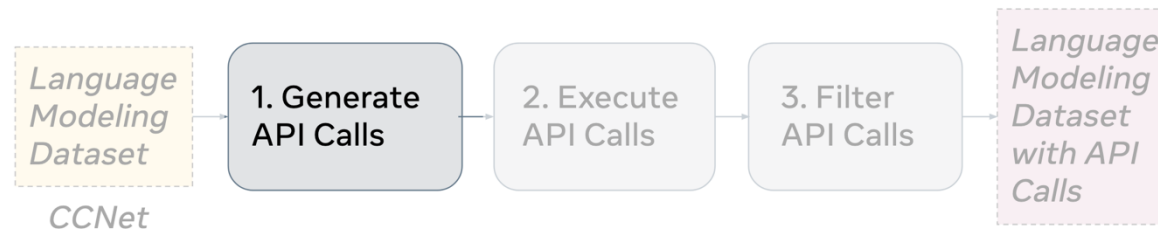
Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:

Input: Joe Biden was born in Scranton, Pennsylvania.

Output: Joe Biden was born in [QA("Where was Joe Biden born?")]
Scranton, [QA("In which state is Scranton?")] Pennsylvania.

Input: $\{\text{input}\}$

Output:



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

Your task is to add calls to a QA API to a piece of text [...]

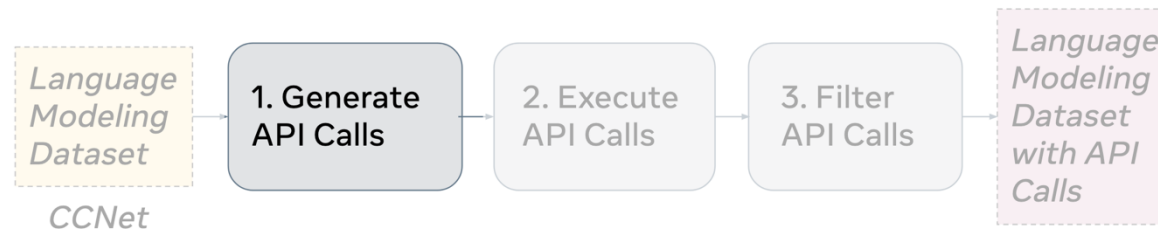
Input: Pittsburgh is known as the Steel City.

Output:

Pittsburgh is known as [QA("In which state is Pittsburgh?")] the Steel City.

Pittsburgh is known as [QA("What other name is Pittsburgh known by?")] the Steel City.

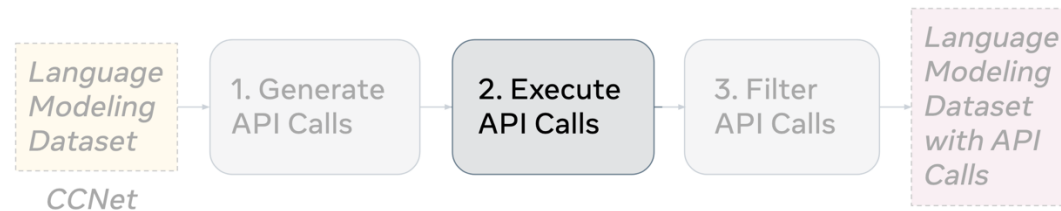
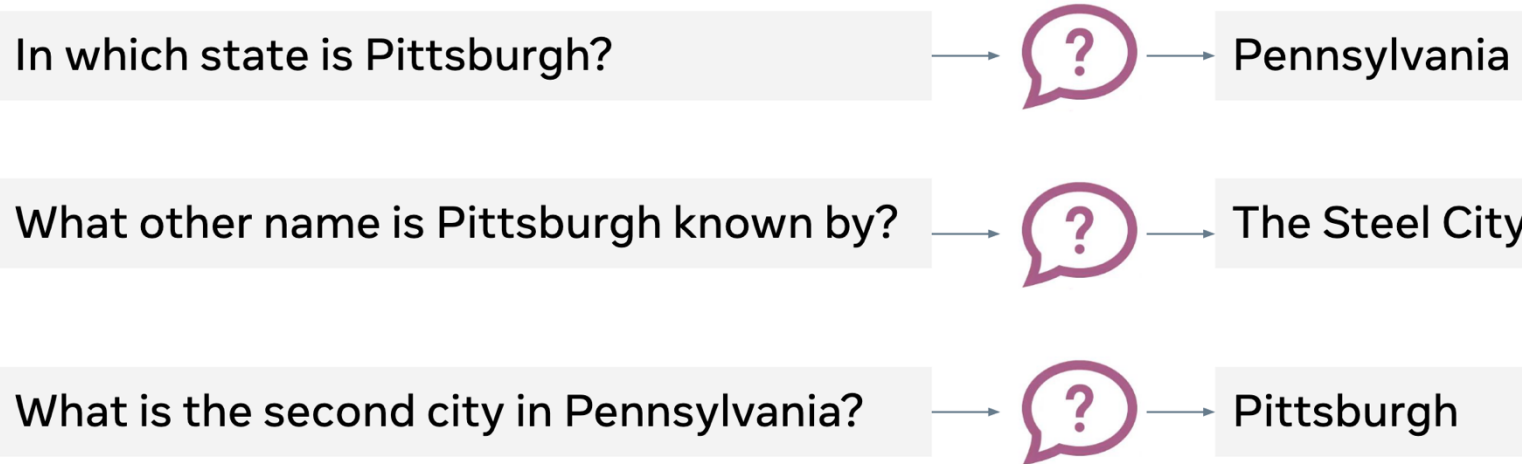
Pittsburgh is known as [QA("What is the second city in Pennsylvania?")] the Steel City.



