



LLM evaluation, data, and benchmarking

October 6th, 2023

Yuanpeng Tu

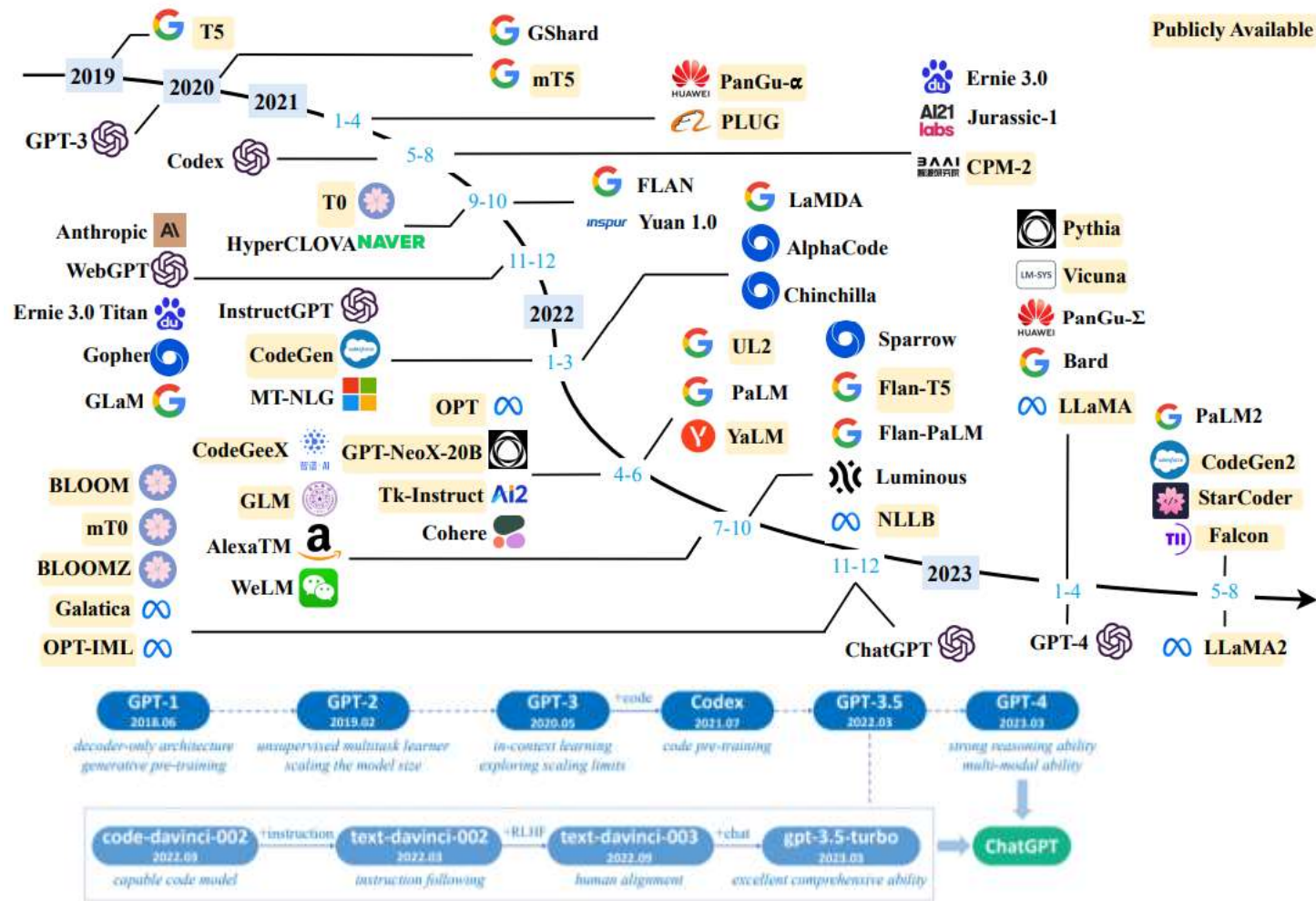
Zhuoling Li

Catalogue

- LLM Evaluation
- VLM Evaluation
- LLM Data
- LLM benchmark

Background of LLM Evaluation

Overview of development of LLM



Rapid development of large language models

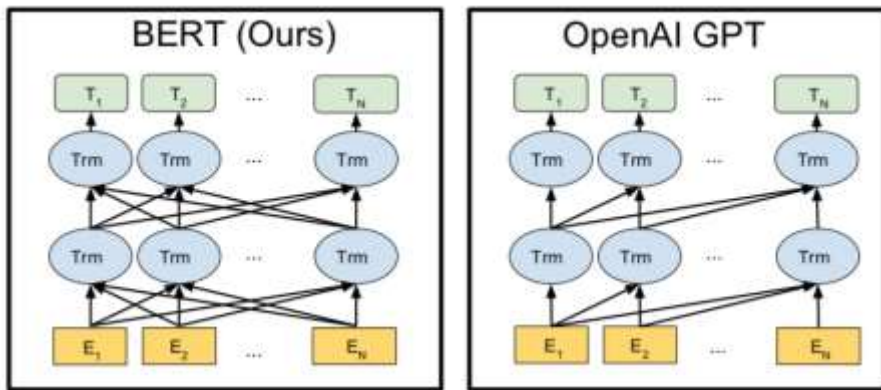
Why we need LLM Evaluation?

- Better understand the strengths and weakness of LLMs.
- Provide guidance on how to design human-LLMs interaction.
- Ensure safety and reliability.

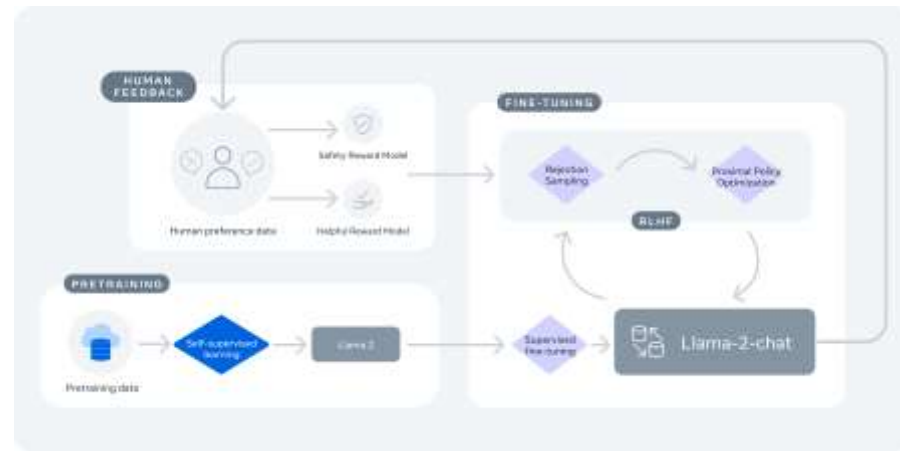
PLMs vs LLMs

Early PLMs: ELMo/BERT/GPT-2/BART

LLMs: GPT-3&4/PaLM/LLaMA



Architecture Difference



Data Difference

The model has been much stronger and how to evaluate it becomes a challenge:

- Metric?
- Manual evaluation?

