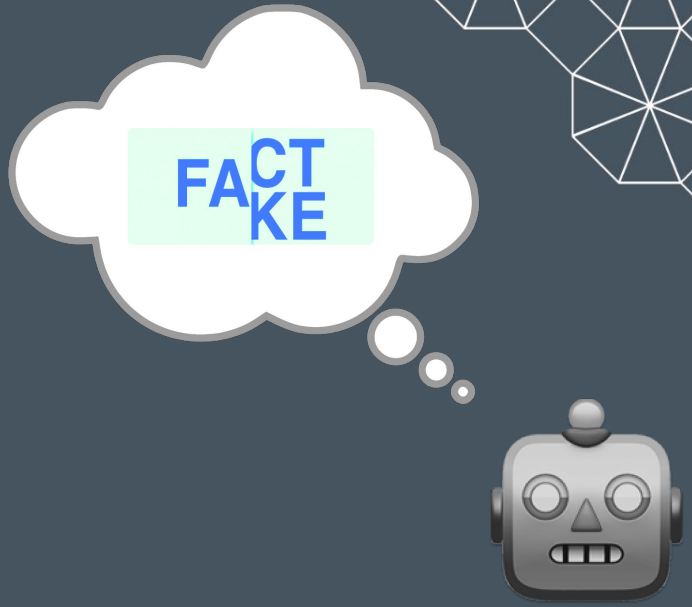
Polyglot or Not?

Measuring Multilingual Encyclopedic Knowledge Retrieval from LLMs



Shreshta Bhat, Daniel Furman, Tim Schott

Advisor: David Bamman

Table of contents

1. Motivation [1.5 min]
2. Background [1.25 min]
3. Methods [1.5 min]
4. Results and Analysis [2 min]
5. Impact [2 min]
6. Takeaways [1 min]
7. Q/A

Motivation

GenAI Expectations...

arXiv

https://arxiv.org › cs

[2303.12712] Sparks of Artificial General Intelligence

Mar 22, 2023 — The latest model developed by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data. In this paper, we report on our ...

...GenAI Reality

The Verge

Google's AI chatbot Bard makes factual error in first demo

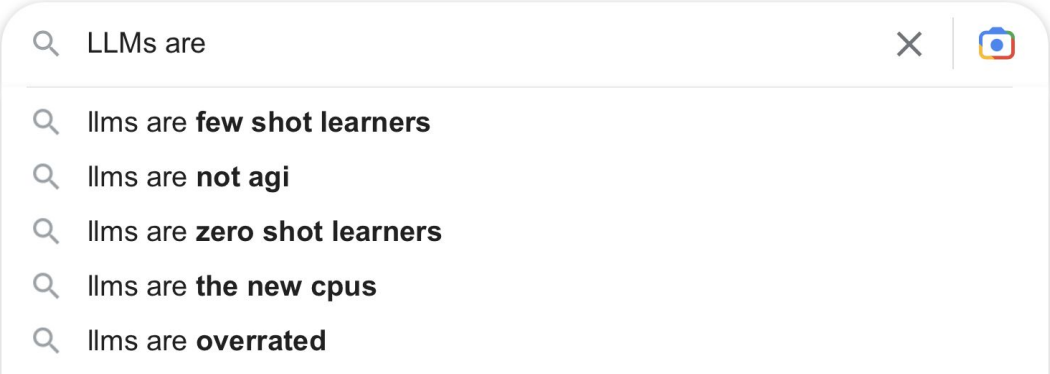
On Monday, Google announced its AI chatbot Bard — a rival to OpenAI's ChatGPT that's due to become “more widely available to the public in...

Feb 8, 2023

Background

What’s an LLM?

- Large Language Model (**LLM**): predict the **next** word using previous words as context



LLMs as knowledge bases

- LLMs can retrieve **factual associations** in training data

Query	Answer	Generation
Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5]
The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7]
Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5]

Petroni et al. (2019), "Language Models as Knowledge Bases? (ACL)

Berkeley
UNIVERSITY OF CALIFORNIA

“The more data, the better. But **what’s available to train a model varies widely across** the thousands of **languages** used today.”

–Viorica Marian, director of the Bilingualism and Psycholinguistics Research Lab at Northwestern

Berkeley
UNIVERSITY OF CALIFORNIA

Research Question

How well do LLMs retrieve facts in **different languages**?