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

Toolformer*

Image credits: <https://neurips.cc/media/neurips-2023/Slides/71288.pdf>



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

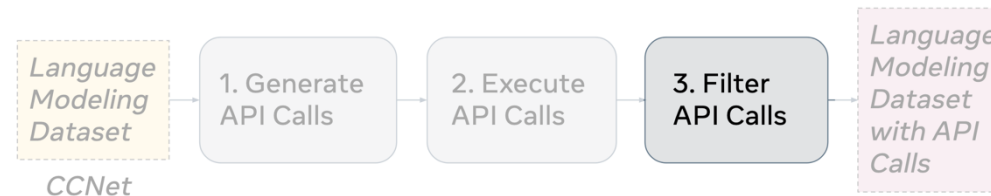
$$L_{\bullet}(PREFIX) = -\log p(\text{the Steel City.} \mid PREFIX)$$

A. No API Call $L_A(\text{Pittsburgh is known as}) = 2.5$

B. Non-executed API Call $L_B(\text{Pittsburgh is known as [QA("What other name is Pittsburgh known by?")} \rightarrow ?]) = 2.1$

C. Executed API Call $L_C(\text{Pittsburgh is known as [QA("What other name is Pittsburgh known by?")} \rightarrow \text{Steel City}]) = 0.8$

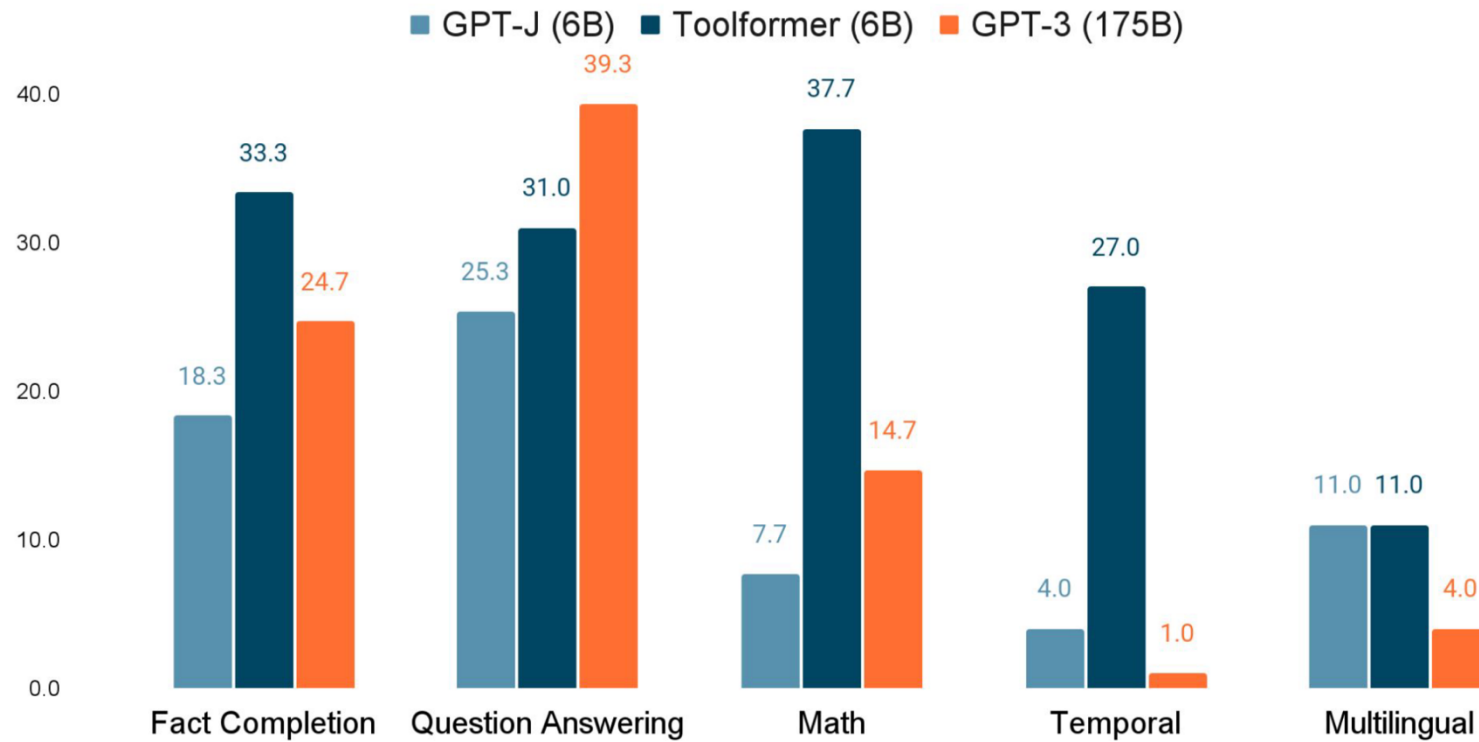
$$\text{Usefulness} = \min(L_A, L_B) - L_C = \min(2.5, 2.1) - 0.8 = 1.3$$