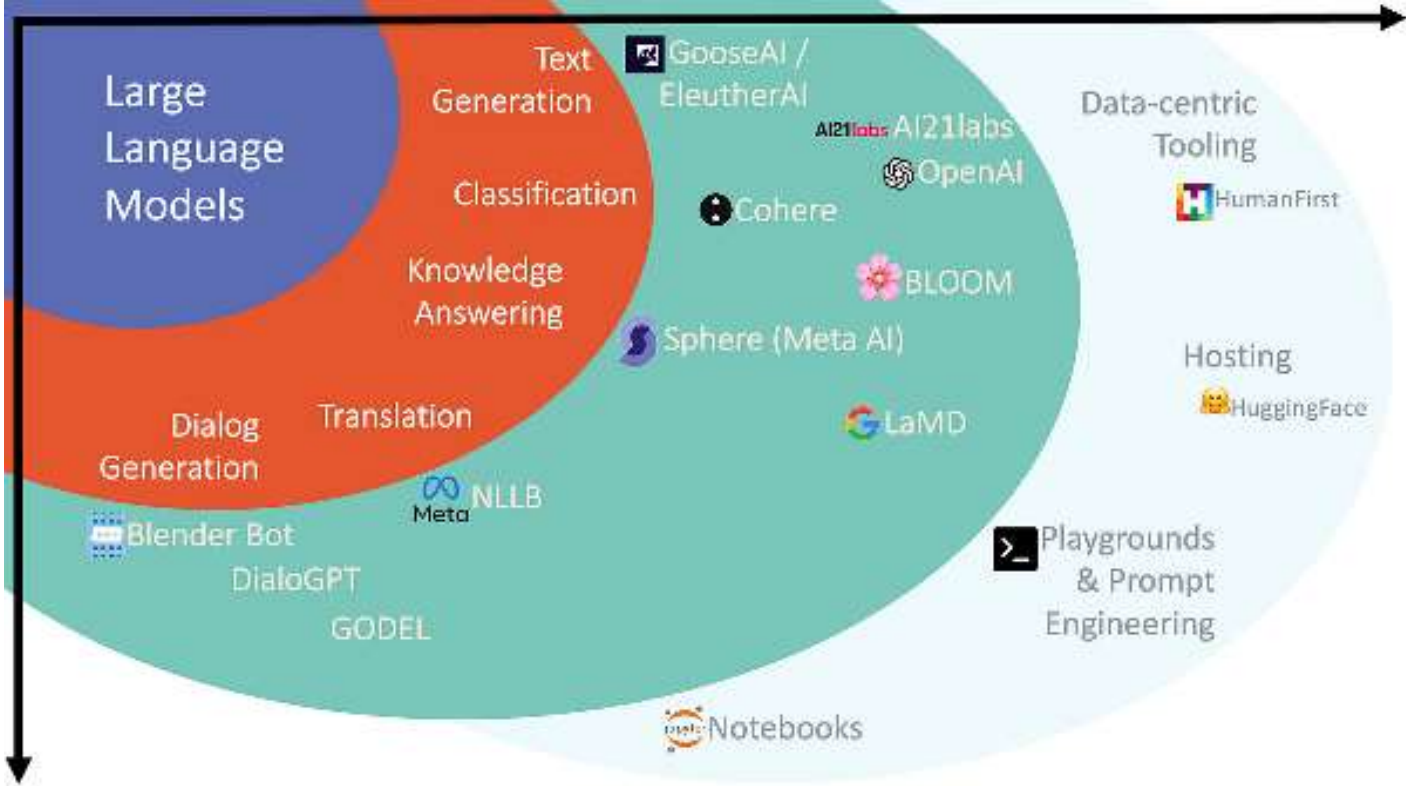
How We Evaluate It Before?

Use independent specific tasks:

- Language Modelling
- Sentiment Analysis
- Text Classification
- Classification
- Retrieval
- Question Answering
- Speech Recognition
- Named Entity Recognition

- Although some models get better scores than human under previous metrics, their expression ability is still worse than human.
- So, how can we evaluate the capability of LLM properly?

Development of LLM Evaluation



From low-level tasks to high-level tasks requiring reasoning

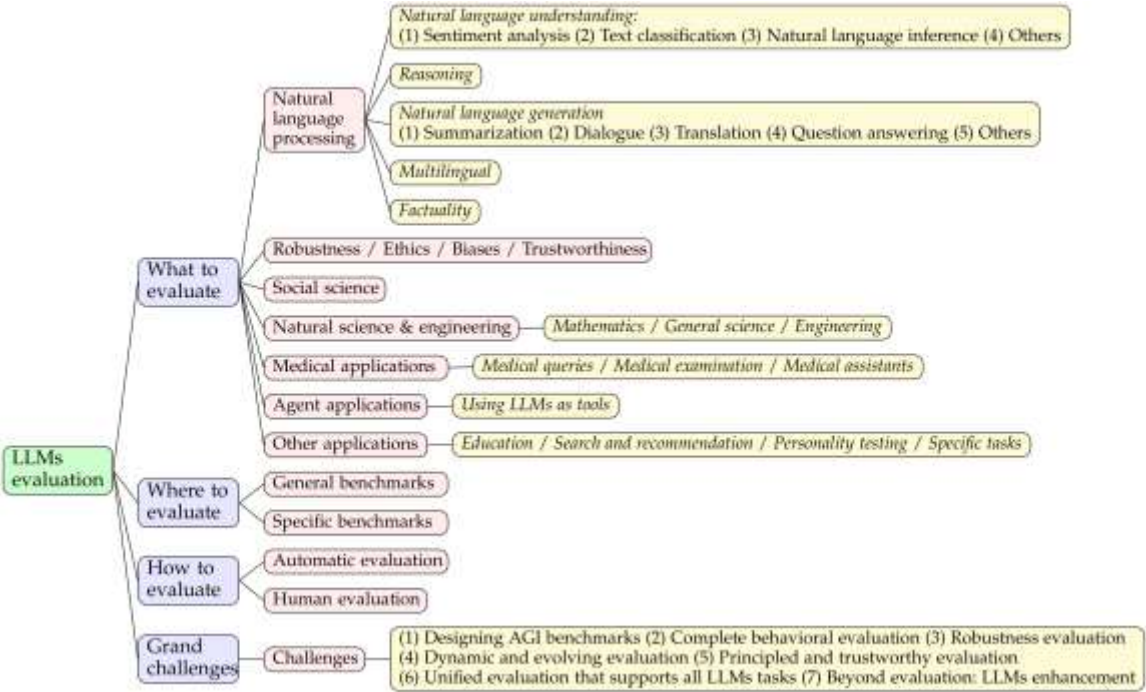
Existing Evaluation Benchmarks

Trend:

- More tasks.
- More high-level tasks.

How to categorize existing benchmarks:

- What to evaluate
- Where to evaluate
- How to evaluate
- Included Grand Challenges



Popular Evaluation Benchmarks

Framework Name	Factors Considered	Url Link
Big Bench	Generalization abilities	https://github.com/google/BIG-bench
GLUE Benchmark	Grammar, Paraphrasing, Text Similarity, Inference	https://gluebenchmark.com/
SuperGLUE Benchmark	Natural Language Understanding, Reasoning, Understanding complex sentences beyond training data,	https://super.gluebenchmark.com/
OpenAI Moderation API	harmful or unsafe content	https://platform.openai.com/docs/api-reference/moderations
MMLU	understanding across various tasks and domains	https://github.com/hendrycks/test

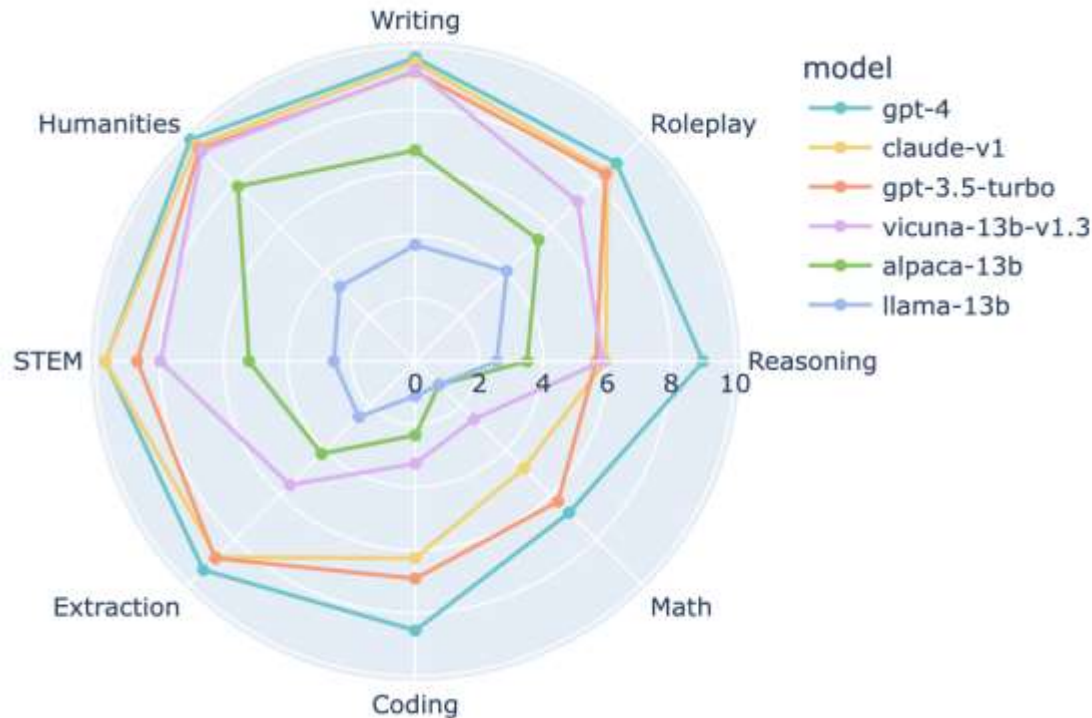
Popular Evaluation Leaderboards

Table 1: LLM Leaderboard (Timeframe: April 24 - June 19, 2023). The latest and detailed version [here](https://lmsys.org/blog/2023-06-22-leaderboard/).

Model	MT-bench (score) ▾	Arena Elo Rating	MMLU	License
GPT-4	8.99	1227	86.4	Proprietary
GPT-3.5-turbo	7.94	1130	70.0	Proprietary
Claude-v1	7.90	1178	75.6	Proprietary
Claude-instant-v1	7.85	1156	61.3	Proprietary
Vicuna-33B	7.12	-	59.2	Non-commercial
WizardLM-30B	7.01	-	58.7	Non-commercial
Guanaco-33B	6.53	1065	57.6	Non-commercial
Tulu-30B	6.43	-	58.1	Non-commercial
Guanaco-65B	6.41	-	62.1	Non-commercial
OpenAssistant-LLaMA-30B	6.41	-	56.0	Non-commercial
PaLM-Chat-Bison-001	6.40	1038	-	Proprietary
Vicuna-13B	6.39	1061	52.1	Non-commercial
MPT-30B-chat	6.39	-	50.4	CC-BY-NC-SA-4.0
WizardLM-13B	6.35	1048	52.3	Non-commercial
Vicuna-7B	6.00	1008	47.1	Non-commercial
Baize-v2-13B	5.75	-	48.9	Non-commercial
Nous-Hermes-13B	5.51	-	49.3	Non-commercial
MPT-7B-Chat	5.42	956	32.0	CC-BY-NC-SA-4.0




<https://lmsys.org/blog/2023-06-22-leaderboard/>

Popular Evaluation Leaderboards



- GPT-4 shows superior performance in Coding and Reasoning compared to GPT-3.5/Claude, while Vicuna-13B lags significantly behind in several specific categories: Extraction, Coding, and Math.

