Toolformer -
How to equip LLMs with tools

Dominik Lewy

Information Sensitivity: General\External

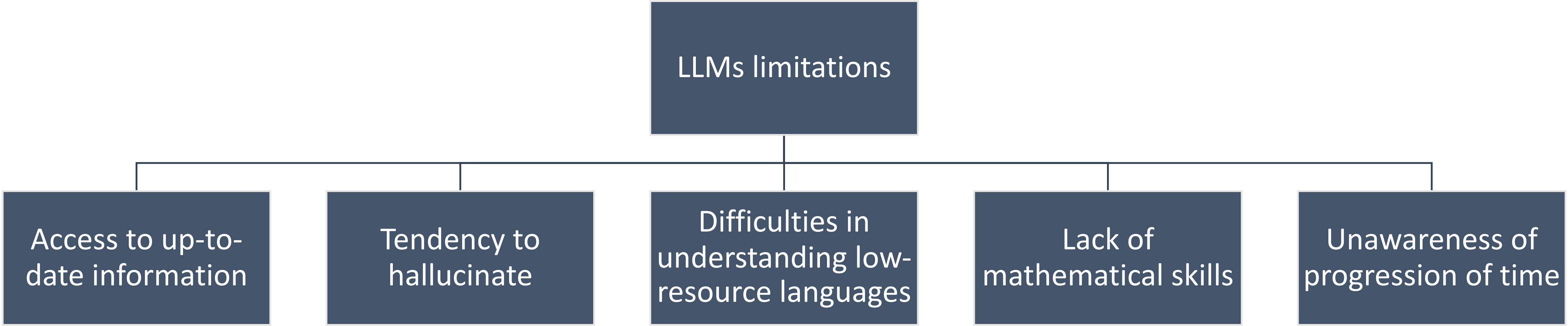
Agenda

1. What LLMs are good at and what they fail at?
2. Toolformer
3. In-context learning
4. Creation of dataset for Toolformer training
5. Tools
6. Downstream tasks
7. Other (preserving LLM core abilities, scaling, limitations)

Information Sensitivity: General\External

What LLMs are good at and what they fail at?

What LLMs are good at and what they fail at?



Information Sensitivity: General\External

What LLMs are good at and what they fail at?

LLM knowledge:

DO What is the flag of Burkina Faso?

The flag of Burkina Faso consists of two horizontal bands of red and green, with a yellow five-pointed star in the center. The red color represents the revolutionary struggle for independence, the green symbolizes the country's rich natural resources and hope for a bright future, and the yellow star represents the guiding light of the revolution. This flag was adopted when Burkina Faso, formerly known as Upper Volta, gained its independence from France in 1984 and changed its name. The flag's design and colors reflect the nation's history and aspirations.

From Wikipedia, the free encyclopedia

The national **flag of Burkina Faso** (French: *drapeau du Burkina Faso*) is formed by two equal horizontal bands of red (top) and green, with a yellow five-pointed star resting in the center.[1] The flag was adopted o 4 August 1984. The flag uses the Pan-African colours of Ethiopia, reflecting both a break with the country's colonial past and its unity with other African ex-colonies. The red is also said to symbolize the revolution an the green the abundance of agricultural and natural riches. The yellow star placed over the red and green stripes represents the guiding light of the revolution. The flag was adopted following the coup of 1983 whicl brought Thomas Sankara to power.[2]

LLM simple math:

2+2*2 how much is it?

Sure! 2 + 2 * 2 = 2 + 4 = 6.

($1,452,190 x 2.6 x 5/12) / 100 how much is this?

To perform the calculation, we need to follow the order of operations (PEMDAS):

1. Multiply $1,452,190 by 2.6: $1,452,190 x 2.6 = $3,735,834
2. Divide $3,735,834 by 12: $3,735,834 ÷ 12 = $311,320
3. Divide $311,320 by 100: $311,320 ÷ 100 = $3,113.20

Therefore, ($1,452,190 x 2.6 x 5/12) / 100 = $3,113.20.

Information Sensitivity: General\External

What LLMs are good at and what they fail at?