Approach

The “Polyglot or Not?” Fact-Completion Test

Contrastive Knowledge Assessment

Prompt:	Paris	is	the	capital	of
Token ids:	7270	83	70	10323	111

Transformer
(Language Model)

Prediction probs:	0.12	0.01	...	0.61
	Spain	vs.		France

Test dataset: 303k facts across 20 languages

True statements ✔ and **counterfactuals** ✘

English: Sundar Pichai works for

- **Google** ✔ vs. **Apple** ✘

French: Sundar Pichai travaille pour

- **Google** ✔ vs. **Apple** ✘

Ukrainian: Сундар Пічаї працює в

- **Google** ✔ vs. **Apple** ✘

Results and Analysis

Multilingual Leaderboard (20 Lang Average)

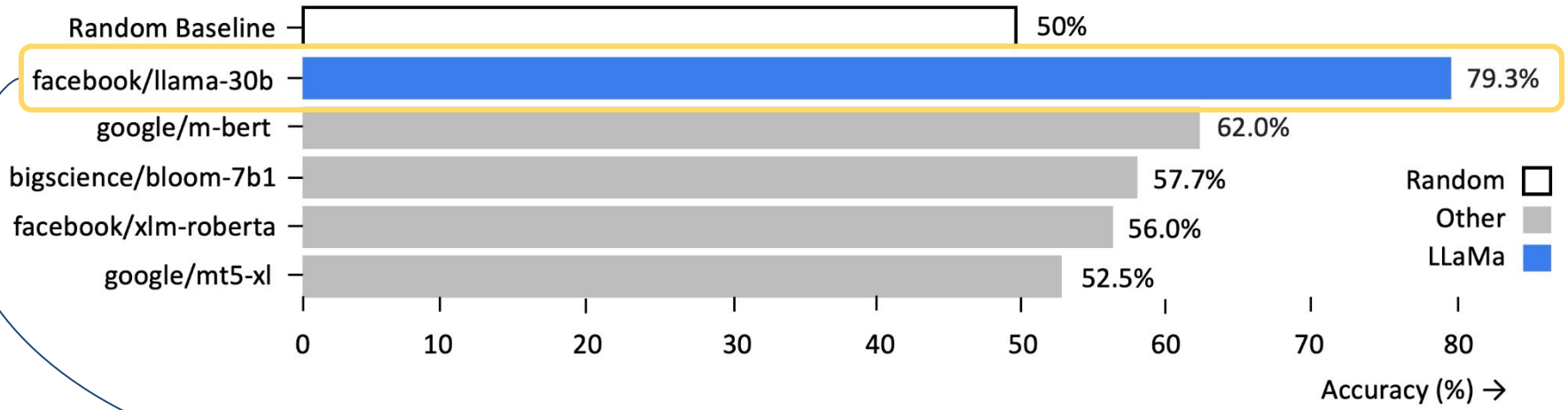


Figure: Meta's LLaMa-30b outperforms other multilingual foundation models by a large margin across 20 languages

