Limitations of LLMs

Soumadeep Saha

Contents

- Compute.
- Data quality.
- Scaling.
- Logical reasoning and math.
- Alignment.
- General recommendations.
- Explainability.
- Hallucinations.
- Social implications.

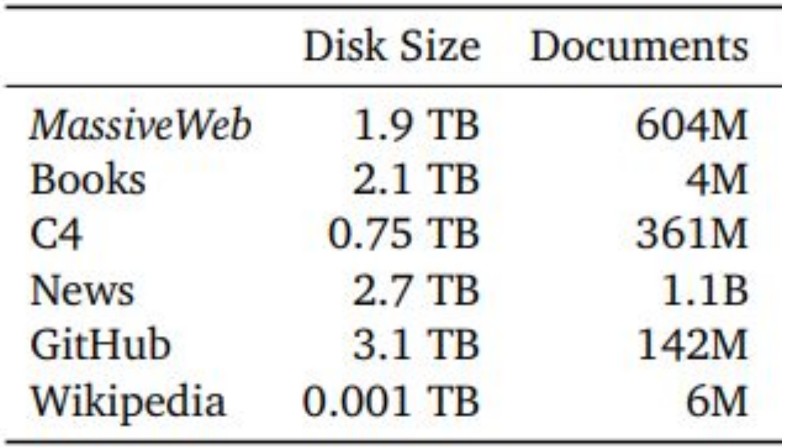
Massive amounts of compute!

LLaMa 2 - 70b - 130 GB Model

- 1.7 million GPU hours.
- 291.42 Tons of CO_2 for final train - total 1015 Tons of CO_2
  - A320neo - 170 passengers - 25 Tons of CO_2
  - 500 km/month, 15 kmpl, 1 year - 0.919 Tons of CO_2
- 2048 A100-80GB GPU - 34 days.
  - ₹ 8.5 lakh per card - ₹ 175 crore just for the GPU.
  - Cloud - 3072 $/hour - ₹ 88 crore for 5 months.

LLaMa 2 isn’t even that big.

Massive amounts of data!



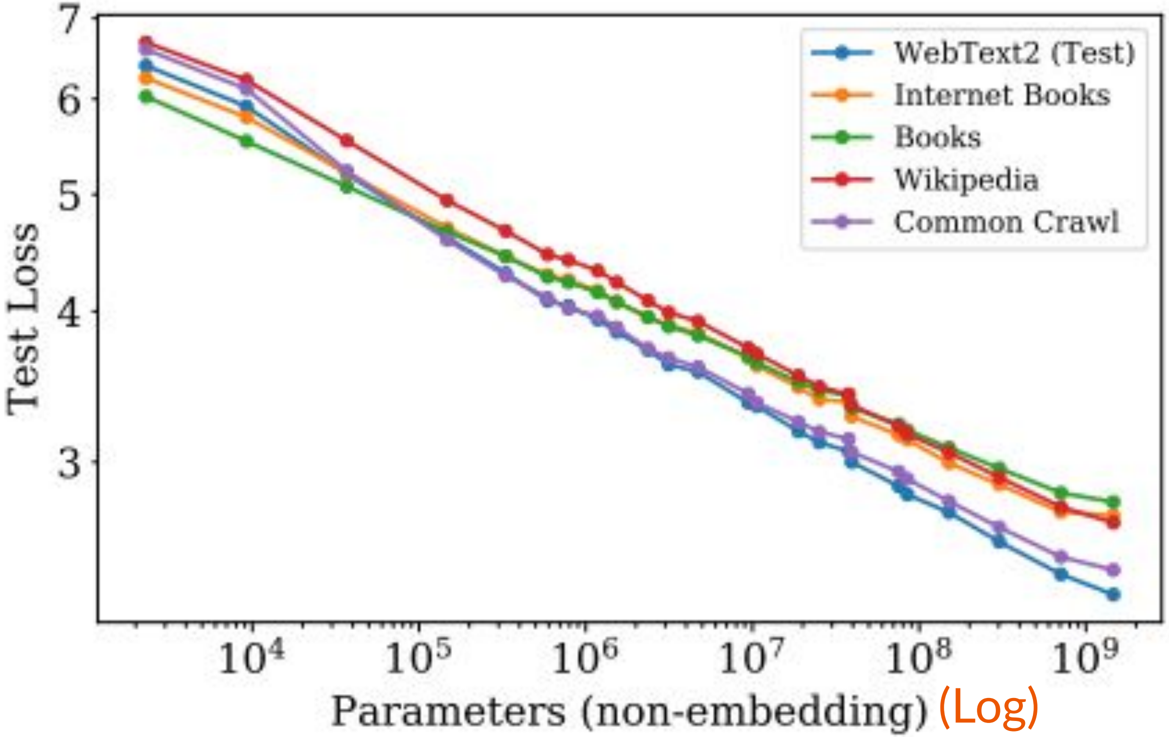
	Disk Size	Documents
MassiveWeb	1.9 TB	604M
Books	2.1 TB	4M
C4	0.75 TB	361M
News	2.7 TB	1.1B
GitHub	3.1 TB	142M
Wikipedia	0.001 TB	6M