DO

List research that adapts mixup data augmentation to NLP and asses the importance based on conference impact factor

Mixup is a data augmentation technique that creates synthetic training examples by linearly interpolating between two examples and their corresponding labels. Here are a few research papers that have adapted mixup for natural language processing (NLP) tasks:

1. "MixUp: A Holistic Approach to Semi-Supervised Learning" by Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. This paper, published at the International Conference on Learning Representations (ICLR) in 2018, proposes the mixup approach and demonstrates its effectiveness on a variety of NLP tasks, including text classification, machine translation, and language modeling. This conference has an impact factor of 18.8.
2. "MixText: Improved Text Classification with Mixup Augmentation" by Xingxing Zhang, Furu Wei, and Ming Zhou. This paper, published at the Conference on Empirical Methods in Natural Language Processing (EMNLP) in 2019, applies mixup to text classification and shows that it can improve performance on multiple datasets. The EMNLP conference has an impact factor of 10.4.
3. "MixUp Augmentation for Neural Machine Translation" by Zhen Yang, Yingce Xia, Yvette Graham, and Qun Liu. This paper, published at the Conference of the Association for Computational Linguistics (ACL) in 2019, investigates the use of mixup for neural machine translation and finds that it can improve translation quality and reduce overfitting. The ACL conference has an impact factor of 10.3.

Information Sensitivity: General\External

What LLMs are good at and what they fail at?

LLM English: ✓

DO Which club messi plays for?

As of my last knowledge update in January 2022, Lionel Messi was playing for Paris Saint-Germain (PSG), a top-tier football club in France. Messi transferred to PSG in August 2021 after spending his entire professional career with Barcelona. Please note that player transfers can occur, and his current club may have changed since then. I recommend checking the latest sports news or official sources to verify Messi's current club as of November 2023.

LLM Suahili: ✓ ✗

DO Answer in English to the following question: messi anachezea klabu gani?

Lionel Messi plays for Paris Saint-Germain (PSG) as of my last knowledge update in January 2022.

Information Sensitivity: General\External

What LLMs are good at and what they fail at?

DO

Which party rules in Poland?

As of my last knowledge update in January 2022, the Law and Justice Party (Prawo i Sprawiedliwość, often abbreviated as PiS) was in power in Poland. PiS is a right-wing political party in Poland, and it was in control of the government at that time. However, please note that political situations can change, and I do not have access to real-time information. I recommend checking the latest news or official government sources for the most up-to-date information on the current ruling party in Poland.

Information Sensitivity: General\External

Toolformer

Toolformer

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 AI Research [†]Universitat Pompeu Fabra

Information Sensitivity: General\External

Toolformer

Definition:

A model trained to decide which API to call, when to call it and how to use the result in future token prediction.

Information Sensitivity: General\External

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.

Figure 1: Exemplary predictions of Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is useful for completing a piece of text.

Assumptions:

- Tools should be learned in self-supervised way (cost, usefulness)
- No loss in generality and independence in API usage decision making (not tied to specific task)

Information Sensitivity: General\External

In-context learning

In-context learning

In-context learning (ICL) is a specific **method of prompt engineering where demonstrations of the task are provided to the model as part of the prompt (in natural language).** With ICL, you can use off-the-shelf large language models (LLMs) to solve novel tasks without the need for fine-tuning. ICL can also be combined with fine-tuning for more powerful LLMs.

The main types of machine learning (supervised ML, unsupervised ML, semi-supervised ML, and reinforcement learning) can only learn with data they are trained on. That is, they can only solve tasks that they are trained to solve. **LLMs that are large enough have shown a new type of machine learning - in-context learning - the ability to learn to solve new tasks by providing "training" examples in the prompt.** In contrast to the aforementioned types of ML, the newly **learnt skill is forgotten directly after the LLM sends its response - model weights are not updated**.

In-context learning (ICL) learns a new task from a small set of examples presented within the context (the prompt) at inference time. LLMs trained on sufficient data exhibit ICL, even though they are trained only with the objective of next token prediction. Much of the interest in LLMs is due to the prompting with examples as it enables applications on novel tasks without the need for fine-tuning the LLM.