LLaMa-30B Multilingual Performance

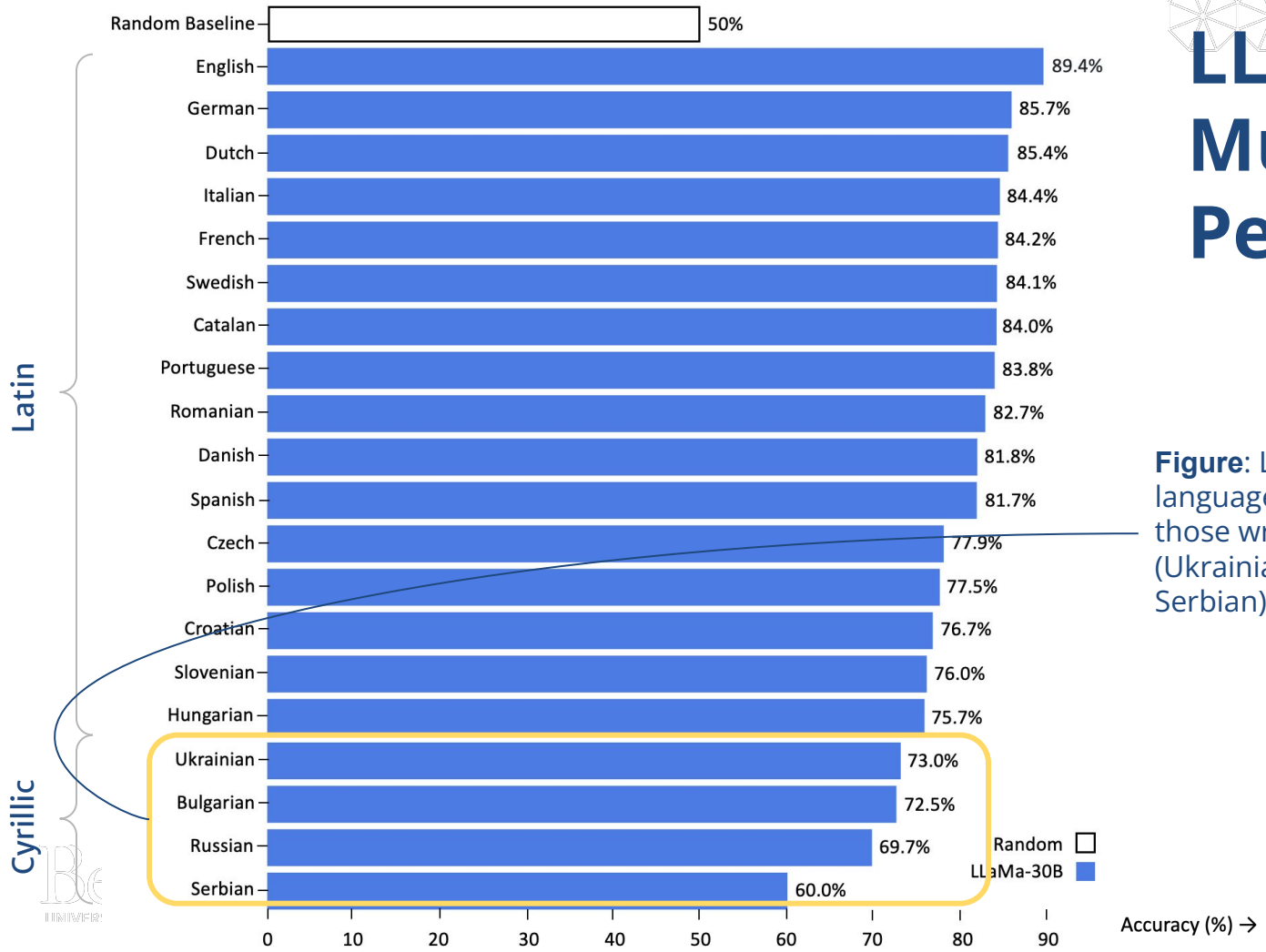


Figure: LLaMa-30b scores higher on languages written in Latin script than those written in Cyrillic script (Ukrainian, Bulgarian, Russian and Serbian) ($p < 0.001$).

Impact

“The **most powerful models** will be trained on about 20 **‘high-resource languages**’

“The more data, the better. But **what’s available to train a model varies widely across** the thousands of **languages** used today.”d

AI will churn out massive amounts of new text **mostly in those languages**. Like invasive species, such **dominant models could drive out languages for which fewer resources exist**.”

–Viorica Marian, director of the Bilingualism and Psycholinguistics Research Lab at Northwestern

Takeaways

Project

- **Size** + **robustness** of **training data** outweighs # of model parameters for fact-completion.
- This is the **first counterfactual** + **multilingual** knowledge assessment for open-source LLMs.

Personal

- **Overcoming setbacks** and work within **limitations**
- **End-to-end, collaborative research** – I School style!

Scan to see code + data!

Github repo

Huggingface dataset

Thanks!