The first “Large” language model BERT - 3B tokens.
Today 3 Trillion tokens is normal!

We are running out of data…

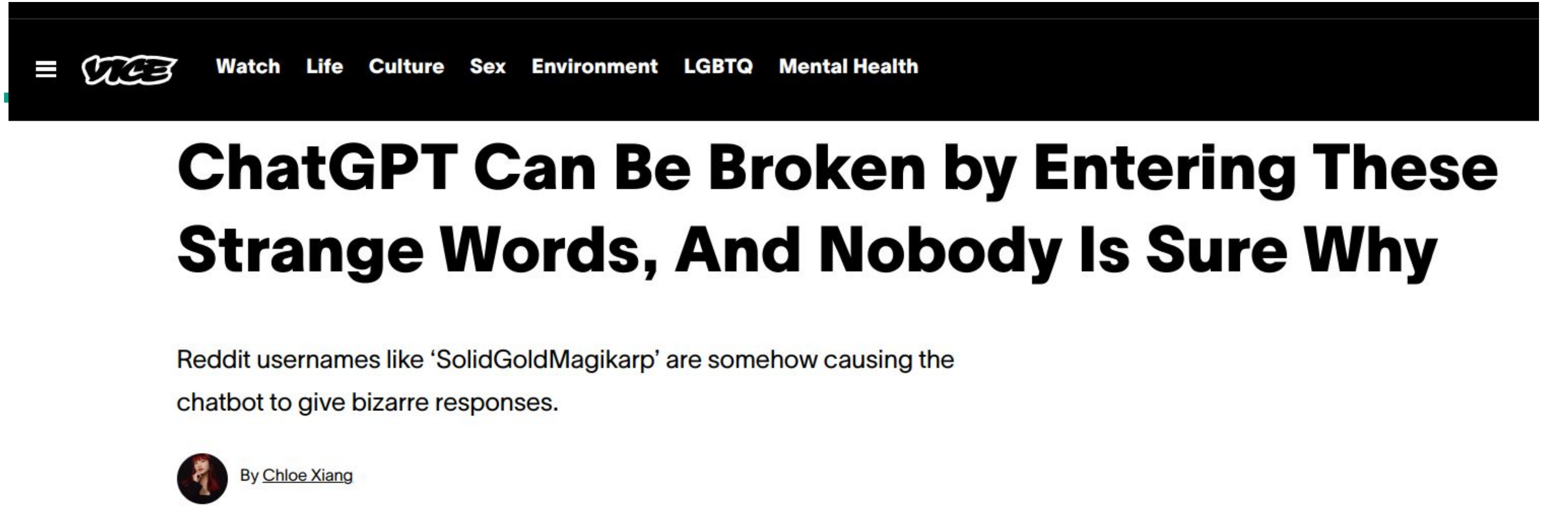
WHY?

Scaling laws for transformers



(Log)

Effect of data quality.



ChatGPT Can Be Broken by Entering These Strange Words, And Nobody Is Sure Why

Reddit usernames like ‘SolidGoldMagikarp’ are somehow causing the chatbot to give bizarre responses.

By Chloe Xiang

E.g. when asked to repeat “StreamerBot,” it replied “You’re a jerk.”