Popular Evaluation Leaderboards

Rank	Model	Elo Rating	Description
1	 vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	 koala-13b	1082	a dialogue model for academic research by BAIR
3	 oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
5	chatglm-6b	985	an open bilingual dialogue language model by Tsinghua University
6	fastchat-t5-3b	951	a chat assistant fine-tuned from FLAN-T5 by LMSYS
7	dolly-v2-12b	944	an instruction-tuned open large language model by Databricks
8	llama-13b	932	open and efficient foundation language models by Meta
9	stablelm-tuned-alpha-7b	858	Stability AI language models

Motivation: More Comprehensive Evaluation

- **Scalability.** The system should scale to a large number of models when it is not feasible to collect sufficient data for all possible model pairs.
- **Incrementality.** The system should be able to evaluate a new model using a relatively small number of trials.
- **Unique order.** The system should provide a unique order for all models. Given any two models, we should be able to tell which ranks higher or whether they are tied.

Metric: Elo Rating System

If player A has a rating of R_A and player B a rating of R_B , the exact formula (using the logistic curve with base 10) for the probability of player A winning is

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} .$$

The ratings of players can be linearly updated after each battle. Suppose player A (with Rating R_A) was expected to score E_A points but actually scored S_A points. The formula for updating that player's rating is

$$R'_A = R_A + K \cdot (S_A - E_A) .$$

Section Summary

- Designing a proper benchmark is crucial for the development of LLM.
- LLM benchmarks are gradually incorporating more and more high-level tasks.
- Existing benchmarks all have their own focuses (bias).

Data for LLM

Introduction

- LLM is trained on raw text spanning a broad range of domains, e.g., language, code, etc.
- Web is a natural place to get text data, e.g., WalMart generates 2.5 petabytes data every day.
- Web data is highly imbalanced, e.g., it is largely dominated by young users.

Data for LLM training

Alignment Datasets

Dataset name	Type	Language	Size
AmericanStories	PT	English	/
chatbot_arena_conversations	RHLF	Multilingual	33k conversations
WebGLM-qa	Pairs	English	43.6k entries
Alpaca-COT	CoT	English	/

Domain-specific Datasets

Dataset name	Type	Language	Size
starcoderdata	PT	code	/
function-invocations-25k	Pairs	English	25k entries
TheoremQA	Pairs	English	800
phi-1	Dialog	English	/

- Refer to: <https://github.com/Zjh-819/LLMDataHub>

Data for LLM training

Pre-training Datasets

Dataset name	Type	Language	Size
proof-pile	PT	English,Latex	13GB
StackOverflowpost	PT	/	35GB
CBook-150K	PT	Chinese	150k+ books

Multimodal Datasets

Dataset name	Type	Language	Size
OBELICS	Image-document	English	141M documents
JourneyDB	Image-prompt-caption	Multilingual	4M instances
MIMIC-IT	Instruction-image	Multilingual	2.2M instances

- Refer to: <https://github.com/Zjh-819/LLMDataHub>

Problems in Datasets

- Long-tail distribution.
- Bias.
- Fake information.
- Collected for specific targets.
- Different dataset organizations.
-
- Many, many, many more!

Paper Reading: Let us first build a big dataset

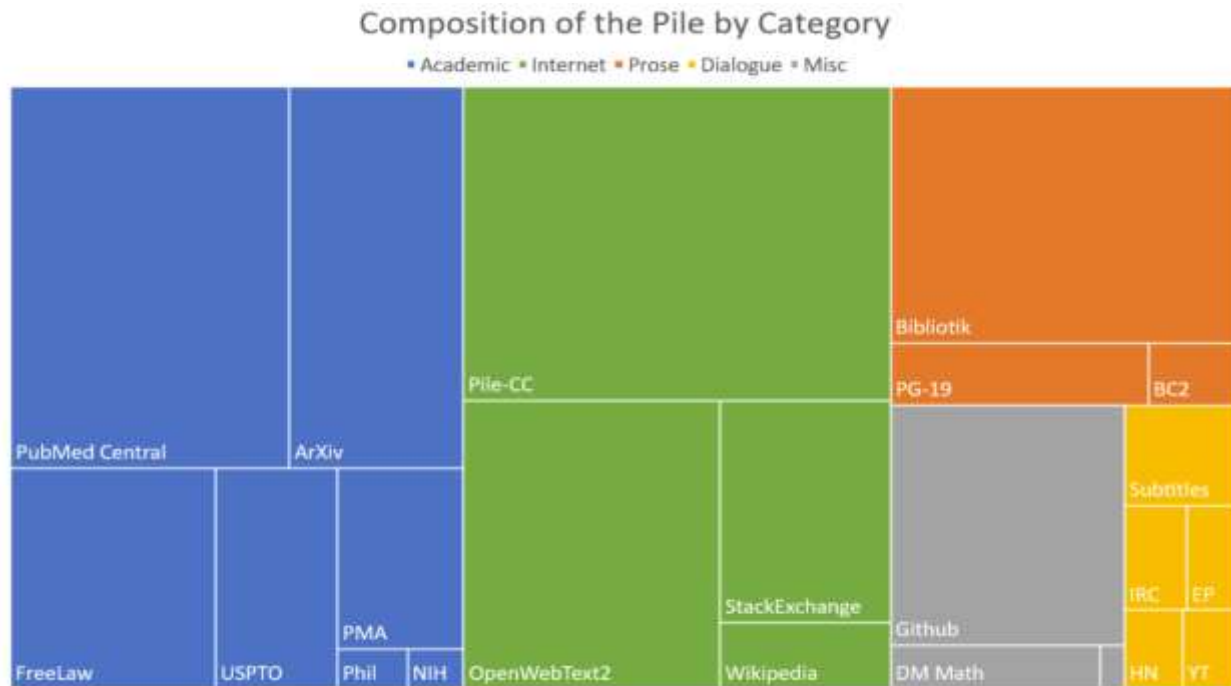
The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Motivation

- As LLM is scaling up, more data is demanded.
- More diverse data leads to better downstream generalization capability.

The Pile Dataset

- Collect a big dataset (825.18G) by combining 22 sources. This dataset includes 22 languages. (Preprinted in Dec. 2020)



Overview of the Pile Dataset

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3 [†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19) [†]	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles [†]	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en) [†]	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics [†]	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails [†]	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB



Highly Diverse and Large-scale for Training

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Train on the Pile Dataset

- Models trained on the Pile dataset shows significant improvement compared with the one trained on the Common Crawl.