Source: https://www.hopsworks.ai/dictionary/in-context-learning-icl

Information Sensitivity: General\External

In-context learning

Standard prompt:

```
prompt = "You are an AI assistant
that helps people translate text.
Translate the below text from
{sourceLanguage} to
{outputLanguage}.
{textForTranslation}
```

In-context propmpt:

```
prompt = "You are an AI assistant that
helps people translate text. Please
translate text provided between ###
from {sourceLanguage} to
{outputLanguage}.
\n ### {textForTranslation} ###
\n For reference you can use below
examples of similar translation that
were made in the past, which should
help you translate text correctly.
\n Text: {originalText1}
\n Translation: {translatedText1}.
\n Text: {originalText2}
\n Translation {translatedText2}."
```

Creation of dataset for Toolformer training

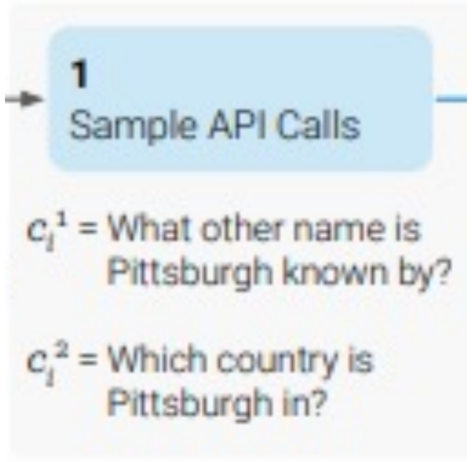
Information Sensitivity: General\External

Creation of dataset for Toolformer training

Figure 2: Key steps in our approach, illustrated for a *question answering* tool: Given an input text $\mathbf{x}$, we first sample a position i and corresponding API call candidates $c_i^1, c_i^2, \ldots, c_i^k$. We then execute these API calls and filter out all calls which do not reduce the loss L_i over the next tokens. All remaining API calls are interleaved with the original text, resulting in a new text $\mathbf{x}^*$.

Information Sensitivity: General\External

Creation of dataset for Toolformer training



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: Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

Output: Coca-Cola, or [QA("What other name is Coca-Cola known by?")] Coke, is a carbonated soft drink manufactured by [QA("Who manufactures Coca-Cola?")] the Coca-Cola Company.

Input: x

Output:

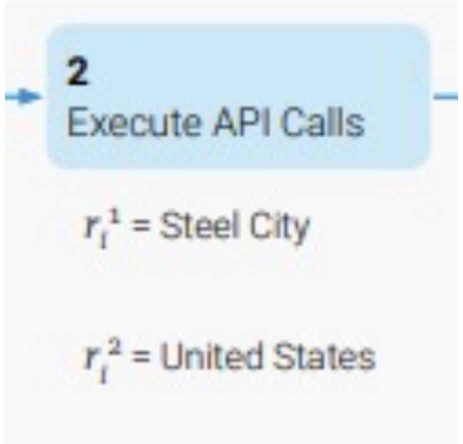
Figure 3: An exemplary prompt $P(\mathbf{x})$ used to generate API calls for the question answering tool.

Comments:

- Human involvement limited to examples from prompt
- Where to introduce API call is decided by model (up to *k* candidates)
- How many calls are generated for each position is a parameter (up to *m*)

Information Sensitivity: General\External

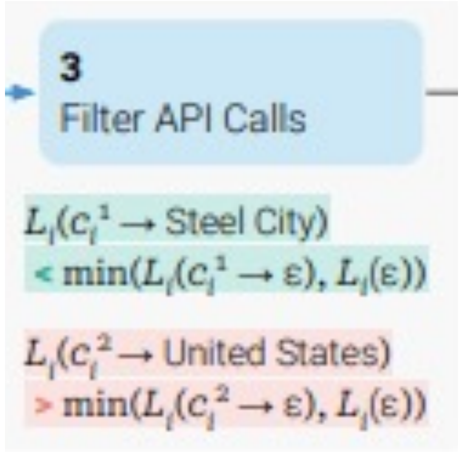
Creation of dataset for Toolformer training



Comments:

- Execution depends on the API, it can be calling NN, search engine or python tool
- Response needs to be a single text sequence

