Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì,
Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer,
Nicola Cancedda, Thomas Scialom

Meta

ChatGPT Limitations

Who is the current President of the United States?

As an AI language model, I don't have real-time data or internet access, and my knowledge was last updated in September 2021. [...]

What day of the week is it today?

As an AI language model, I do not have the capability to access current time or date information. [...]

What is the result of 3435 * 235 / 9?

This is approximately 89,937.22.

True answer: 89691.66

Toolformer Tools/APIs

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

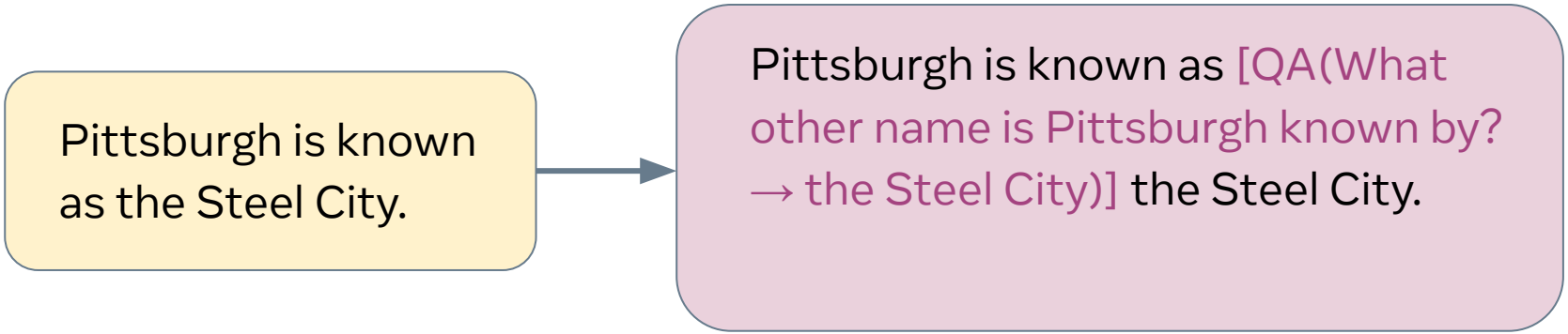
Thursday, March 10, 2019

Os Melhores Escolas em Jersey

The Best Schools in Jersey