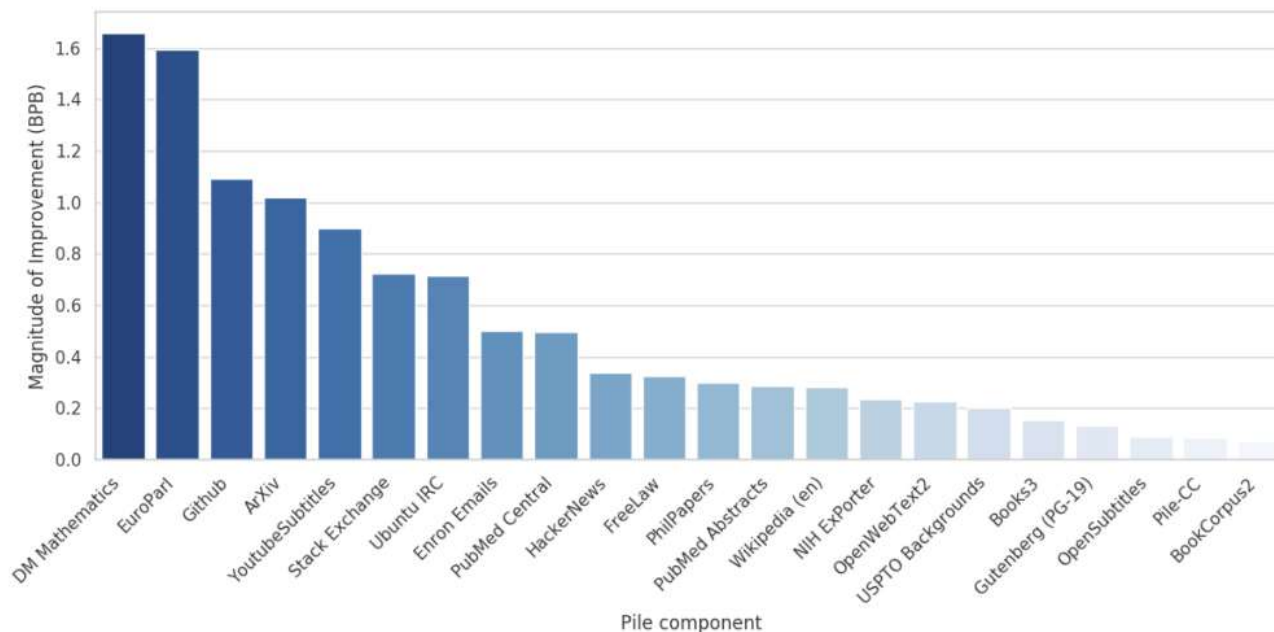


Figure 4: Magnitude of BPB improvement of Pile model over CC-100 model on each test set.

Evaluation using the Pile Dataset

- The result of training GPT=2/3 using the Pile dataset satisfies the scaling law.

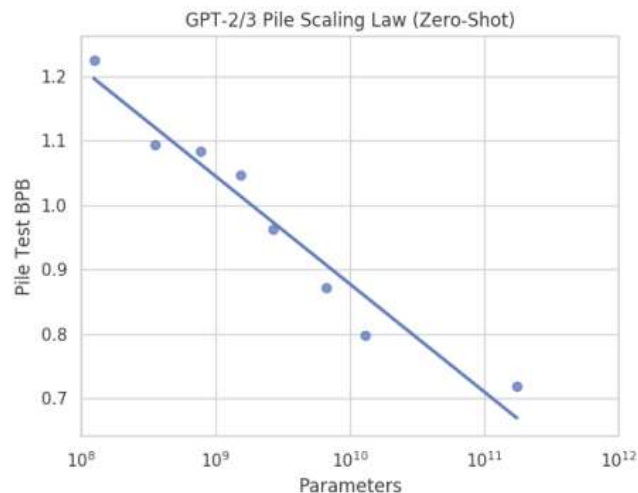


Figure 2: Scaling law for performance of GPT-2/3 models. ‘Zero-shot’ refers to the fact that none of the models have been fine-tuned on data from the Pile.

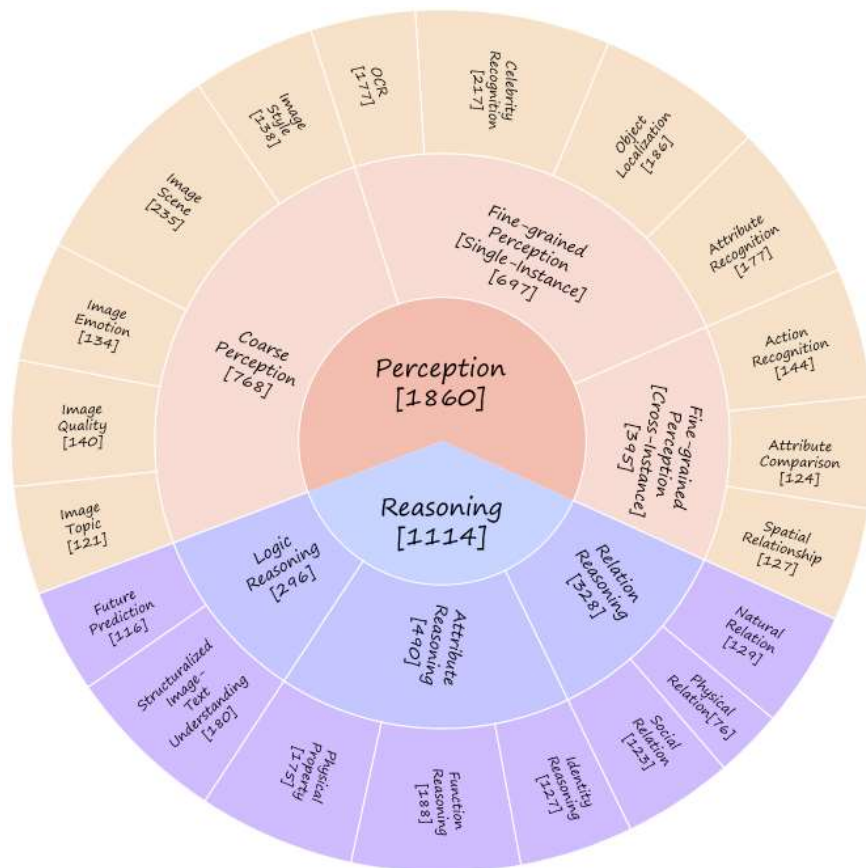
Paper Reading: Multi-modal Evaluation

MMBench: Is Your Multi-modal Model an All-around Player?

Motivation

- No proper benchmark for evaluating large vision-language model (limited data volume and task number).
- Existing benchmarks often use fixed label, e.g., bike \neq bicycle.

Task Formulation: Perception and Reasoning



MMBench: Is Your Multi-modal Model an All-around Player?

