GPT-2

Language Models are Unsupervised Multi-task Learners

Kejian Shi

Jan.24, 2023

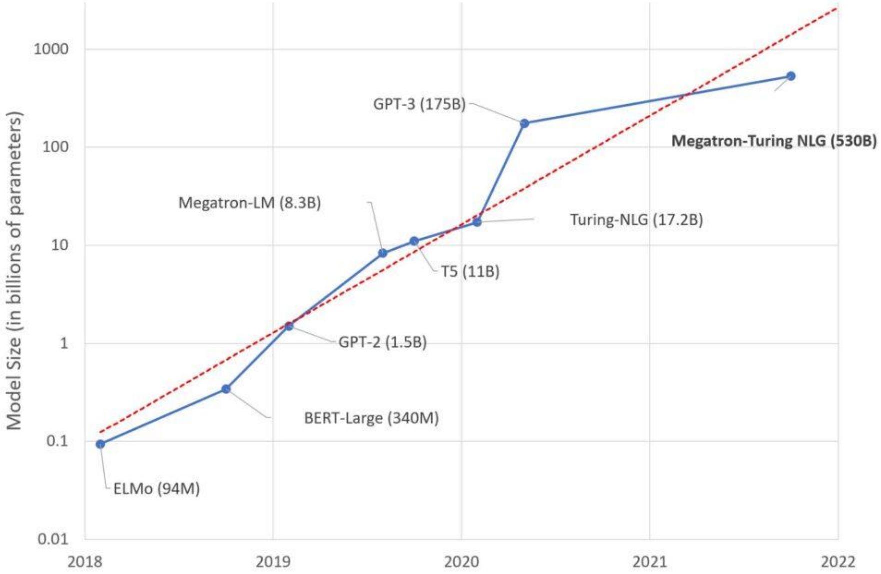
📖 Outline

- Overview & Background
- Architecture
- Training & Data
- Experiments
- Limitations
- Discussion

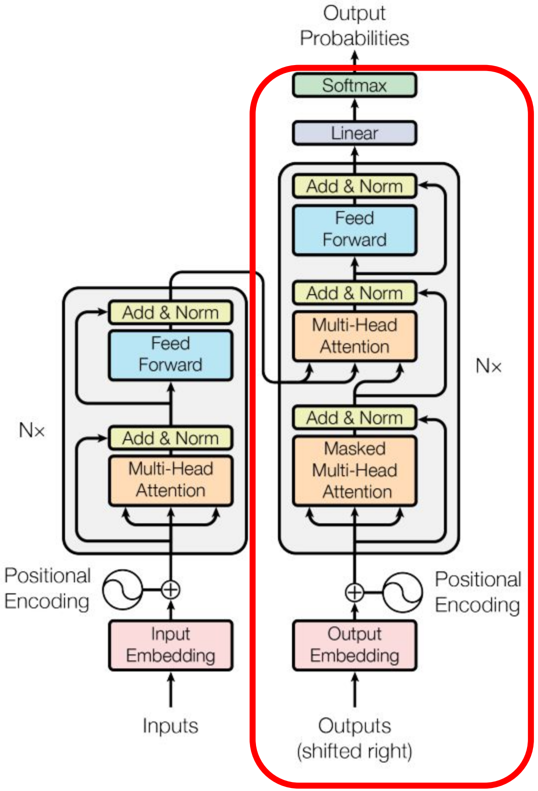
Est. time: 15 min talk + 15 min discussion

🚀 Overview & Background

- Lack of Generalization of SOTA.
- Self-supervised learning.
- Respond to BERT paper.
- Zero-shot transfer.