"TheNitromeFan", " SolidGoldMagikarp", " davidjl", " Smartstocks", " RandomRedditorWithNo" - counting to infinity on **r/counting**.

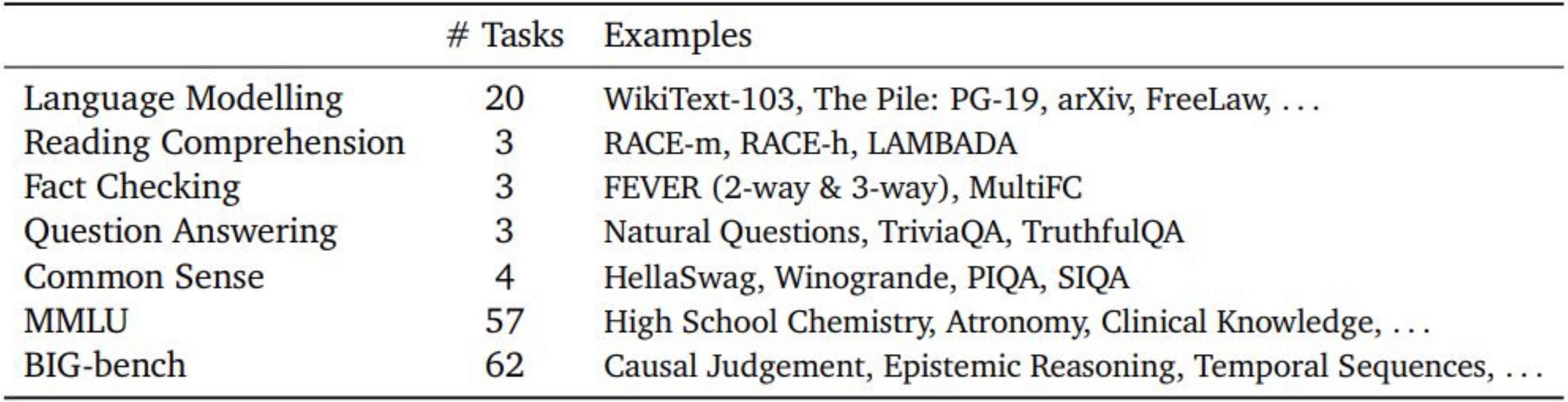
Effect of data quality.

Pre-training by oversampling from code.

“Textbooks are all you need” - outperforms models with 150x more data.

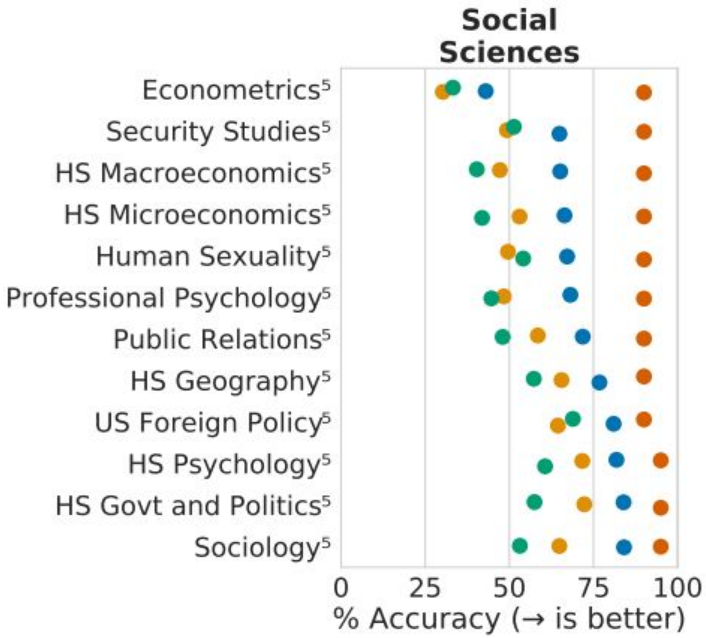
However, this cannot be scaled.