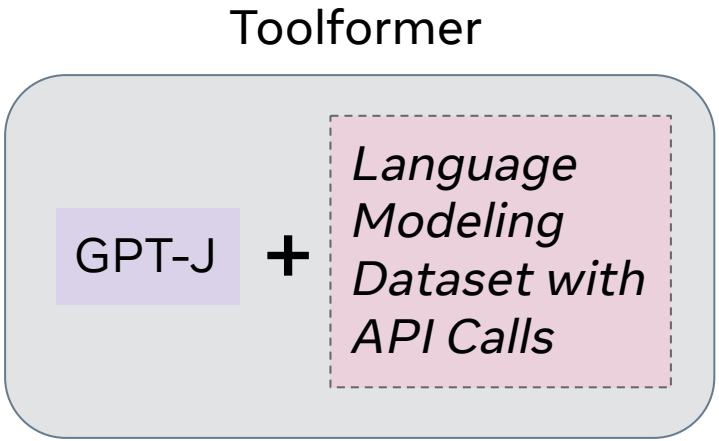
We want to train a model to learn…

- Which tool to use
- When to use the tool
- How to use the tool

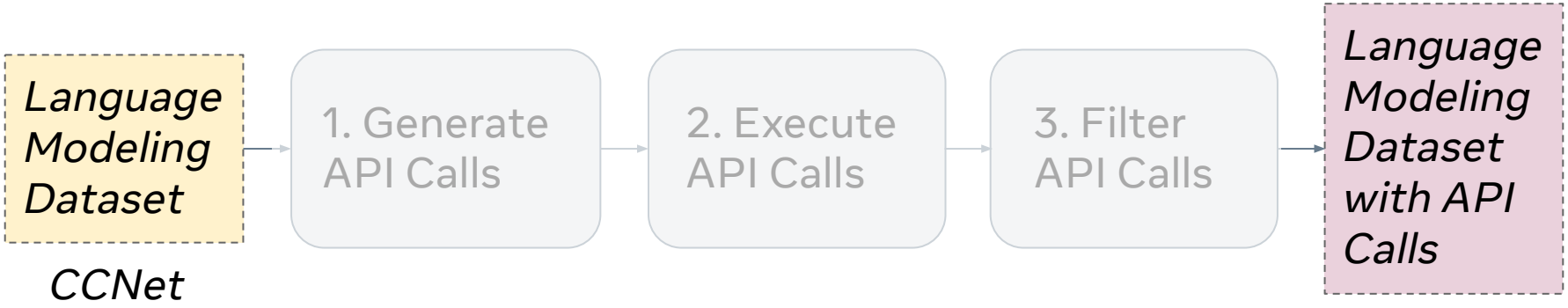


Steps to Creating Toolformer

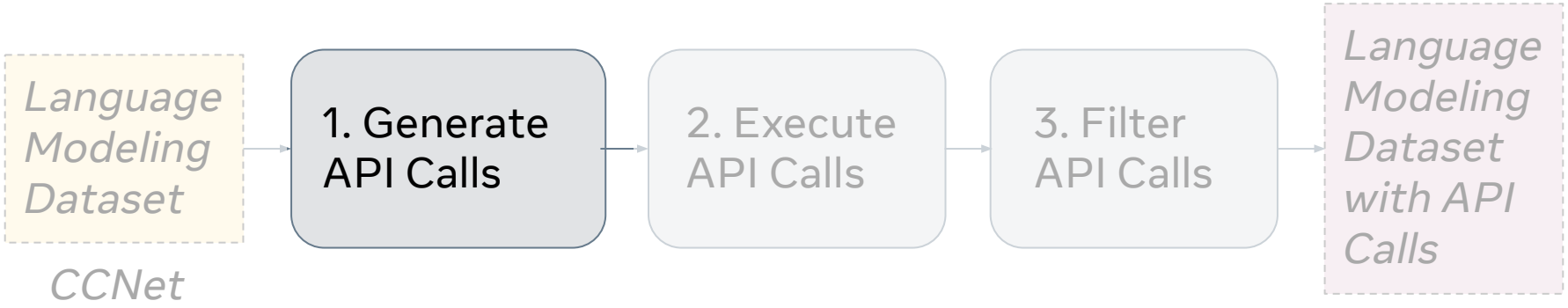
1. Creating a new training dataset augmented with API calls
2. Finetune GPT-J using this new dataset



Creating the Training Dataset



Creating the Training Dataset



Prompting the Model to Generate API Calls

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: *${input}*
Output:

Examples of Generated API Calls

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

Input: Pittsburgh is known as the Steel City.

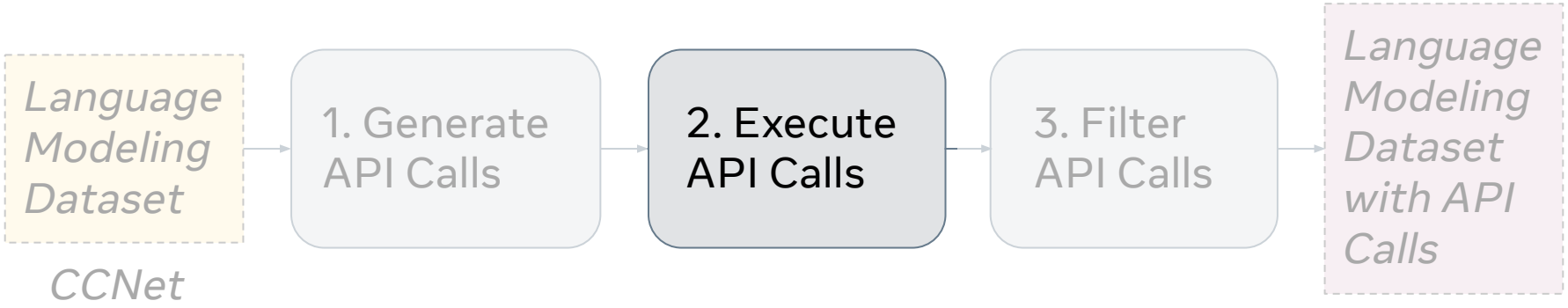
Output:

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.