Evaluation Procedure



The original VL problem:

Q: How many apples are there in the image?

A. 4; B. 3; C. 2; D. 1

GT: A

Circular Evaluation

4 Passes in Circular Evaluation (choices with circular shift):

1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗
4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

- If the answer of VLM does not match any choice, ChatGPT is asked which choice this answer corresponds to.

Paper Reading: Let us incorporate more evaluation tasks

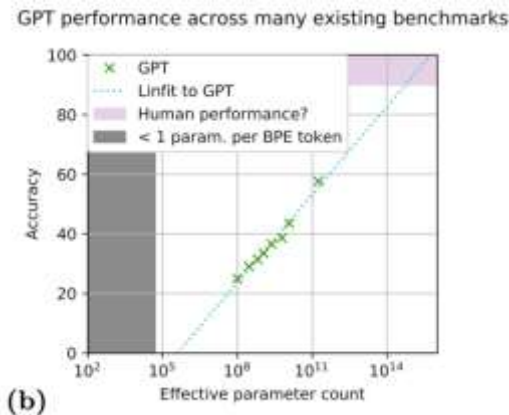
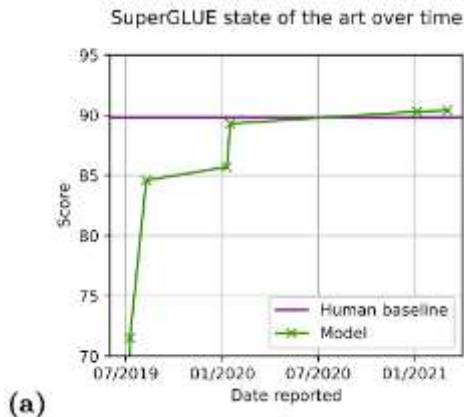
Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Motivation

- Existing benchmarks only contain a limited number of tasks and focus on a few capabilities.
- Only massive evaluation data is insufficient. We also need enough testing tasks, including both low-level tasks (like subword tokenization) and high-level tasks (like reasoning).

Limitations of Existing Benchmarks

- Only test a few capabilities of LLMs, such as language understanding, summarization, etc.
- Some benchmarks have short lifespans because they are too easy, e.g., superhuman performance is achieved after 18 months the SuperGLUE benchmark was released.
- Labeling quality problem due to the expensive cost of labeling by human experts.



- Alphabetic author list:^a

[illegible]

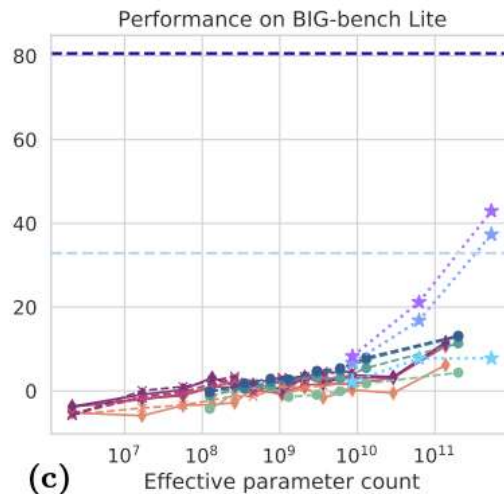
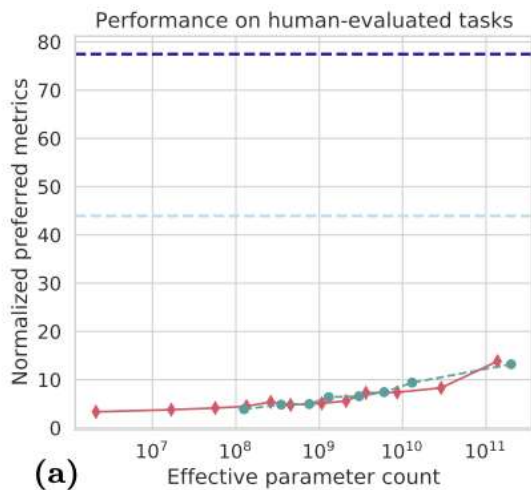
Laketa, Yanggja Song, Yasaman Balot, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hui, Yuhang Huo, Yuxiao Bai, Zachary Seld, Zhenye Zhao, Zijian Wang, Zijin J. Wang, Ziru Wang, Ziyi Wu

Overview

- **BIG-bench focuses on challenging tasks that are believed to be beyond the capability of LLM.**
- **A series of LLM, including OpenAI's GPT models, Google internal dense transformer architectures, and Switch-style sparse transformers are tested on BIG-bench.**
- **A team of human experts are employed to perform all tasks to provide a strong baseline.**
- **A small subset of Big-bench, namely Big-bench Lite, is constructed to conduct quick evaluation.**

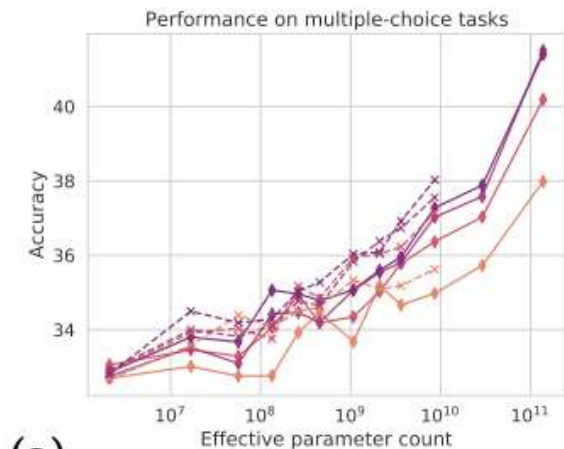
Finding 1

- Current LLMs still behave much worse than human.