Creation of dataset for Toolformer training



Filtering API Calls Let i be the position of the API call c_i in the sequence $\mathbf{x} = x_1, \ldots, x_n$, and let r_i be the response from the API. Further, given a sequence $(w_i \mid i \in \mathbb{N})$ of *weights*, let

$$L_i(\mathbf{z}) = -\sum_{j=i}^{n} w_{j-i} \cdot \log p_M(x_j \mid \mathbf{z}, x_{1:j-1})$$

be the weighted cross entropy loss for M over the tokens $x_i, \ldots, x_n$ if the model is prefixed with $\mathbf{z}$. We compare two different instantiations of this loss:

$$L_i^+ = L_i(\mathrm{e}(c_i, r_i))$$
$$L_i^- = \min\left(L_i(\varepsilon), L_i(\mathrm{e}(c_i, \varepsilon))\right)$$

where ε denotes an empty sequence. The former is the weighted loss over all tokens $x_i, \ldots, x_n$ if the API call and its result are given to M as a prefix;[3] the latter is the minimum of the losses obtained from (i) doing no API call at all and (ii) doing an API call, but not providing the response. Intuitively,

Comments:

- We want to measure if API call with response helps the model reduce uncertainty
- Only calls that reduce it are used to create new data set

Information Sensitivity: General\External

Creation of dataset for Toolformer training

Example	$L_i^- - L_i^+$	Useful
The Flodden Window (a war memorial dedicated to The Middleton Archers), in the Grade I-listed Church of St Leonard in Middleton is said to be the oldest war memorial in the United Kingdom. <API> **WikiSearch(War memorial Flodden) → Battle of Flodden > Commemoration > The stained-glass Flodden Window in Middleton Parish Church [. . .] was constructed by Sir Richard Assheton in memory of the Battle of Flodden and the archers from Middleton who fought in it.** </API> Sir Richard Assheton of Middleton (who built St Leonard) was granted knighthood [. . .]	5.49	✓
Note: The WL will be open on Friday, <API> **Calendar() → Today is Thursday, March 9, 2017.** </API> March 10, and Sunday, March 19 for regular hours.	2.11	✓
The Nile has an approximate length of <API> **QA(What is the approximate length of the Nile?) → 6,853 km** </API> 6,853 kilometers, the White Nile being its main source.	2.08	✓
If Venus had an atmosphere similar to Earth's then you would expect Venus' mean temperature to be 499 K (1.74 x 287) rather than 735 K which is <API> **Calculator(735 / 499) → 1.47** </API> 1.47 (735 / 499) times hotter than it should be.	1.59	✓
You are here: Home / Featured / Catch this fast train to success! <API> **WikiSearch(Fast train success) → Fast Train > It also peaked at #23 on the Canadian CHUM singles chart, on June 26, 1971. The success of this single established Myles Goodwyn as the band's main songwriter, and made it possible for April Wine to record a second album.** </API> Don't wait weeks and [. . .]	0.92	✗

Information Sensitivity: General\External

Creation of dataset for Toolformer training

Os Melhores Escolas em Jersey 2020 <API> **MT(Os Melhores Escolas em Jersey) → The Best Schools in Jersey** </API> On this page you can search for Universities, Colleges and Business schools in Jersey	0.70	✓
Enjoy these pictures from the <API> **Calendar() → Today is Friday, April 19, 2013.** </API> Easter Egg Hunt.	0.33	✓
85 patients (23%) were hospitalised alive and admitted to a hospital ward. Of them, <API> **Calculator(85 / 23) → 3.70** </API> 65% had a cardiac aetiology [. . .]	−0.02	✗
But hey, after the <API> **Calendar() → Today is Saturday, June 25, 2011.** </API> Disneyland fiasco with the fire drill, I think it's safe to say Chewey won't let anyone die in a fire.	−0.41	✗
The last time I was with <API> **QA(Who was last time I was with?) → The Last Time** </API> him I asked what he likes about me and he said he would tell me one day.	−1.23	✗

Table 10: Examples of API calls for different tools, sorted by the value of $L_i^- - L_i^+$ that is used as a filtering criterion. High values typically correspond to API calls that are intuitively useful for predicting future tokens.