Creating the Training Dataset



Execute the API Calls

In which state is Pittsburgh? →  → Pennsylvania

What other name is Pittsburgh known by? →  → The Steel City

What is the second city in Pennsylvania? →  → Pittsburgh

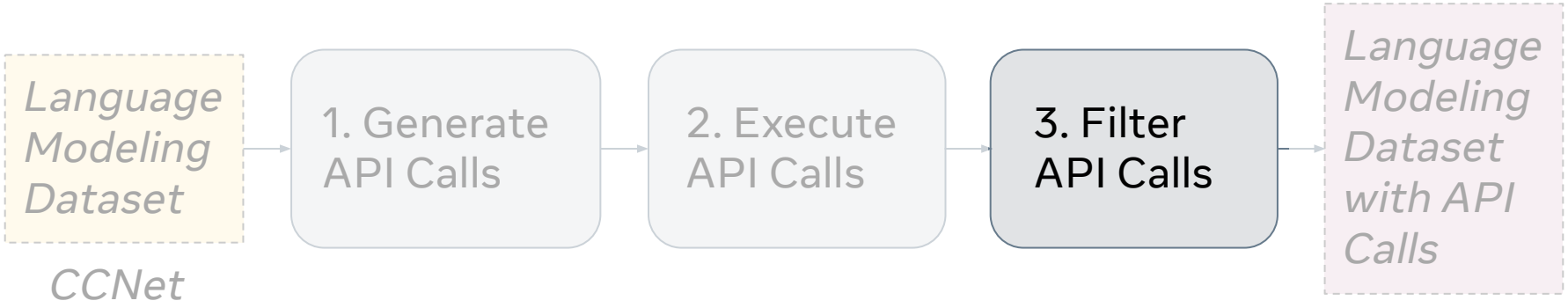
Execute the API Calls

In which state is Pittsburgh? →  → Pennsylvania

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

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

Creating the Training Dataset



Filter the API Calls using Model-based Perplexity

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