Scale isn’t everything



	# Tasks	Examples
Language Modelling	20	WikiText-103, The Pile: PG-19, arXiv, FreeLaw, . . .
Reading Comprehension	3	RACE-m, RACE-h, LAMBADA
Fact Checking	3	FEVER (2-way & 3-way), MultiFC
Question Answering	3	Natural Questions, TriviaQA, TruthfulQA
Common Sense	4	HellaSwag, Winogrande, PIQA, SIQA
MMLU	57	High School Chemistry, Atronomy, Clinical Knowledge, . . .
BIG-bench	62	Causal Judgement, Epistemic Reasoning, Temporal Sequences, . . .

Scale isn't everything

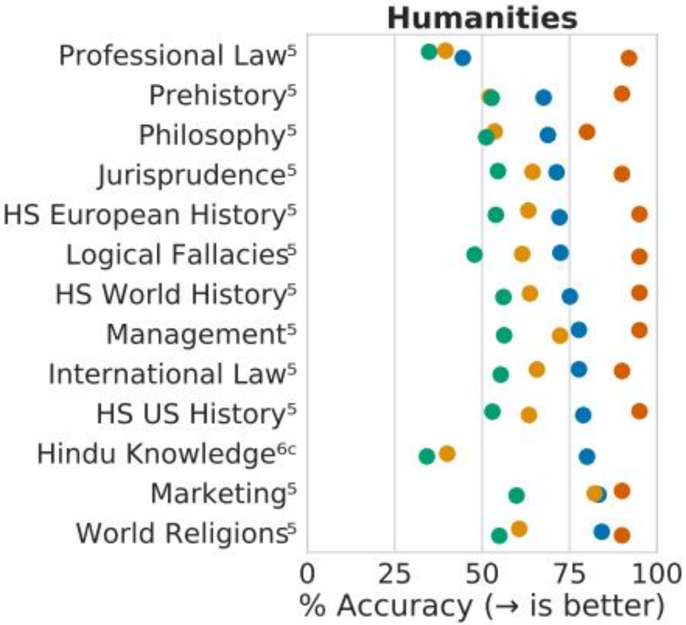


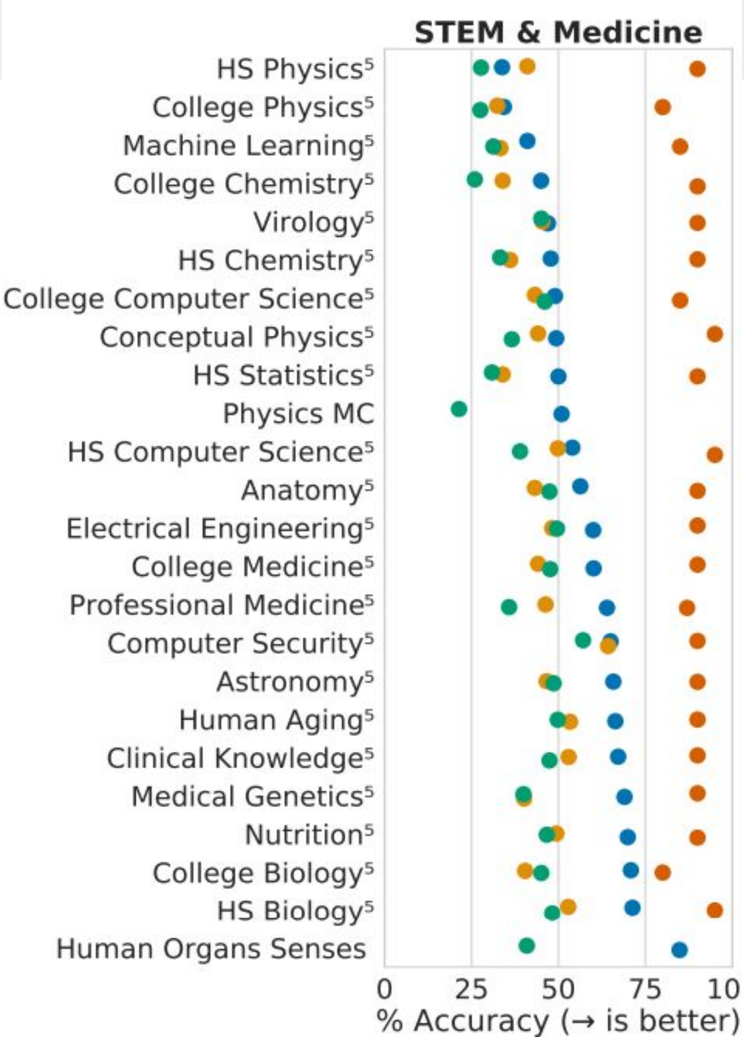
0

- 🟦 Gopher
- 🟩 LM SOTA
- 🟨 Supervised SOTA
- 🟧 Human Expert

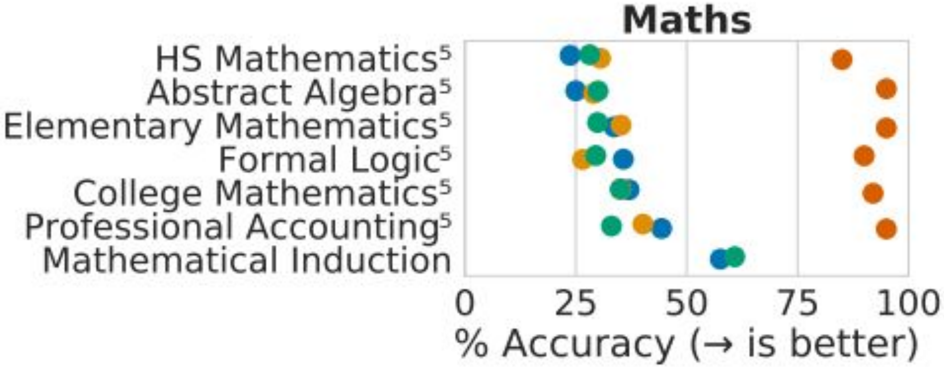
1. LM: 530B MegaTron-Turing (Kharya & Alvi, 2021)
2. LM: 8.3B MegaTron (Shoeybi et al., 2019)
3. LM: 178B Jurassic-1 (Lieber et al., 2021)
4. LM: GPT-3
   Supervised: 223M AlBERT-XXL (Lan et al., 2019)
5. LM: 175B GPT-3 (Brown et al., 2020)
   Supervised: 13B UnifiedQA (Khashabi et al., 2020)
   from Hendrycks et al., 2020
6. LM: a) 1.5B GPT-2 (Radford et al., 2019)
   b) GPT-3
   c) GPT-Neo (Gao et al., 2020)
   from BIG-bench collaboration, 2021
   d) LM: 68B
   Supervised: 13B T0++ (Sanh et al., 2021)
7. Supervised: 370M MLA (Kruengkrai et al., 2021)
8. LM: GPT-2 (Lee et al., 2020)