Information Sensitivity: General\External

LLM

Finetuning:
The newly created data set is used to finetune LLM using a standard language modelling objective.

The ”augmented” data set is a version of original data used for training, which helps preserve LLM general capability.

API calls are inserted in positions where they are needed and helpful motivating model to learn when to use them.

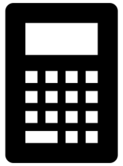
Inference:
At inference stage the model behaves normally up to a point it sees special token (->) at which point it executes the api to get response. The generation process continues on previous text with API response added at the end.

Information Sensitivity: General\External

Tools

Information Sensitivity: General\External

Tools

What other name is Pittsburg known by?	War memorial FLodden	3435*235/9	None	Os Melhores Escolas em Jersey
↓	↓	↓	↓	↓
				
↓	↓	↓	↓	↓
The Steel City	[…] was created in memory of the Battle of FLodden	89691.67	November 15th, 2023	The best schools in Jersey

<u>Constraints:</u>

- Both input and output is represented as text
- The intended use can be demonstrated and acquired

Information Sensitivity: General\External

Tools

API Name	Example Input	Example Output
Question Answering	Where was the Knights of Columbus founded?	New Haven, Connecticut
Wikipedia Search	Fishing Reel Types	Spin fishing > Spin fishing is distinguished between fly fishing and bait cast fishing by the type of rod and reel used. There are two types of reels used when spin fishing, the open faced reel and the closed faced reel.
Calculator	27 + 4 * 2	35
Calendar	ε	Today is Monday, January 30, 2023.
Machine Translation	sûreté nucléaire	nuclear safety

Table 1: Examples of inputs and outputs for all APIs used.

Information Sensitivity: General\External

Tools

What other name is Pittsburg known by?

The Steel City

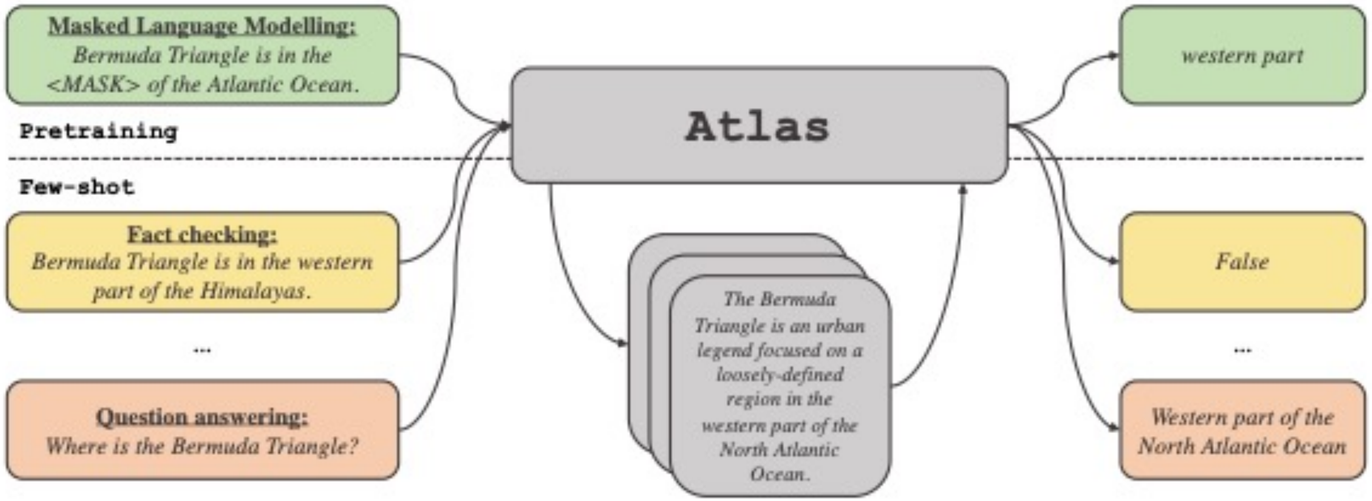


Figure 1: We introduce ATLAS, a retrieval-augmented language model that exhibits strong few-shot performance on knowledge tasks, and uses retrieval during both pre-training and fine-tuning.