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

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

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

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

Usefulness Examples

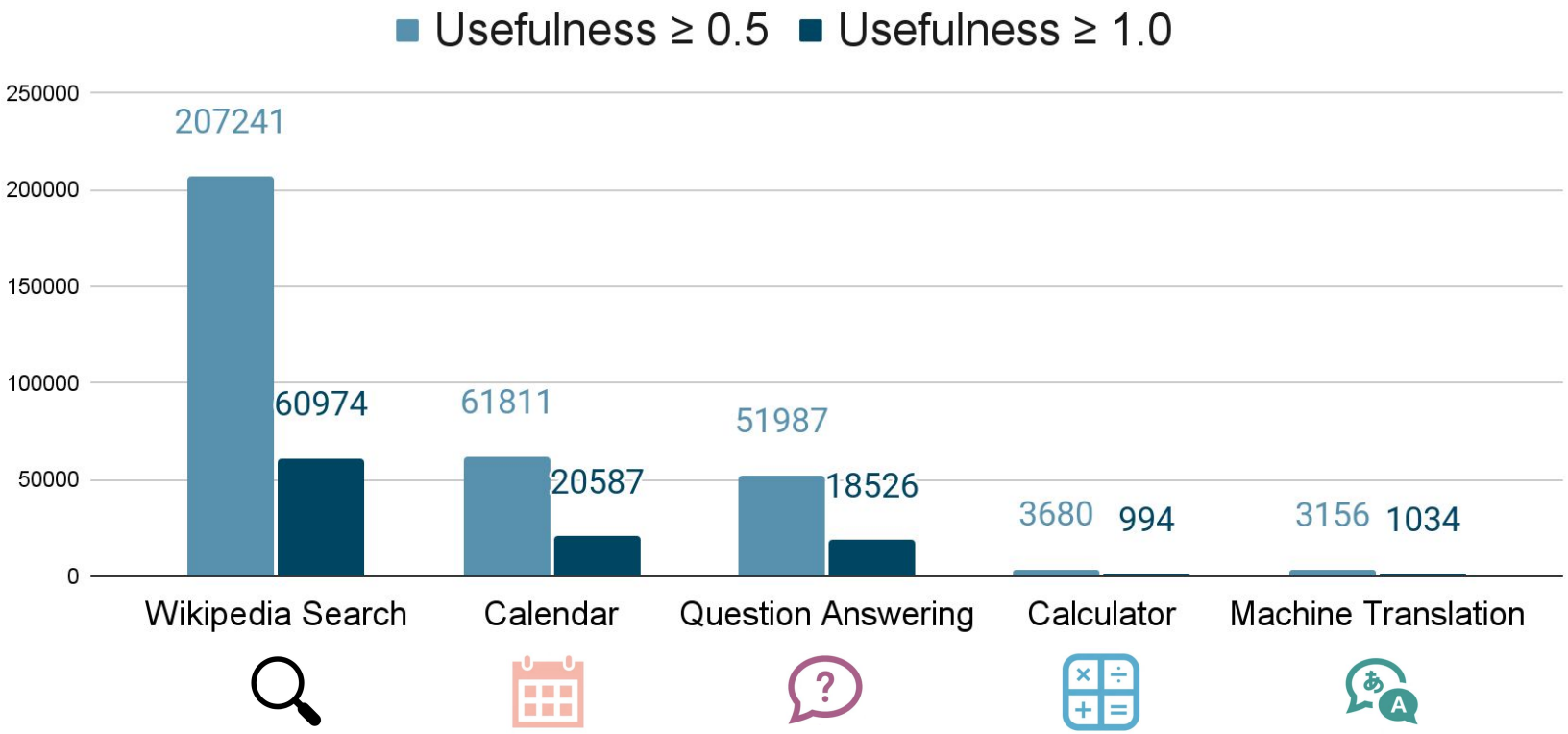
2.11

The WL will be open on Friday, [Calendar() → Today is Thursday, March 9, 2017.] March 10, and Sunday, March 19 for regular hours.

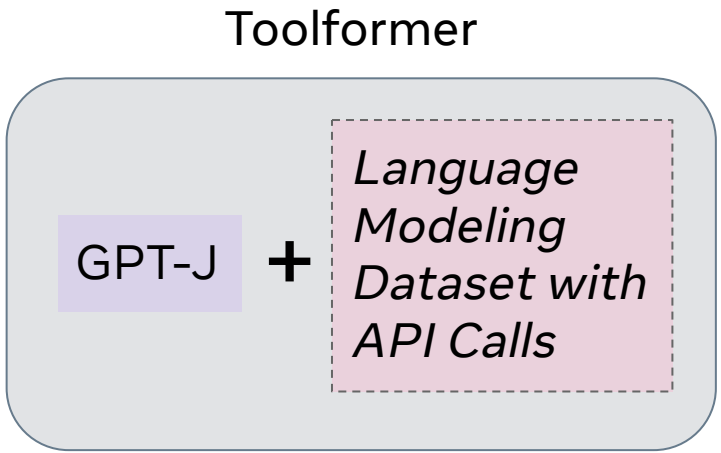
-0.02

85 patients (23%) were hospitalised alive and admitted to a hospital ward. Of them, [Calculator(85 / 23) → 3.70] 65% had a cardiac aetiology.