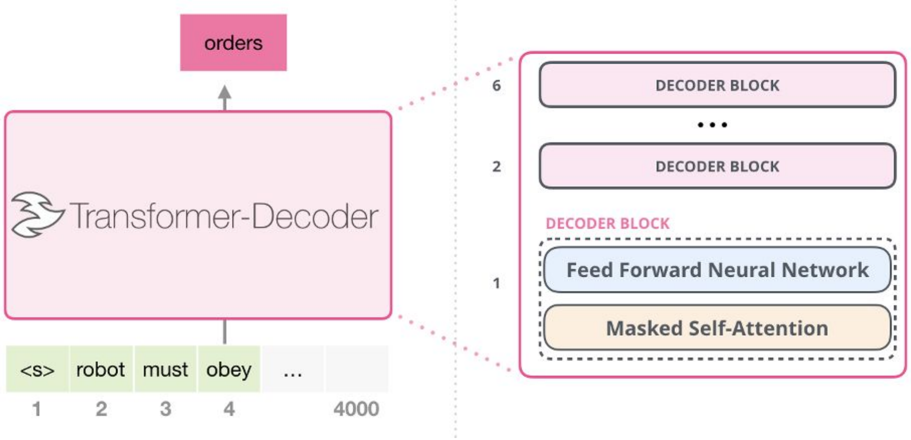


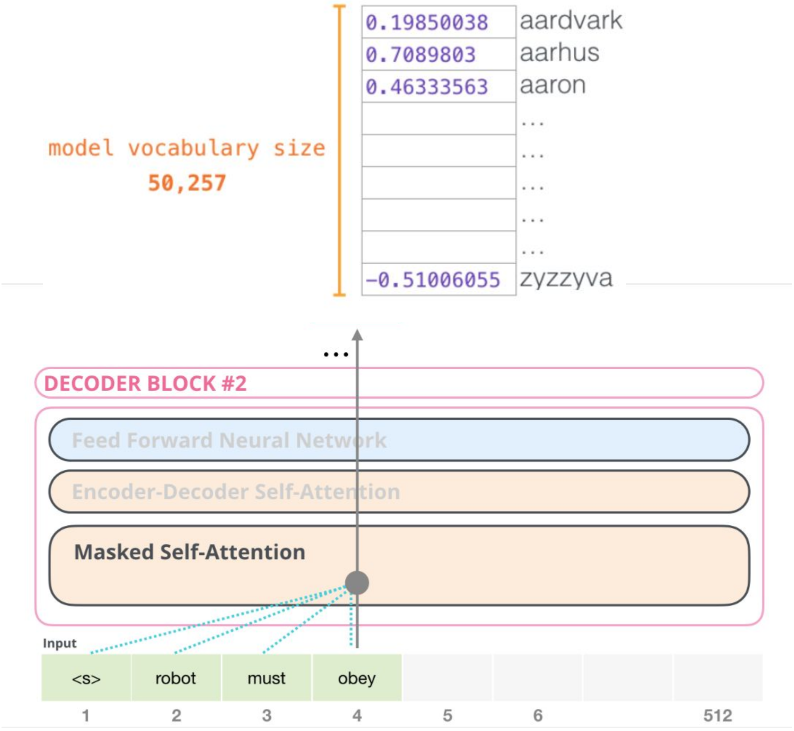
Architecture



🔨 Architecture

- Transformer Decoder
  - 12 ~ 48 layers
- Masked Self-Attention
  - Compute efficient
- Sampling



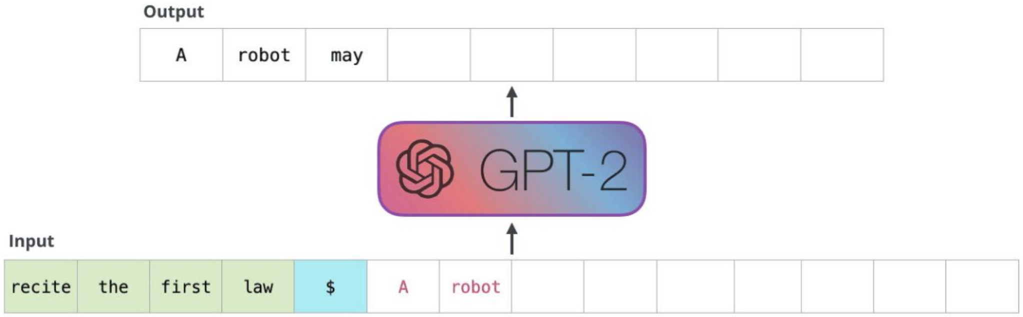


0.19850038	aardvark
0.7089803	aarhus
0.46333563	aaron
	...
	...
	...
	...
	...
-0.51006055	zyzzyva