Source: https://arxiv.org/pdf/2208.03299.pdf

Information Sensitivity: General\External

Tools

War memorial FLodden

[…] was created in memory of the Battle of FLodden

BM25 also known as the Okapi BM25, is a ranking function used in information retrieval systems to estimate the relevance of documents to a given search query.

3435*235/9

Simple Python script supporting 4 basic operations: +, -, *, /

89691.67

Information Sensitivity: General\External

Tools

Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis

Wenhao Zhu[1,2*], Hongyi Liu[3*], Qingxiu Dong[4], Jingjing Xu[2]
Shujian Huang[1], Lingpeng Kong[2,5], Jiajun Chen[1], Lei Li[6],
[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2]Shanghai AI Lab [3] Shanghai Jiao Tong University [4] Peking University
[5] The University of Hong Kong [6] University of California, Santa Barbara
zhuwh@smail.nju.edu.cn, liu.hong.yi@sjtu.edu.cn, dqx@stu.pku.edu.cn, jingjingxu@pku.edu.cn
huangsj@nju.edu.cn, lpk@cs.hku.hk, chenjj@nju.edu.cn, lilei@cs.ucsb.edu,

LLM vs Supervised:

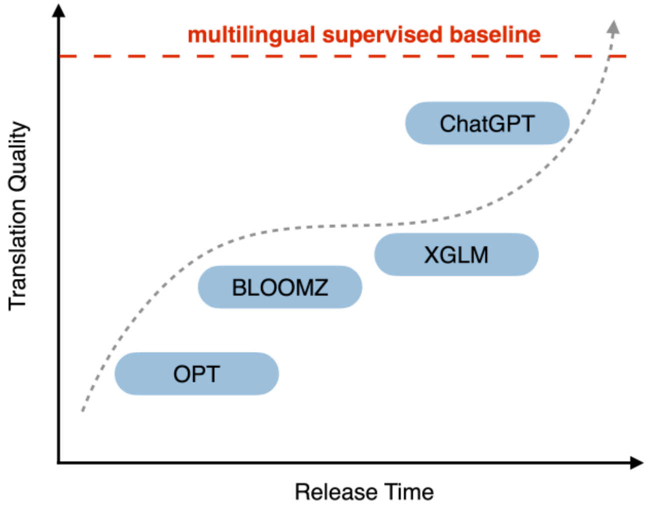


Figure 1: Averaged multilingual translation performance of popular LLMs. These models demonstrated great potential in multilingual machine translation.

et al., 2022a; Chowdhery et al., 2022). On the other hand, compared with a widely-used supervised baseline NLLB (Costa-jussà et al., 2022), ChatGPT only achieves higher performance on 16.67% translation directions. We further analyze the trans-

Source: https://arxiv.org/pdf/2304.04675.pdf, https://arxiv.org/pdf/2207.04672.pdf, https://github.com/facebookresearch/fairseq/tree/nllb

Information Sensitivity: General\External

Downstream tasks

Downstream tasks

Baseline models:

- **GPT-J**: A regular GPT-J model without any finetuning.
- **GPT-J + CC**: GPT-J finetuned on $\mathcal{C}$, our subset of CCNet *without* any API calls.
- **Toolformer**: GPT-J finetuned on $\mathcal{C}^*$, our subset of CCNet augmented with API calls.
- **Toolformer (disabled)**: The same model as Toolformer, but API calls are disabled during decoding.[5]

Comments:

- GPT-J is an opensource 6 billion parameter GPT-2 like model
- Additionally, the models are compared against big general models OPT (66B parameters) and GPT-3 (175B parameters)

Information Sensitivity: General\External

Downstream tasks

Subset of LAMA benchmark (fill statement with missing fact):

Model	SQuAD	Google-RE	T-REx
GPT-J	17.8	4.9	31.9
GPT-J + CC	19.2	5.6	33.2
Toolformer (disabled)	22.1	6.3	34.9
Toolformer	**33.8**	**11.5**	**53.5**
OPT (66B)	21.6	2.9	30.1
GPT-3 (175B)	26.8	7.0	39.8

Table 3: Results on subsets of LAMA. Toolformer uses the question answering tool for most examples, clearly outperforming all baselines of the same size and achieving results competitive with GPT-3 (175B).