Number of Tool-Usage Samples After Filtering



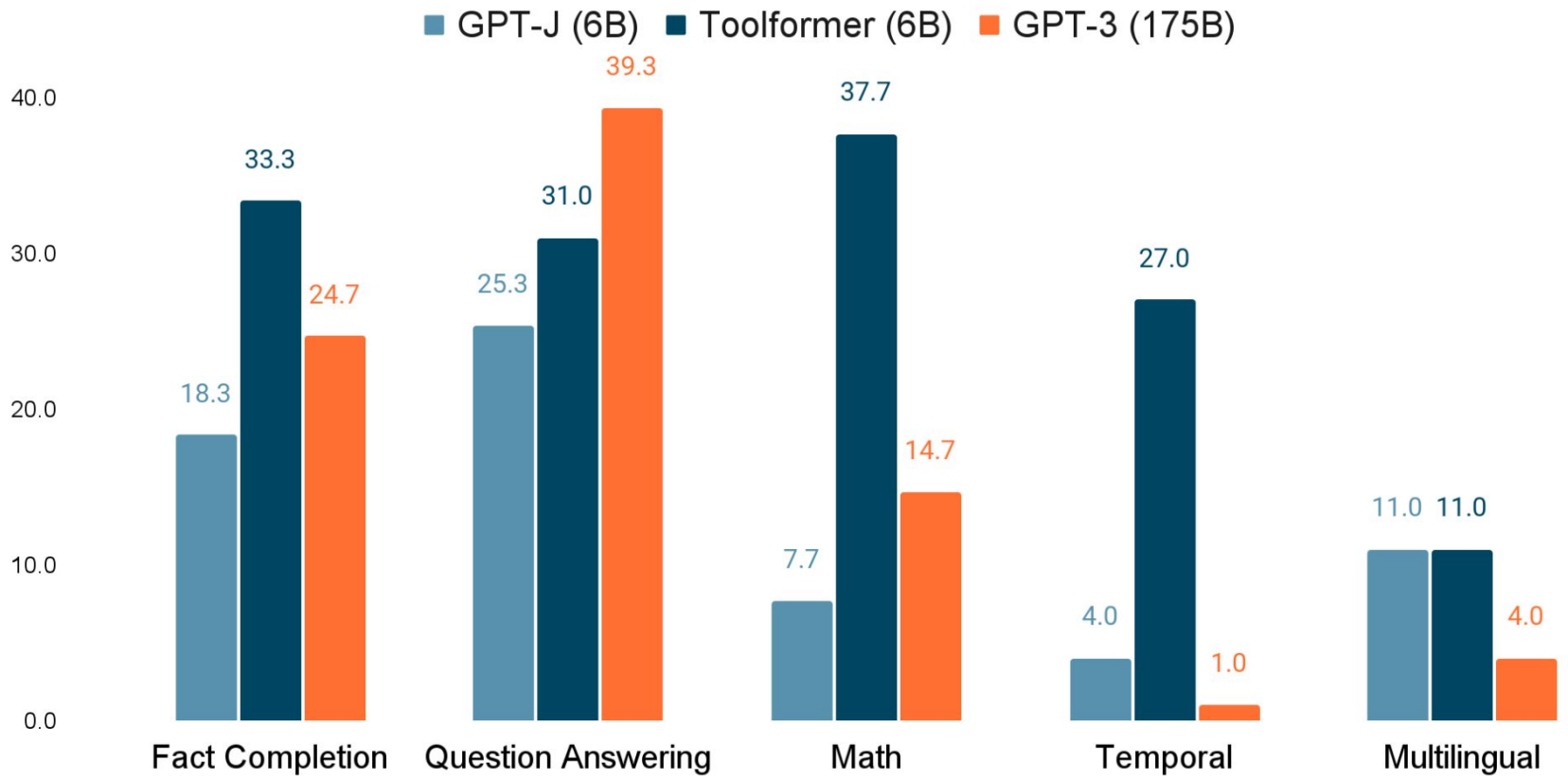
Fine-tuning Toolformer



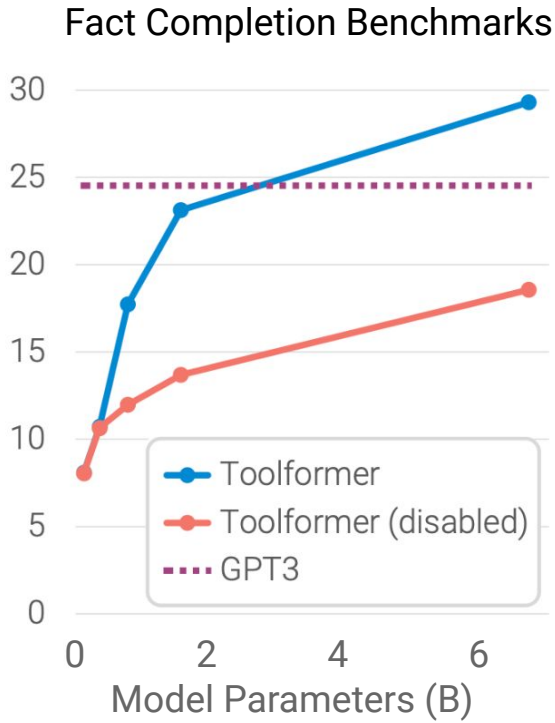
Evaluation Tasks

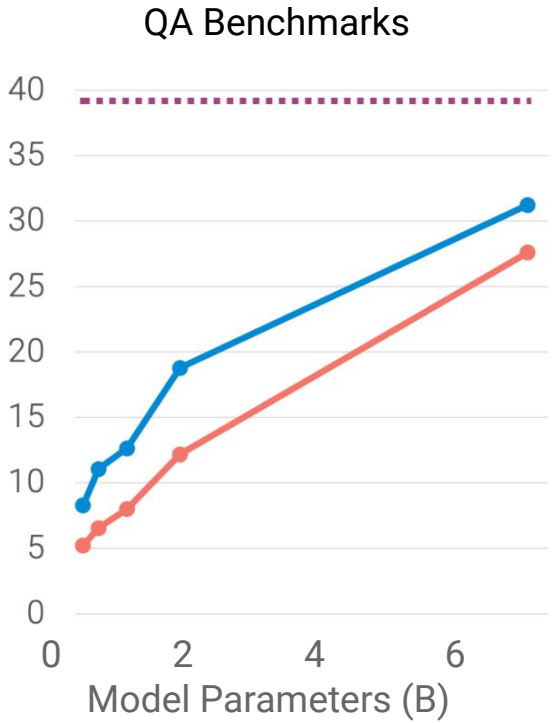
1. Fact Completion/Question Answering
    a. “The theory of relativity was developed by _____”
    b. “In Greek Mythology, who is the goddess of spring growth?”
2. Math Computations