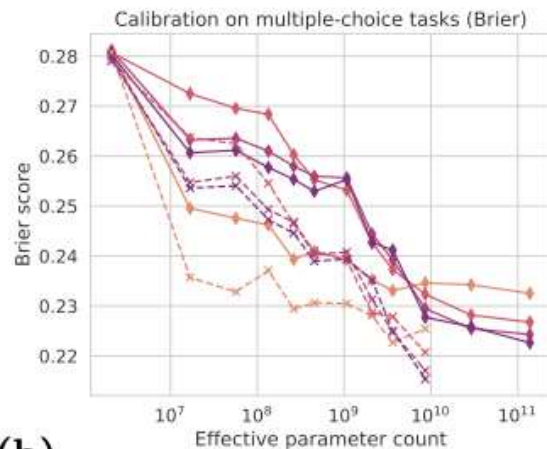


Finding 2

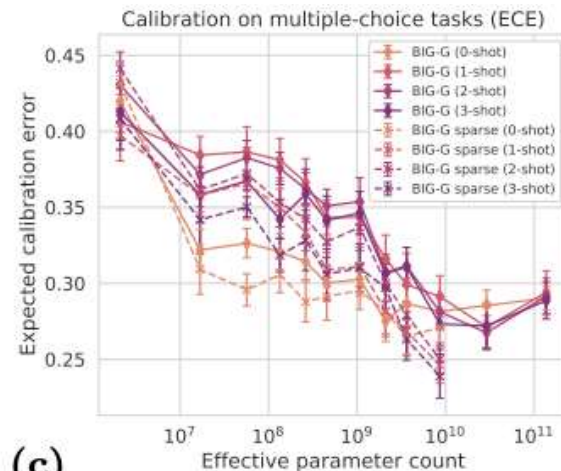
- Calibration improves as the growth of the model volume.



(a)



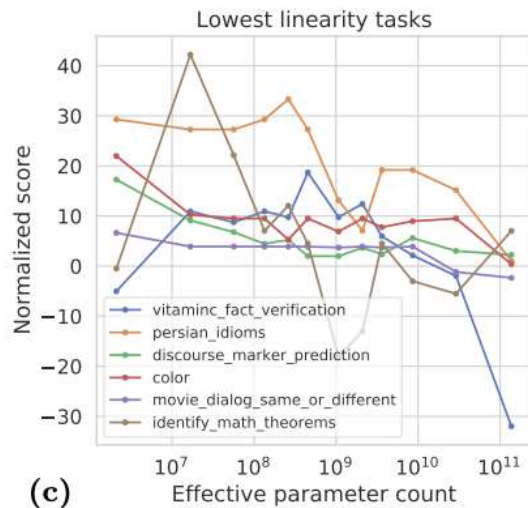
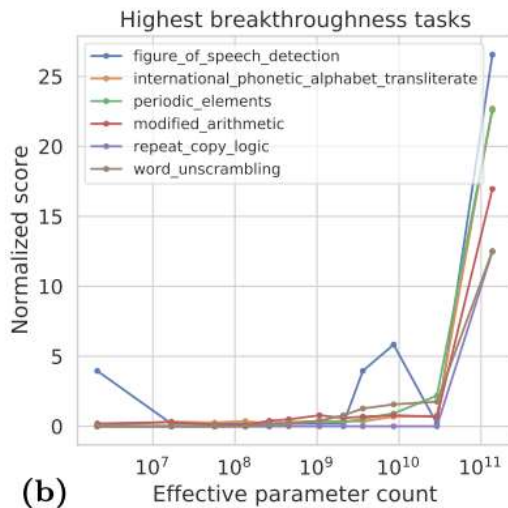
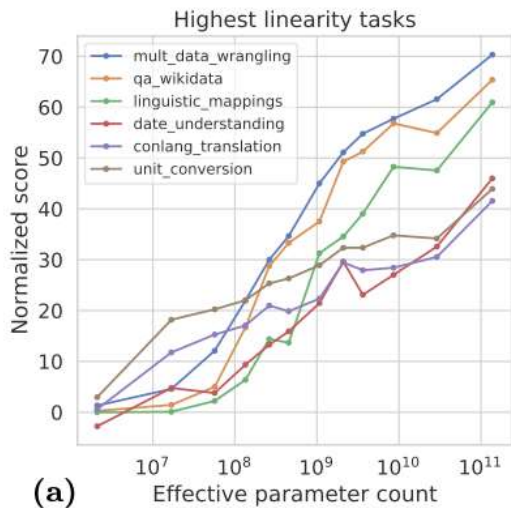
(b)



(c)

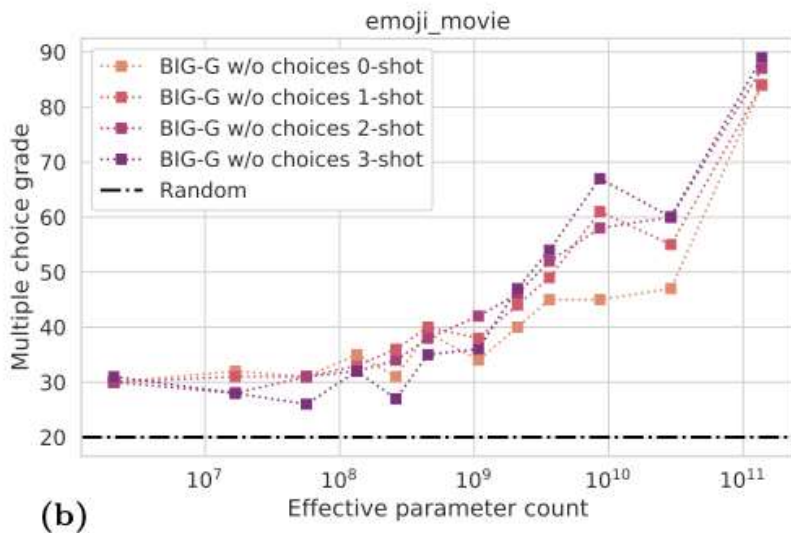
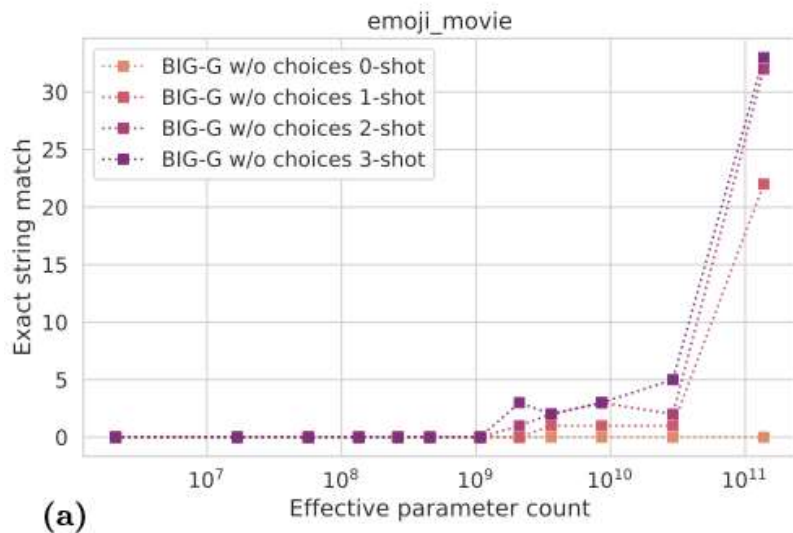
Finding 3

- Not all tasks obey the scaling law or shows breakthroughness.



