AU24 CSE 5539 Presentation

GPT-3: Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et. al.

NeurIPS 2020

Paper: arxiv.org/abs/2005.14165

✍ **Sriram Sai Ganesh**

Authors

Ilya Sutskever
Co-inventor of AlexNet
Co-founder of OpenAI

Dario Amodei
Co-founder & CEO,
Anthropic

Alec Radford
ML @ OpenAI,
GPT 1, 2, 3 & 4, PPO

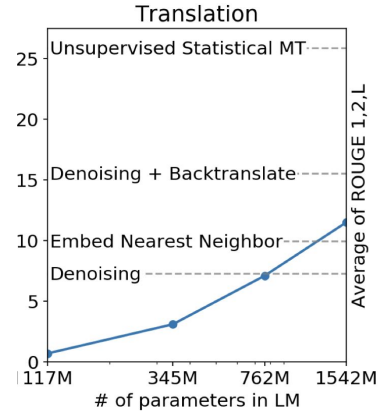
Aditya Ramesh
Scientist @ OpenAI
DALL·E, DALL·E 2

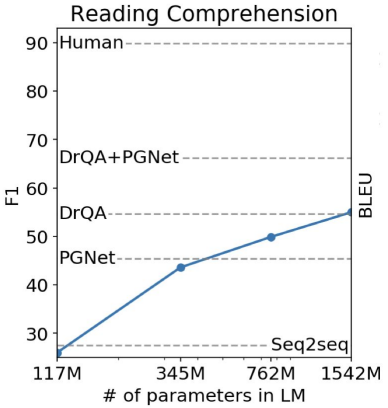
Introduction

- Previously, NLP research tended to:
  - Design task-specific model architectures.
  - Curate language representations & data to specific tasks.
- Recent paradigm shift –
  - Task-agnostic models.
  - Generalized pre-training & architectures.
- Final step (?) –
  - Adapting these task-agnostic models to specific tasks.

How necessary is finetuning?

- Prior work shows:
  - A single pre-trained model has good zero-shot performance. Not SoTA…yet.
  - Performance scales with parameter count* (!)
- Contributions of this work:
  - Empirically test performance scaling, ranging up to **175B parameters (GPT-3.)**
  - Clarify and systematize **"in-context learning."**
  - **Promising** experimental results.



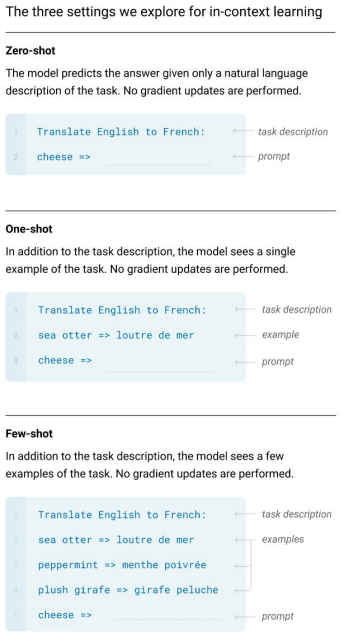


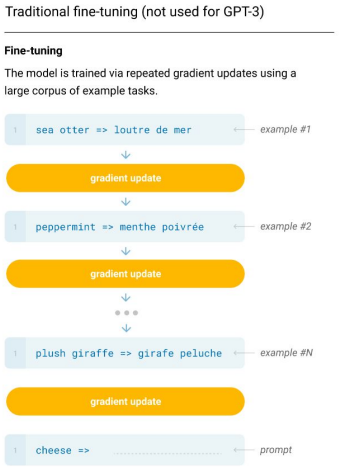
*within experimental constraints

Approach

- **Fine-tuning: update weights** based on data.
  - + Good benchmark performance.
  - − Poor OOD generalization.
- **Few-shot:** task description along with ***K* examples** of samples/completions.
  - + Major reduction in task-specific data.
  - − Worse performance than SoTA *(so far.)*
- **One-shot:** few-shot with **K=1.**
- **Zero-shot:** Task *description* only, **K=0.**

Learning Settings

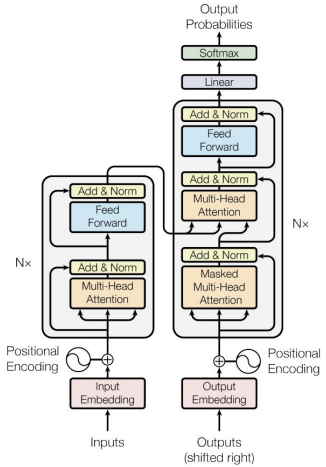




Approach

- **Architecture:**
  - Identical to GPT-2, except for transformer attention pattern.
  - 8 different model sizes – 125M to 175B
  - Model & data partitioned across GPUs to efficiently handle memory constraints
- **Training Dataset:**
  - Filtered CommonCrawl
  - Deduplication to prevent redundancy & ensure integrity of held-out validation set.
  - Augmented with reference corpora: WebText, Books1 & 2, English Wikipedia.