3. Multilingual Questions
    a. Context is given in English, question is multilingual.
4. Temporal Questions
    a. How many days is it until Christmas?

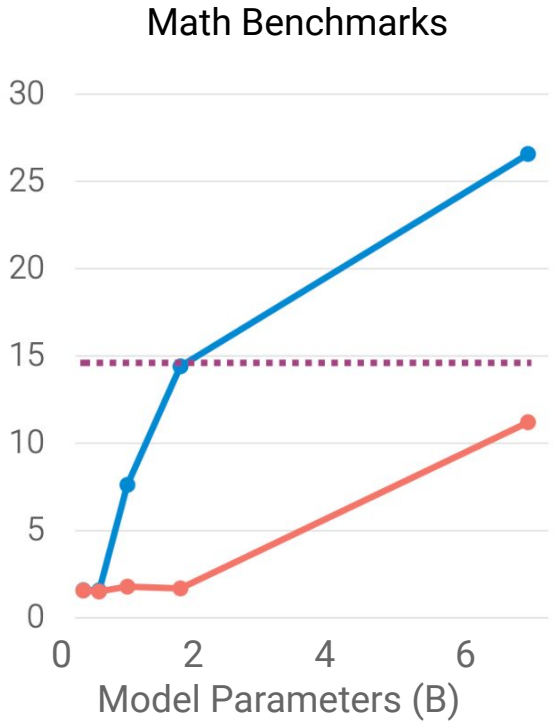
Results


GPT-J (6B)
Toolformer (6B)
GPT-3 (175B)
40.0
30.0
20.0
10.0
0.0
18.3
33.3
24.7
25.3
31.0
39.3
7.7
37.7
14.7
4.0
27.0
1.0
11.0
11.0
4.0
Fact Completion
Question Answering
Math
Temporal
Multilingual