Shreshta Bhat

social impact and ethics

Tim Schott

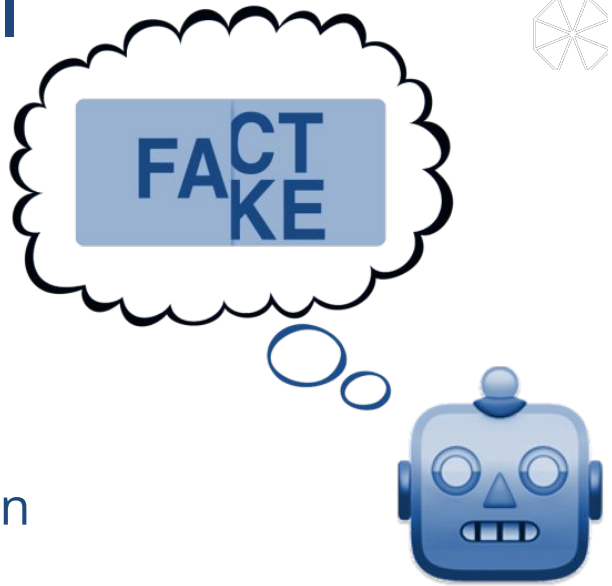
cultural analytics

Daniel Furman

commercial analytics

Berkeley
UNIVERSITY OF CALIFORNIA

Polyglot or Not? Measuring Multilingual Encyclopedic Knowledge Retrieval from LLMs



Shreshta Bhat, Tim Schott, and Daniel Furman

Berkeley
UNIVERSITY OF CALIFORNIA

Code links

probe_t5 helper
compare _models routine
Runnable notebook
Huggingface dataset

Berkeley
UNIVERSITY OF CALIFORNIA

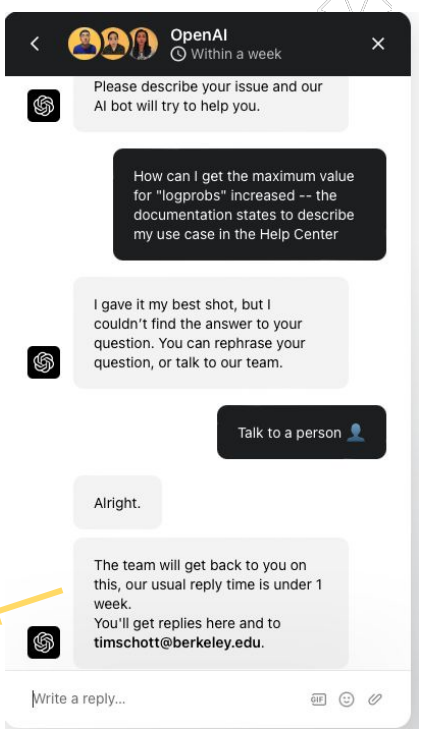
Why couldn't we study GPT-3/4/Chat-GPT?

- OpenAI's models are closed-source, and the API for GPT-3 does not support our CKA method.
- We can't see all the probabilities for the tokens their models predict, as CKA requires.

`logprobs` integer Optional Defaults to null

Include the log probabilities on the `logprobs` most likely tokens, as well the chosen tokens. For example, if `logprobs` is 5, the API will return a list of the 5 most likely tokens. The API will always return the `logprob` of the sampled token, so there may be up to `logprobs+1` elements in the response.

The maximum value for `logprobs` is 5. If you need more than this, please contact us through our **Help center** and describe your use case.



You don't need to be a data scientist to predict if I got a response..

Berkeley
UNIVERSITY OF CALIFORNIA

English-only Test Leaderboard

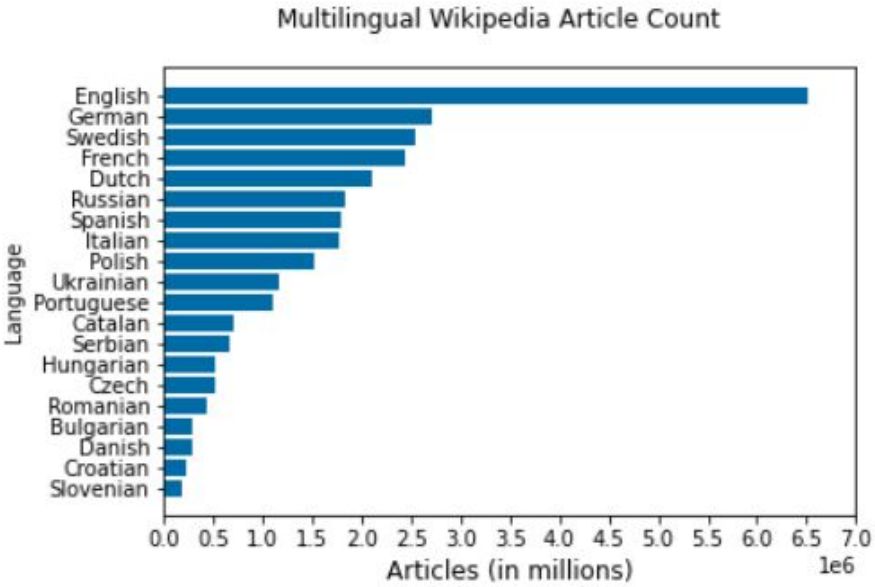
Accuracy (%) →

Future Directions

- Vocabulary-wide probabilities for close sourced inference?
- Evaluating more open-source families as they are released by
- Additional error analysis with new types of metadata
- Increasing size + adding more languages to test dataset
- Model editing (MEMIT) in a multilingual setting