Model & Dataset



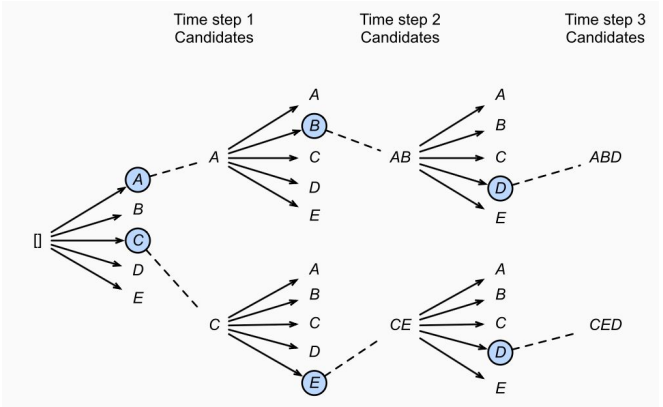
Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Approach

- **Training Process:**
  - Model parallelism both within each matrix multiply & across layers.
- **Evaluation:**
  - One/Few-shot: draw *K* samples from training or dev set as conditioning.
  - Some tasks – additional natural language prompt.
  - Results reported on test set when possible.

Training & Evaluation

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}



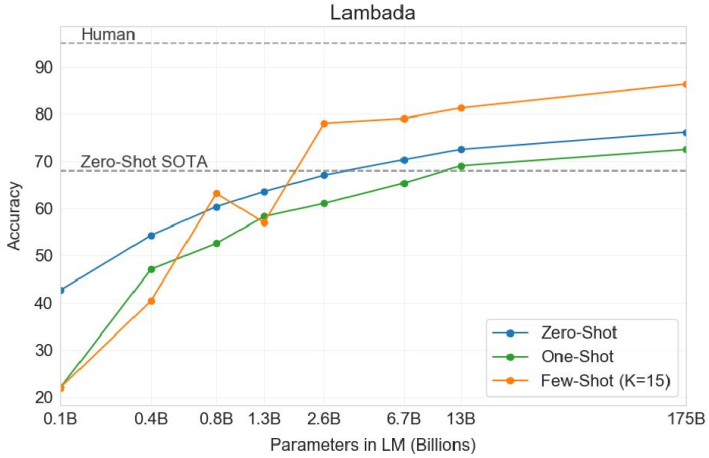
✍ Sriram Sai Ganesh

Results

Language Modeling & Cloze

- Penn Treebank:
  - New SoTA by 15 points.
  - Zero-shot perplexity of 20.5 on POS labeling.
- LAMBADA:
  - Predicting terminal word in a sentence/paragraph.
  - Framed in a few-shot setting – 86.4% (+18%).
  - One-shot – not as effective.
- HellaSwag & StoryCloze – lower than fine-tuned SoTA.

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3



Results

QA & Translation

- Closed-book (no document/info access)
  - GPT-3 nears or exceeds SoTA pre-trained and fine-tuned RAG models on 2 datasets.
  - ARC multiple choice – approaches baselines; much worse than SoTA.
- Reading comprehension – approach human baselines but worse than SoTA NNs.
- Translation:
  - Underperforms SoTA on 0-shot.
  - Few-shot – approaches SoTA when translating to En.

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Setting	ARC (Easy)	ARC (Challenge)	CoQA	DROP
Fine-tuned SOTA	**92.0**[a]	**78.5**[b]	**90.7**[c]	**89.1**[d]
GPT-3 Zero-Shot	68.8	51.4	81.5	23.6
GPT-3 One-Shot	71.2	53.2	84.0	34.3
GPT-3 Few-Shot	70.1	51.5	85.0	36.5

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

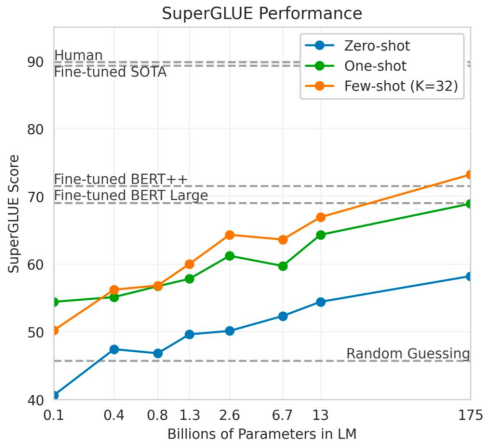
Results

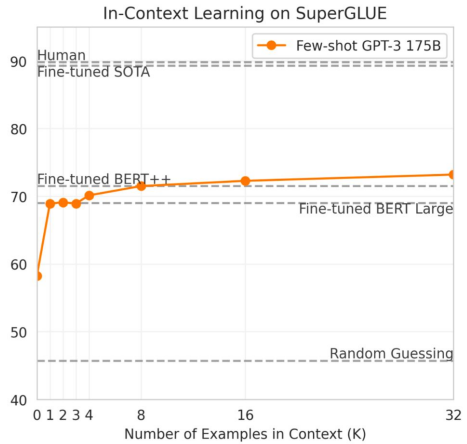
SuperGLUE

- A standardized collection of datasets.
- Few-shot results –
  - Steady improvement through K=32.
  - Large variance in GPT-3 performance.
  - Weak at comparing sentences