Comments:

- Without tools all models achieve similar performance correlated with size
- Usage of tools makes the model clearly outperform all baselines
- In 98.1% of cases API was called

Information Sensitivity: General\External

Downstream tasks

Math dataset:

Model	ASDiv	SVAMP	MAWPS
GPT-J	7.5	5.2	9.9
GPT-J + CC	9.6	5.0	9.3
Toolformer (disabled)	14.8	6.3	15.0
Toolformer	**40.4**	**29.4**	**44.0**
OPT (66B)	6.0	4.9	7.9
GPT-3 (175B)	14.0	10.0	19.8

Table 4: Results for various benchmarks requiring mathematical reasoning. Toolformer makes use of the calculator tool for most examples, clearly outperforming even OPT (66B) and GPT-3 (175B).

Comments:

- Interestingly Toolformer even with API disabled achieves good performance (probably due to training on more mathematical samples)
- Allowing API usage more than doubles model performance and clearly outperforms big general purpose models
- In 97.9% of cases API was called

Downstream tasks

Question answering:

Model	WebQS	NQ	TriviaQA
GPT-J	18.5	12.8	43.9
GPT-J + CC	18.4	12.2	45.6
Toolformer (disabled)	18.9	12.6	46.7
Toolformer	**26.3**	**17.7**	**48.8**
OPT (66B)	18.6	11.4	45.7
GPT-3 (175B)	<u>29.0</u>	<u>22.6</u>	<u>65.9</u>

Table 5: Results for various question answering dataset. Using the Wikipedia search tool for most examples, Toolformer clearly outperforms baselines of the same size, but falls short of GPT-3 (175B).

Comments:

- 99.3% cases API call was made
- For Multilingual depending on the language but between 7.3% (Hindu) and 94.9%

Multilingual question answering:

Model	Es	De	Hi	Vi	Zh	Ar
GPT-J	15.2	**<u>16.5</u>**	1.3	8.2	**<u>18.2</u>**	**<u>8.2</u>**
GPT-J + CC	15.7	14.9	0.5	8.3	13.7	4.6
Toolformer (disabled)	19.8	11.9	1.2	10.1	15.0	3.1
Toolformer	**<u>20.6</u>**	13.5	**<u>1.4</u>**	**<u>10.6</u>**	16.8	3.7
OPT (66B)	0.3	0.1	1.1	0.2	0.7	0.1
GPT-3 (175B)	3.4	1.1	0.1	1.7	17.7	0.1
GPT-J (All En)	24.3	27.0	23.9	23.3	23.1	23.6
GPT-3 (All En)	24.7	27.2	26.1	24.9	23.6	24.0

Table 6: Results on MLQA for Spanish (Es), German (De), Hindi (Hi), Vietnamese (Vi), Chinese (Zh) and Arabic (Ar). While using the machine translation tool to translate questions is helpful across all languages, further pretraining on CCNet deteriorates performance; consequently, Toolformer does not consistently outperform GPT-J. The final two rows correspond to models that are given contexts and questions in English.

Information Sensitivity: General\External

Downstream tasks

Temporal dataset:

Model	TEMPLAMA	DATESET
GPT-J	13.7	3.9
GPT-J + CC	12.9	2.9
Toolformer (disabled)	12.7	5.9
Toolformer	**16.3**	**27.3**
OPT (66B)	14.5	1.3
GPT-3 (175B)	15.5	0.8

Table 7: Results for the temporal datasets. Toolformer outperforms all baselines, but does not make use of the calendar tool for TEMPLAMA.

Comments:

- Dataset is created by authors and contains simple temporal questions (e.g. What day of the week was 30 days ago?)
- API usage was different for those datasets 0.2% for TempLama and 54.8% for Dataset

Information Sensitivity: General\External

Other (preserving LLM core abilities, scaling, limitations)

Other (preserving LLM core abilities, scaling, limitations)

Language modelling:

Model	WikiText	CCNet
GPT-J	**9.9**	10.6
GPT-J + CC	10.3	**10.5**
Toolformer (disabled)	10.3	**10.5**

Table 8: Perplexities of different models on WikiText and our validation subset of CCNet. Adding API calls comes without a cost in terms of perplexity for language modeling without any API calls.