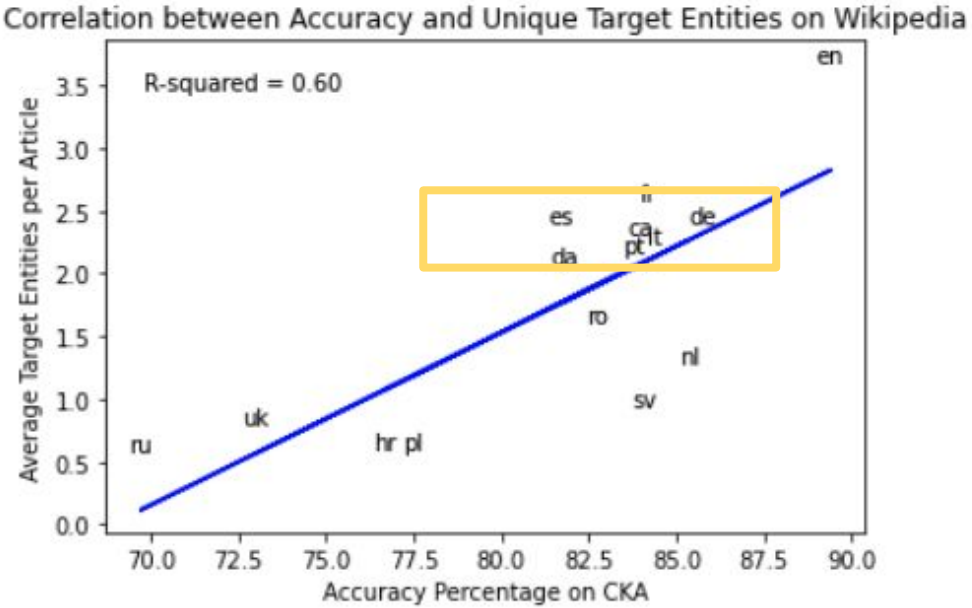
Berkeley
UNIVERSITY OF CALIFORNIA

How big is wikipedia, per language?



number of pages as of June 2022 (LLaMa's training period)

Data Quality > Data Quantity



Why CKA?