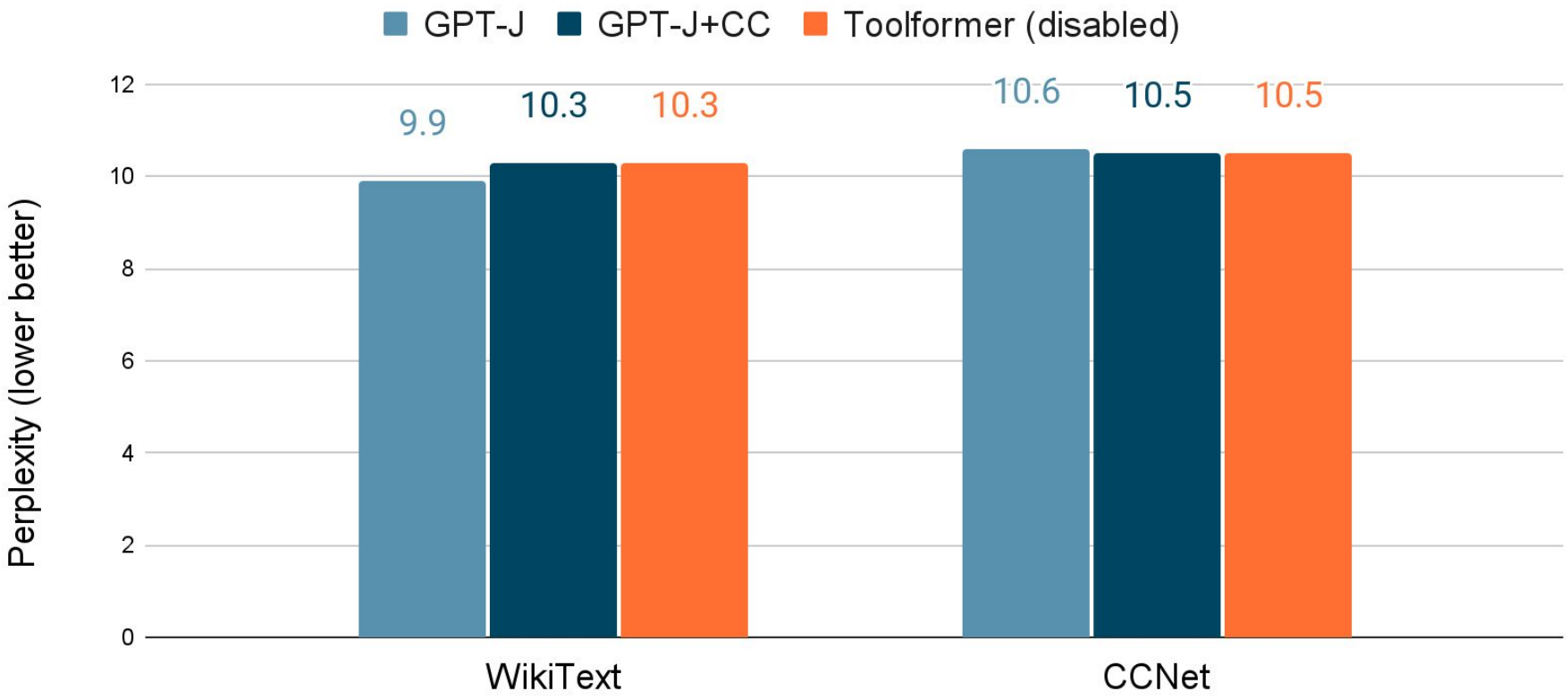
Can small models effectively use tools?







Is Toolformer still a good language model?



Toolformer: Language Models Can Teach Themselves to Use Tools

Poster Session 3
332


Timo Schick

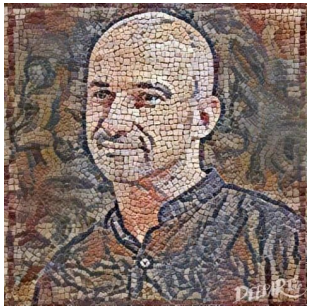

Jane Dwivedi-Yu
janeyu@meta.com


Roberto Dessi


Roberta Raileanu


Maria Lomeli


Luke Zettlemoyer


Nicola Cancedda


Thomas Scialom