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

Toolformer*

Image credits: <https://neurips.cc/media/neurips-2023/Slides/71288.pdf>



*Toolformer: Language Models Can Teach Themselves to Use Tools, Schick et al., 2023

Summary

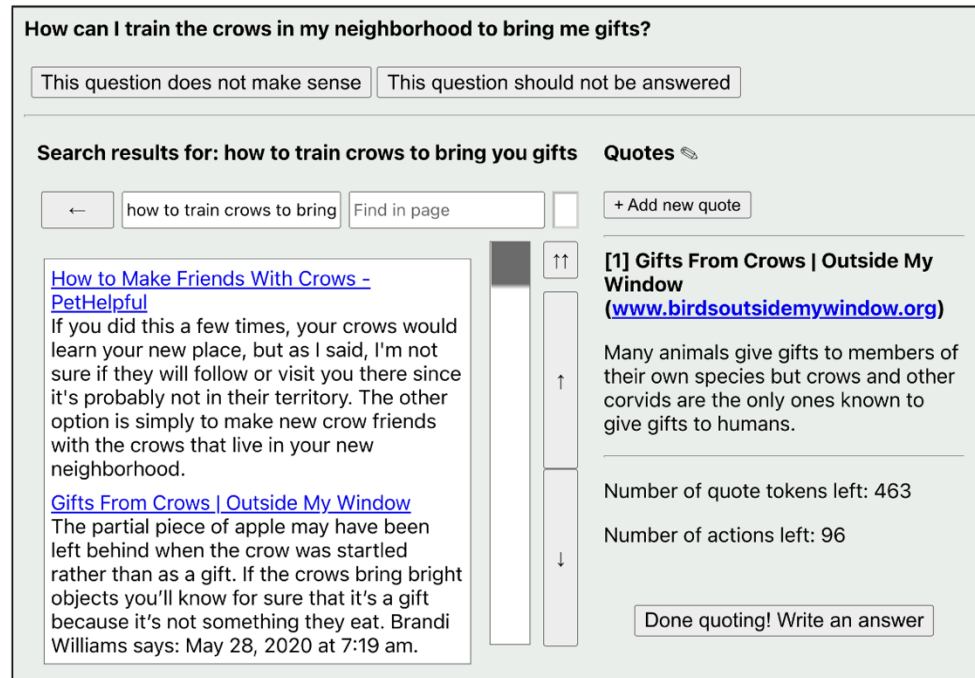
1. Training Verifiers to Solve Math Word Problems
 - Introduced tool augmentation in LLMs during finetuning
 - Use verifiers when the data is less
2. TALM
 - Introduced a general framework
 - Cannot generalize and requires human annotations
3. PAL
 - Requires fewer human annotations
 - Cannot generalize to unseen tools during test
4. Toolformer
 - Generalizes to multiple tasks
 - Requires fewer human annotations
 - Cannot generalize to unseen tools during test