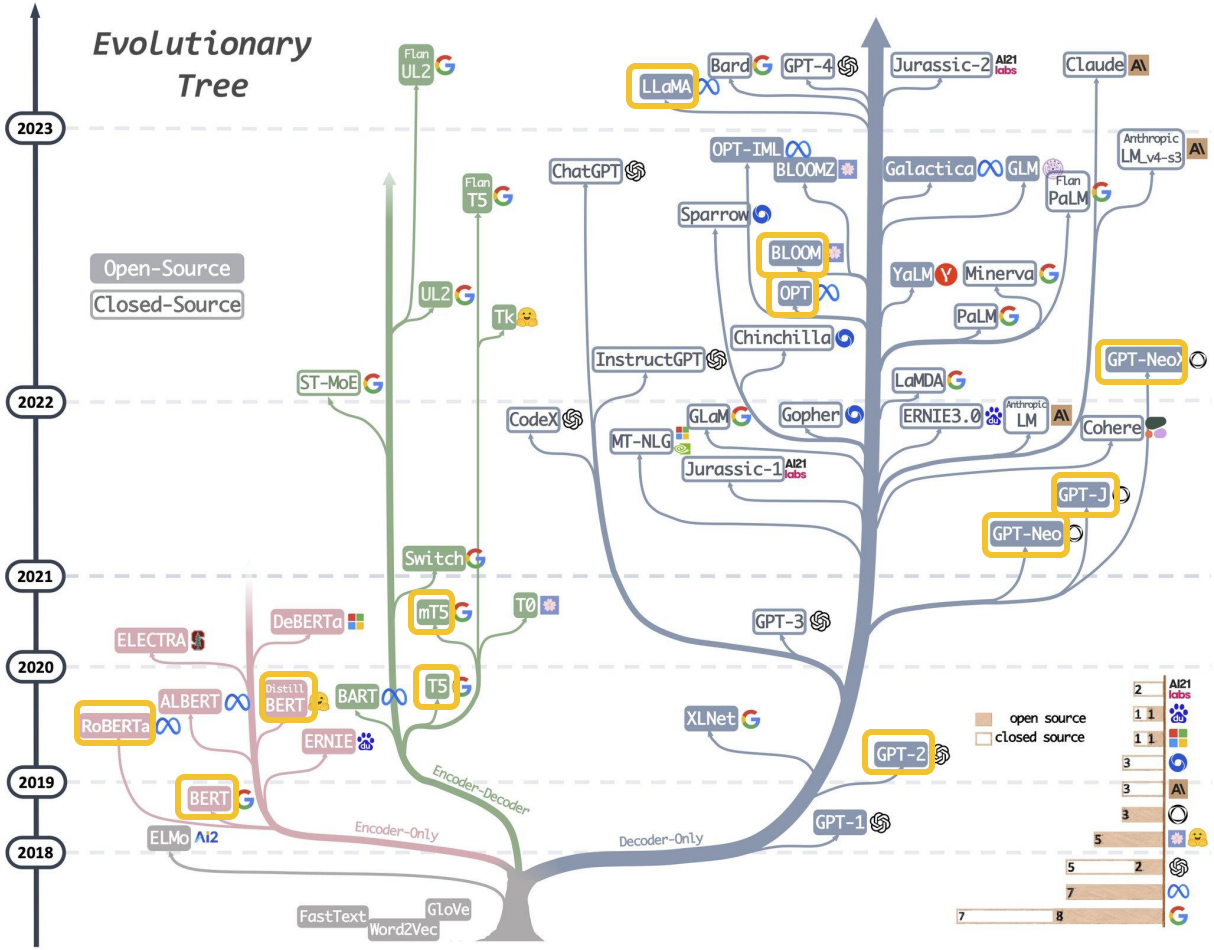
Why not look at top predicted tokens?

- Examining the contrast allows us to quantify the confidence level
- Sometimes, there are multiple right answers, some of which are not present in top tokens
- High probability counterfacts provide insights about model

Berkeley
UNIVERSITY OF CALIFORNIA

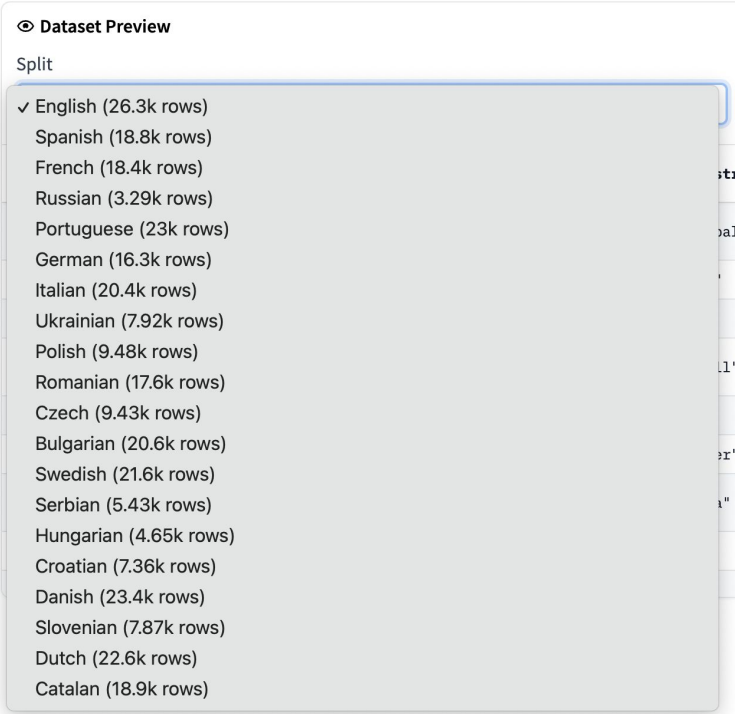
LLMs Included

20 foundation LLMs are including in our project's results



Berkeley
UNIVERSITY OF CALIFORNIA

Dataset Counts



Dataset Preview

Datasets: CalibraGPT / **Fact-Completion** like 4

Tasks: Text Generation | Fill-Mask | Text2Text Generation | Languages: English | French | Spanish | +17 | Multilinguality: multilingual | Size Categories: 100K<n<1M