🎯 Objective

- Language modeling objective
  - i.e. NLL, MLE objective
  - “Predict the next word”
- vs. MLM (BERT)

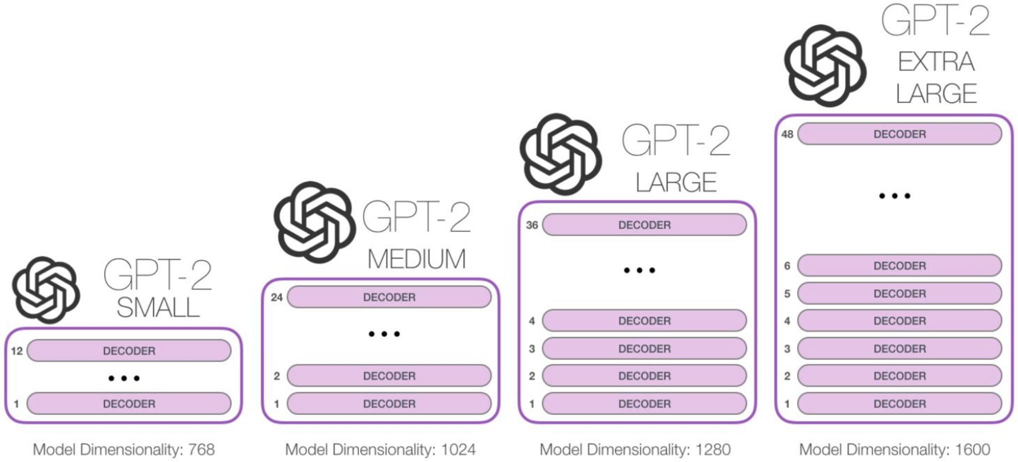
$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

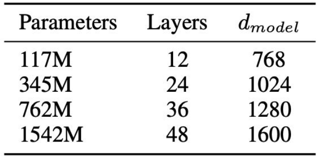


$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

🔨 Architecture

- 4 model variants.
- Compare to GPT-1
    - 10x parameters (XL model)
    - Vocabulary: 40k → 50k
    - Context size: 512 → 1024
    - Batch size: 64 → 512
    - Layer norm position
    - Weight initialization



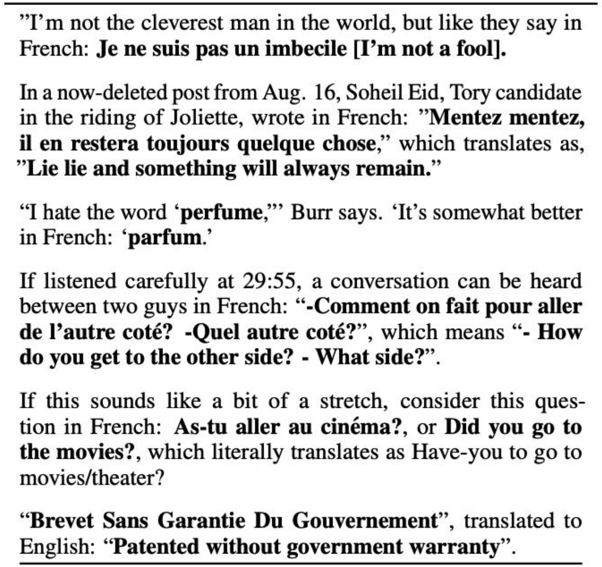


Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

Data

- Data plays a big role.
- WebText
  - ~ 8 million web pages scraped and filtered from reddit.
    - (> 3 "karma" upvote score)
  - wikipedia explicitly filtered out.
- 40GB of data.
- Naturally occurring task-related data