Search Engine as a Tool: WebGPT*

Command	Effect
Search <query>	Send <query> to the Bing API and display a search results page
Clicked on link <link ID>	Follow the link with the given ID to a new page
Find in page: <text>	Find the next occurrence of <text> and scroll to it
Quote: <text>	If <text> is found in the current page, add it as a reference
Scrolled down <1, 2, 3>	Scroll down a number of times
Scrolled up <1, 2, 3>	Scroll up a number of times
Top	Scroll to the top of the page
Back	Go to the previous page
End: Answer	End browsing and move to answering phase
End: <Nonsense, Controversial>	End browsing and skip answering phase

*WebGPT: Browser-assisted question-answering with human feedback, Nakano et al., June 2022

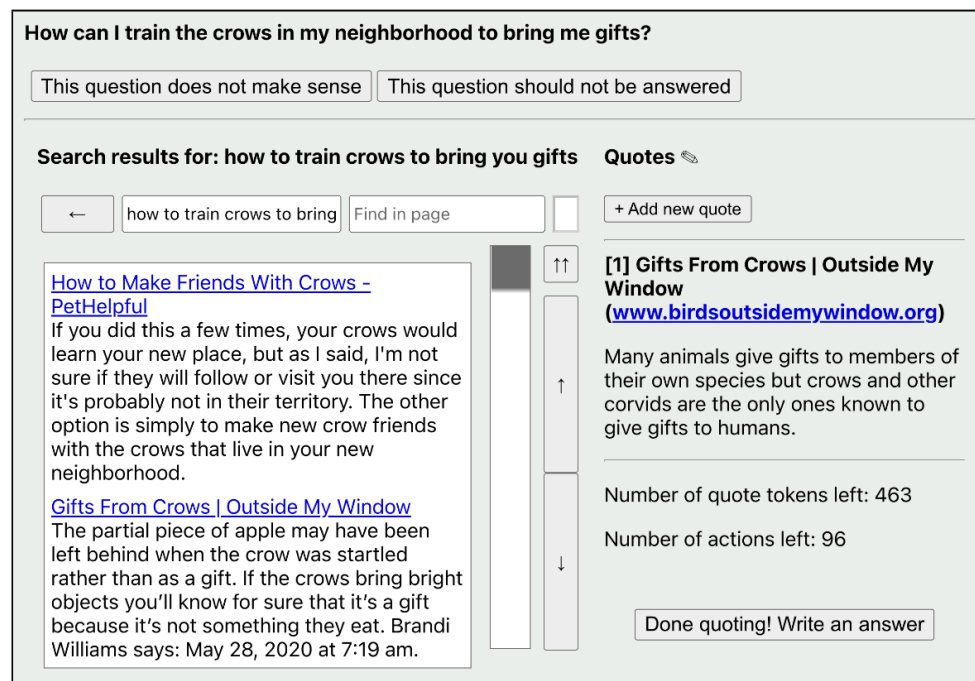
Search Engine as a Tool: WebGPT*



(a) Screenshot from the demonstration interface.

*WebGPT: Browser-assisted question-answering with human feedback, Nakano et al., June 2022

Search Engine as a Tool: WebGPT*



(a) Screenshot from the demonstration interface.

◆Question
How can I train the crows in my neighborhood to bring me gifts?

◆Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemymywindow.org)
> Many animals give gifts to members of their own species but crows and other corvids are the only ones known to give gifts to humans.

◆Past actions
Search how to train crows to bring you gifts
Click Gifts From Crows | Outside My Window www.birdsoutsidemymywindow.org
Quote
Back

◆Title
Search results for: how to train crows to bring you gifts

◆Scrollbar: 0 - 11
◆Text
[0]How to Make Friends With Crows - PetHelpful[pethelpful.com]
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.
[1]Gifts From Crows | Outside My Window[www.birdsoutsidemymywindow.org]
The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat. Brandi Williams says: May 28, 2020 at 7:19 am.

◆Actions left: 96
◆Next action

(b) Corresponding text given to the model.

*WebGPT: Browser-assisted question-answering with human feedback, Nakano et al., June 2022

Search Engine as a Tool: WebGPT*

[WebGPT Answer Viewer](#)

*WebGPT: Browser-assisted question-answering with human feedback, Nakano et al., June 2022

Summary

1. Training Verifiers to Solve Math Word Problems

- Introduced tool augmentation in LLMs during finetuning
- Use verifiers when the data is less

2. TALM

- Introduced a general framework
- Cannot generalize and requires human annotations

3. PAL

- Requires fewer human annotations
- Cannot generalize the tool use to similar tasks

4. Toolformer

- Generalizes to multiple tasks
- Requires fewer human annotations
- Cannot generalize to unseen tools during test

5. WebGPT

- Specific for web search as a tool