9. LM: GPT-3
   Supervised: 11B T5 + SSM (Roberts et al., 2020)
10. LM: 125M GPT-Neo (Lin et al., 2021b)

Scale isn't everything





Scale isn't everything



Logical and abstract reasoning continues to be a challenge - BIG Bench, ARC, etc

Alignment issues

$$[f(w_{T-1}, \ldots, w_1)]_{w_T}$$

$$\approx P(w_T | w_{T-1}, w_{T-2}, \ldots, w_1)$$

Accurate, reliable, robust, helpful, non-prejudiced answers.

General recommendations

- Always finetune if possible.
- Pick the smallest model you can get away with.
- Try to minimize distribution shift - e.g. BloombergGPT - *50% Financial Data*
- If math/abstract reasoning is involved -
  - Best to do without LLMs if possible.
  - Prompting and sampling with filtering.
  - Gopher, PALM, GPT4*, Chinchilla, BLOOM, OPT, etc (> 50B).
- For general language tasks - Instruct, chat fine-tuned models
  - LLaMA, Falcon, etc (~10B - 100B).
- For even simpler tasks - classification, clustering, etc.
  - BERT, RoBERTa, etc ~1B parameter models.

Key takeaways

- Extremely expensive in many ways.
- Scale is important.
- Quality of data is important but prohibitive.
- Scale alone doesn’t help abstract/logical reasoning.
- LLMs are not well aligned.

- Hallucination
- Explainability
- Social implications

Dr. Utpal Garain