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

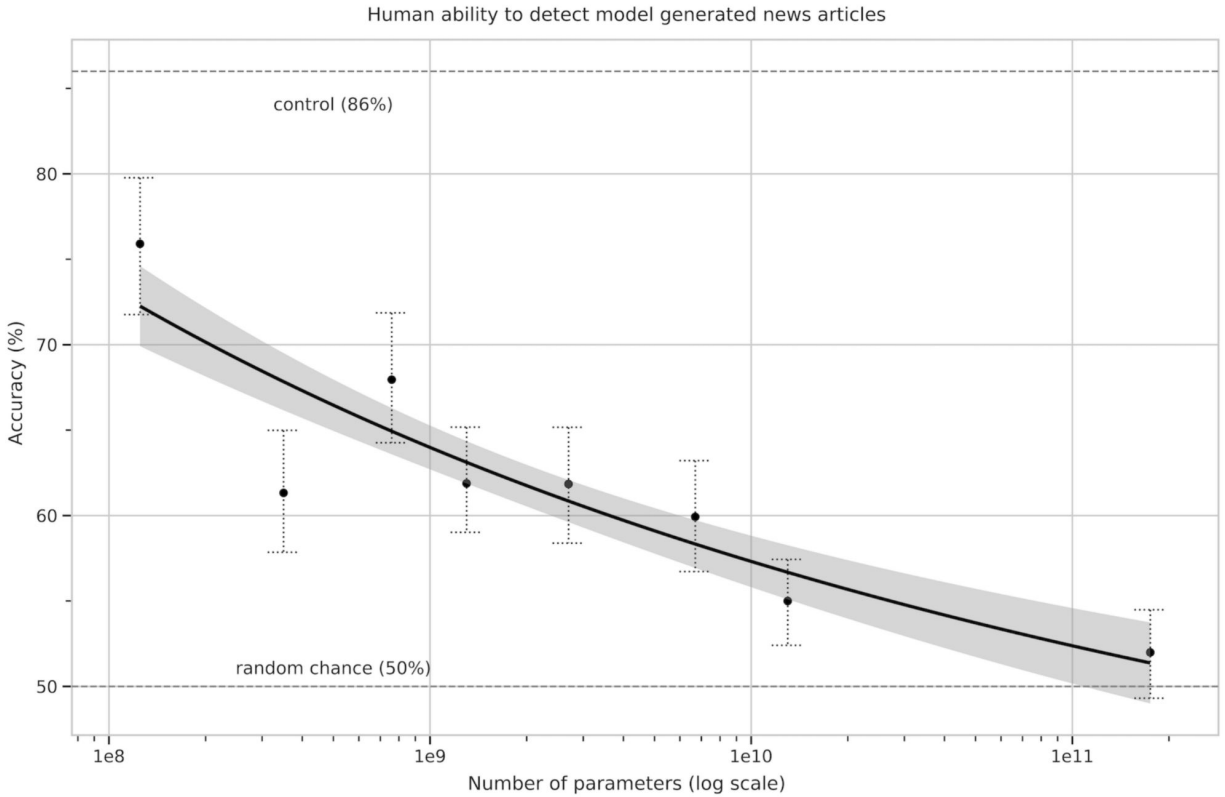




✍ Sriram Sai Ganesh

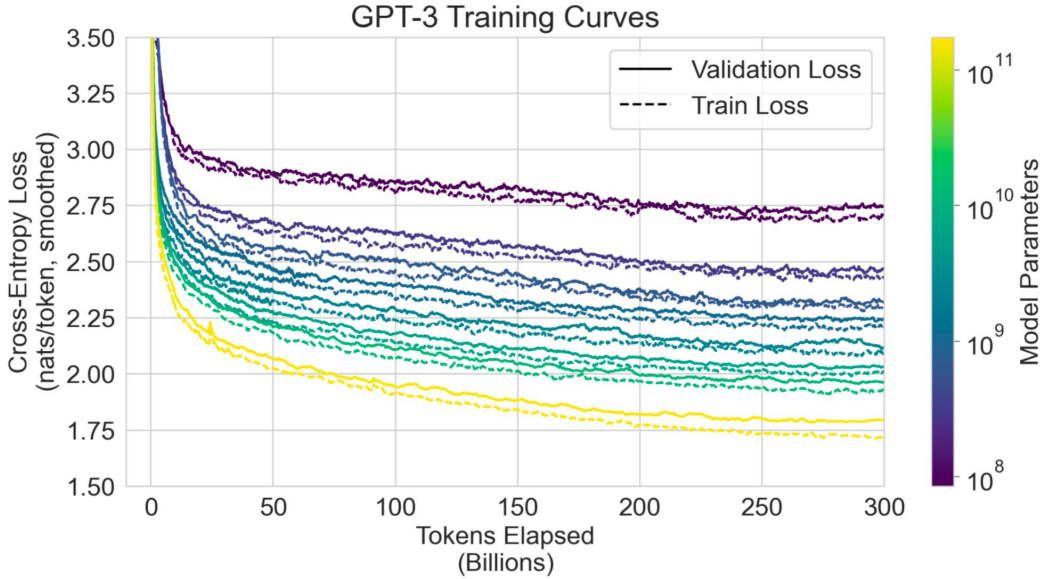
Results

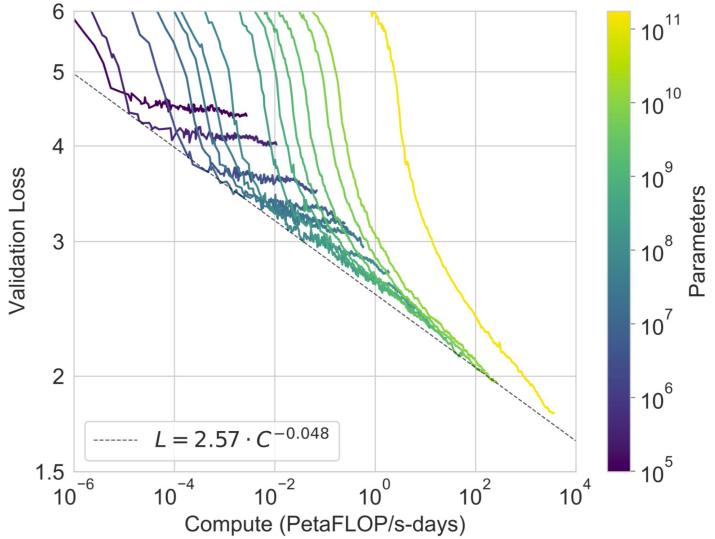
Misc.



Results

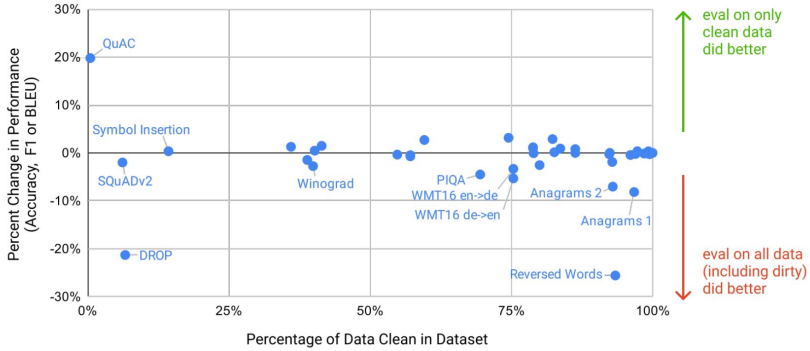
Misc.

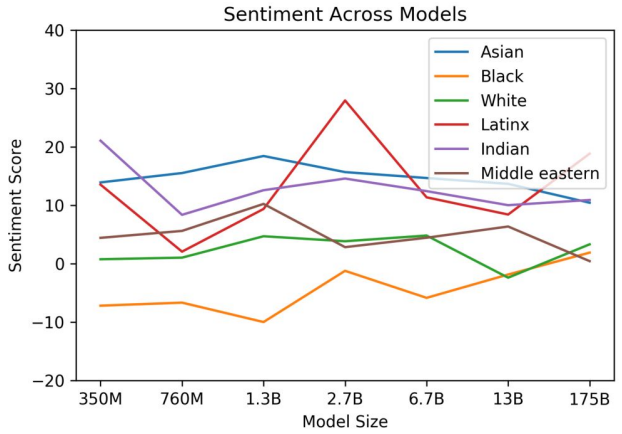




Limitations & Conclusion

- Potential *test set contamination* from the internet-scale dataset.
- Model limitations:
  - Semantic self-repetition.
  - Weakness at “common-sense” and comparative tasks.
  - Lack of interpretability.
  - Poor sample efficiency.
  - What does ICL actually do?
- 175B model; towards general language systems; empirical scaling results; ethical considerations.





Thank you!

Questions?

GPT-3: Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et. al.

Paper: arxiv.org/abs/2005.14165