Language Creators: expert-generated | machine-generated | ArXiv: arxiv:2302.13971 | arxiv:2210.03329 | arxiv:2210.07229 | Tags: natural-language-understanding | License: apache-2.0

Dataset card | Files and versions | Community | Settings

Dataset Preview Size: 27.1 MB | API

Split

English (26.3k rows)

dataset_id (string)	stem (string)	true (string)	false (string)	relation (string)	subject (string)	object (string)
"rome_21844"	"What sport does Megan Rapinoe play? They play"	"soccer"	"basketball"	"P641"	"Megan Rapinoe"	"soccer"
"rome_9500"	"Trailer Park Boys was from"	"Canada"	"Norway"	"P495"	"Trailer Park Boys"	"Canada"
"rome_9881"	"Joni Mitchell performs"	"folk"	"opera"	"P136"	"Joni Mitchell"	"folk"
"rome_11754"	"Wilt Chamberlain is a professional"	"basketball"	"football"	"P641"	"Wilt Chamberlain"	"basketball"
"calinet_8922"	"Prius is produced by"	"Toyota"	"Honda"	"P176"	"Prius"	"Toyota"
"rome_9037"	"How I Met Your Mother is a"	"sitcom"	"thriller"	"P136"	"How I Met Your Mother"	"sitcom"
"calinet_2820"	"Rwanda, which has the capital"	"Kigali"	"Kampala"	"P36"	"Rwanda"	"Kigali"
"rome_10452"	"Suwayq, which is located in"	"Oman"	"India"	"P17"	"Suwayq"	"Oman"
"rome_5025"	"Sundar Pichai works for"	"Google"	"Apple"	"P108"	"Sundar Pichai"	"Google"
"rome_15553"	"In Quebec City, an official language is"	"French"	"English"	"P37"	"Quebec City"	"French"

< Previous 1 2 3 ... 263 Next >

Almost every “massive” LLM is closed source

List of large language models

Name	Release date[a]	Developer	Number of parameters[b]	Corpus size	License[c]	Notes
GPT-4	March 2023	OpenAI	Unknown[f]	Unknown	public web API	Available for ChatGPT Plus users and used in several products.
GLaM (Generalist Language Model)	December 2021	Google	1.2 trillion[42]	1.6 trillion tokens[42]	Proprietary	Sparse mixture-of-experts model, making it more expensive to train but cheaper to run inference compared to GPT-3.
PanGu-Σ	March 2023	Huawei	1.085 trillion	329 billion tokens[65]	Proprietary	
PaLM (Pathways Language Model)	April 2022	Google	540 billion[48]	768 billion tokens[47]	Proprietary	aimed to reach the practical limits of model scale
Minerva	June 2022	Google	540 billion[52]	38.5B tokens from webpages filtered for mathematical content and from papers submitted to the arXiv preprint server[52]	Proprietary	LLM trained for solving "mathematical and scientific questions using step-by-step reasoning".[53] Minerva is based on PaLM model, further trained on mathematical and scientific data.
Megatron-Turing NLG	October 2021[36]	Microsoft and Nvidia	530 billion[37]	338.6 billion tokens[37]	Restricted web access	Standard architecture but trained on a supercomputing cluster.
Gopher	December 2021	DeepMind	280 billion[43]	300 billion tokens[44]	Proprietary	
Ernie 3.0 Titan	December 2021	Baidu	260 billion[38]	4 Tb	Proprietary	Chinese-language LLM. Ernie Bot is based on this model.
GPT-3	2020	OpenAI	175 billion[10]	499 billion tokens[29]	public web API	A fine-tuned variant of GPT-3, termed GPT-3.5, was made available to the public through a web interface called ChatGPT in 2022.[31]
OPT (Open Pretrained Transformer)	May 2022	Meta	175 billion[49]	180 billion tokens[50]	Non-commercial research[d]	GPT-3 architecture with some adaptations from Megatron

Wikipedia, Large Language Models