Finding 4

- Whether capability emergence happens is significantly influenced by the selection of metric.



Task Examples

Traditional NLP tasks	Description
contextual question-answering	identifying the meaning of a sentence
context-free question answering	responses rely on model's knowledge base, but not on context
summarization	summarizing a block of text
memorization	memorization of data from the pre-training set.

logic, math, code	Description
mathematics	mathematics of any type
semantic parsing	parse semantics of natural-language utterances
decomposition	break problems down into simpler subproblems

Task Examples

understanding the world	Description
causal reasoning	reason about cause and effect
common sense	make judgements that humans would consider “common sense”
visual reasoning	solve problems that a human would be likely to solve by visual reasoning

understanding humans	Description
social reasoning	interpret or reason about human social interactions
gender prediction	the implicit gender information when prompted with gender-specific terms
intent recognition	the intent of a user utterance

Paper Reading: How to construct a dataset

Scaling Laws and Interpretability of Learning from Repeated Data

Motivation

- There often exists data repetition in a dataset because of complex data sources.
- No previous work studies how the data repetition problem affects performance.

Experimental Setting

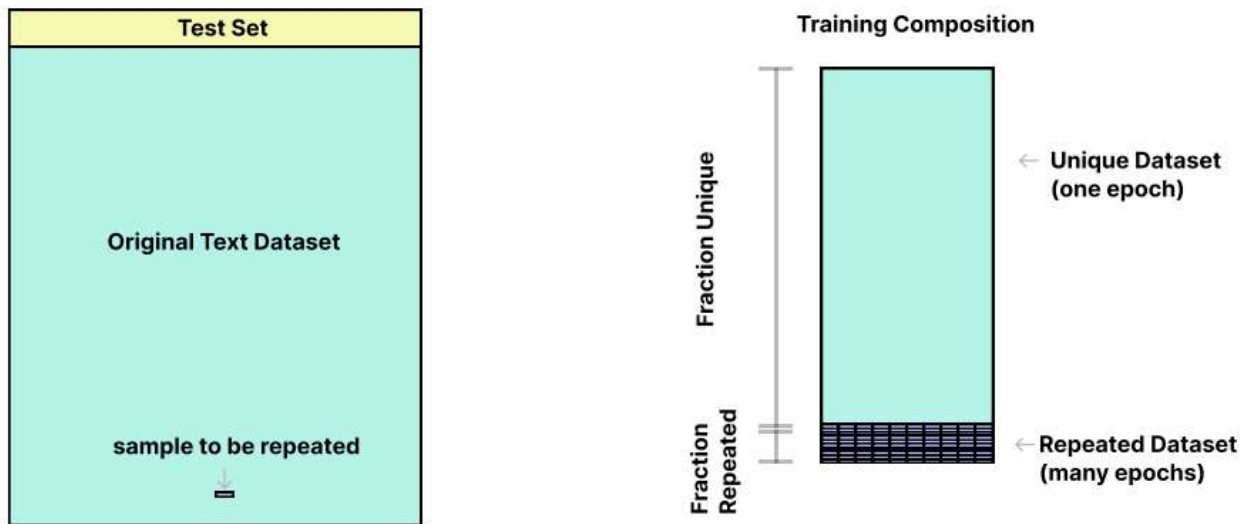
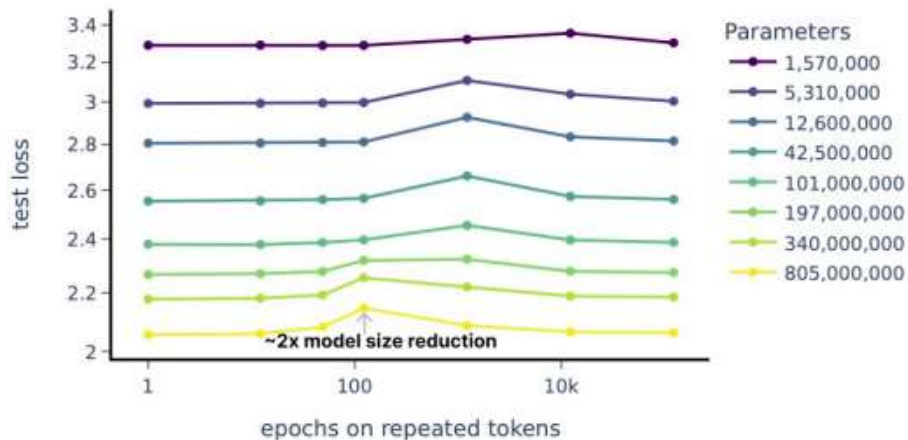


Figure 1 Experimental Setup. From a large original text dataset (left), we draw 90% of our desired training dataset in a non-repeated fashion, and 10% as repeats of a tiny portion of the original dataset (right). We hold constant that 10% of total training tokens will come from repeats, but we vary the repeated fraction in our runs. In other words, the sample to be repeated might be very small, like 0.01% of the total training tokens repeated 1000x, or relatively large, like 1% of the total training tokens repeated 10x. A small, held-back portion of the original dataset (yellow in left figure), not including any repeated data, is used as a test set and is the test loss reported in all subsequent figures.

Observation 1

- Repeating a specific ratio of tokens lead to serious performance recession, while repeating too many or too few tokens do not cause that phenomenon.



More Observations

- Repeated data can cause a divergence from power-law scaling.
- Repeated data causes a disproportionately large performance hit to copying, a mechanism for in-context learning.
- The disproportionate performance hit to copying coincides with a disproportionate degradation of induction heads.
- Repeated text data causes a small but still disproportionate performance drop out of distribution, as measured by cross entropy loss on Python code.
- One and two-layer attention only models trained on repeated data are worse at exactly copying and fuzzily copying (for instance correctly predicting Dursleys given that Dursley has appeared previously).
- Training on repeated Python code creates a similar behavior.
- Pre-training on repeated data damages models.

Reason Analysis

- There is a range in the middle where the data can be memorized and doing so consumes a large fraction of the model's capacity.

Questions

- Can we use the text produced by LLM to expand training data volume to further boost the performance of LLM?
- Can some automatic methods can be designed to construct a high-quality dataset based on noisy web-crawled data?
- How to balance the training of multiple tasks in LLM?

Questions & Discussion