Comments:

- Training on dataset augmented with API calls does not change the metrics significantly

Other (preserving LLM core abilities, scaling, limitations)

Scaling Laws:

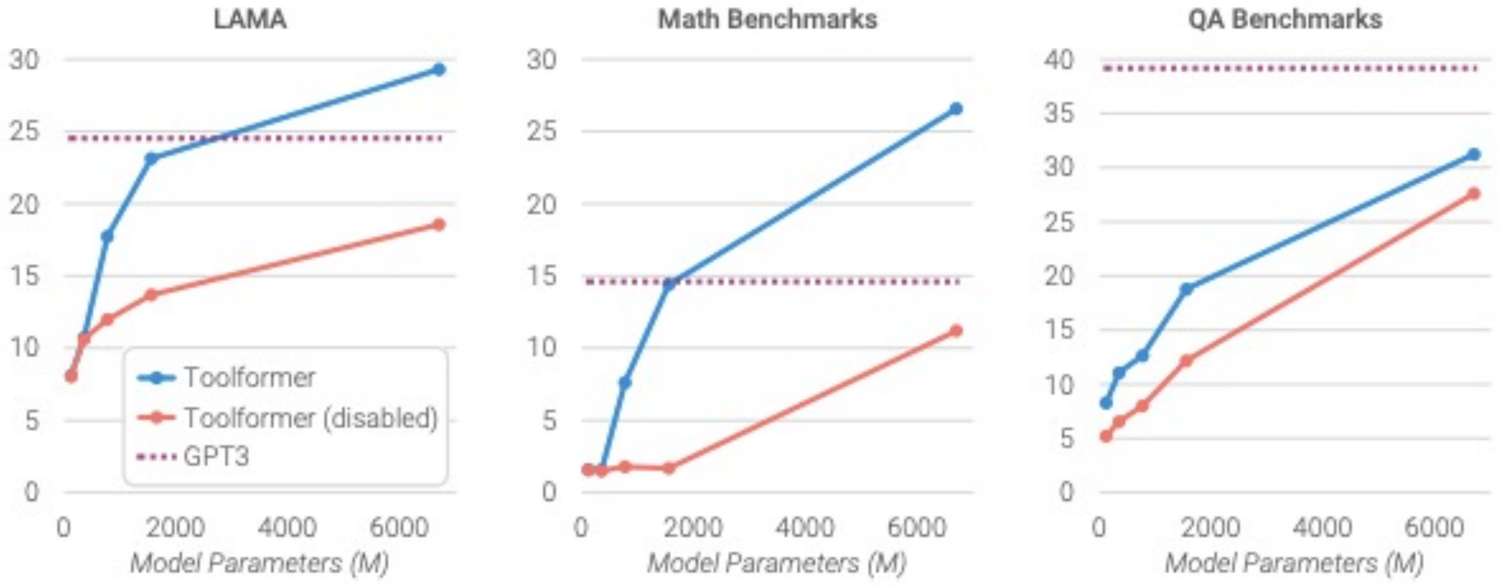


Figure 4: Average performance on LAMA, our math benchmarks and our QA benchmarks for GPT-2 models of different sizes and GPT-J finetuned with our approach, both with and without API calls. While API calls are not helpful to the smallest models, larger models learn how to make good use of them. Even for bigger models, the gap between model predictions with and without API calls remains high.

Comments:

- Models generally become better at solving tasks with increase in size
- Small models do not learn to use tools
- Ability to make good use of tools increases with size as well

Information Sensitivity: General\External

Other (preserving LLM core abilities, scaling, limitations)

- Toolformer limitations
  - No chaining of API calls
  - In some cases data acquisition method is sample inefficient
  - No interactive API usage (refining answer based on subsequent calls)
  - Sensitivity to exact wording on input when deciding to use API
  - No knowledge of cost of using external tool

Information Sensitivity: General\External