”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

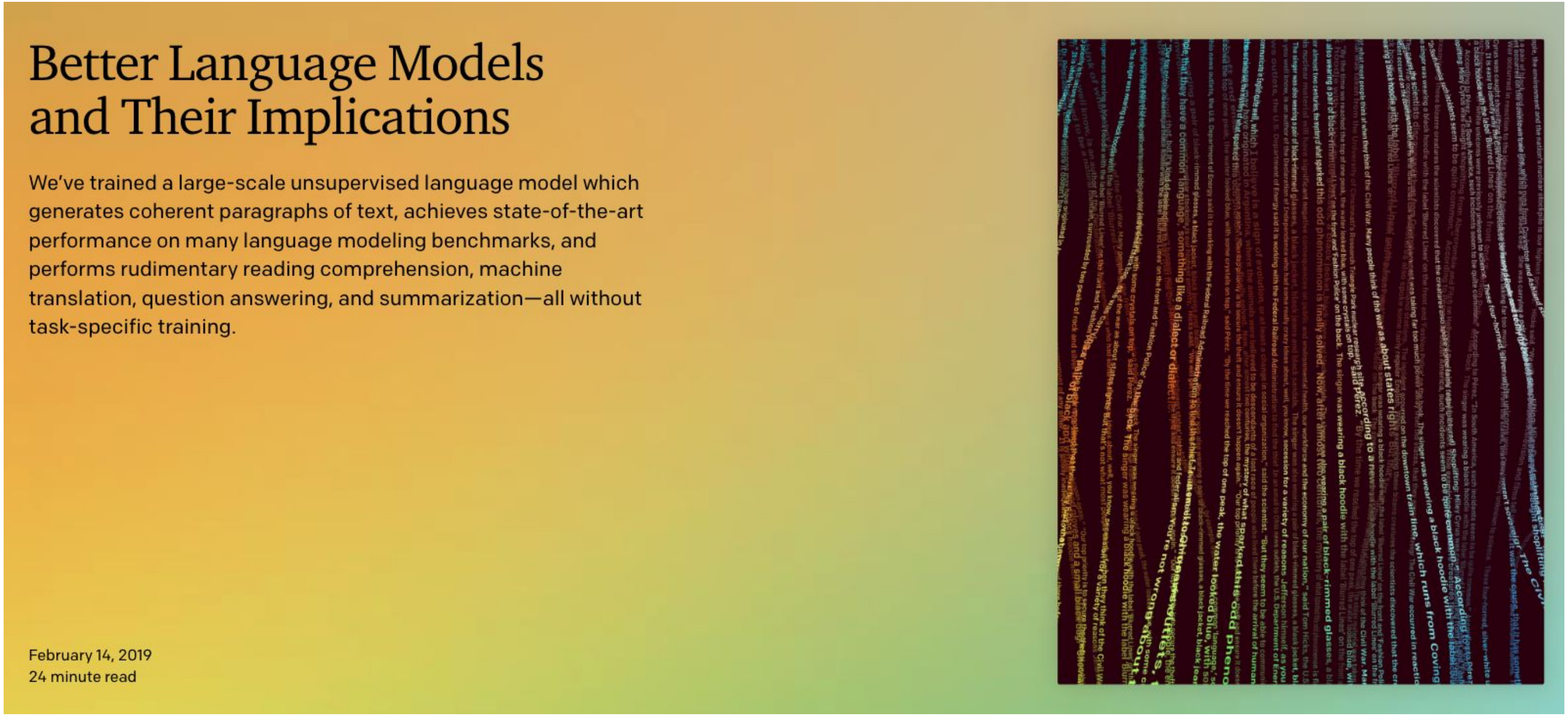
“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

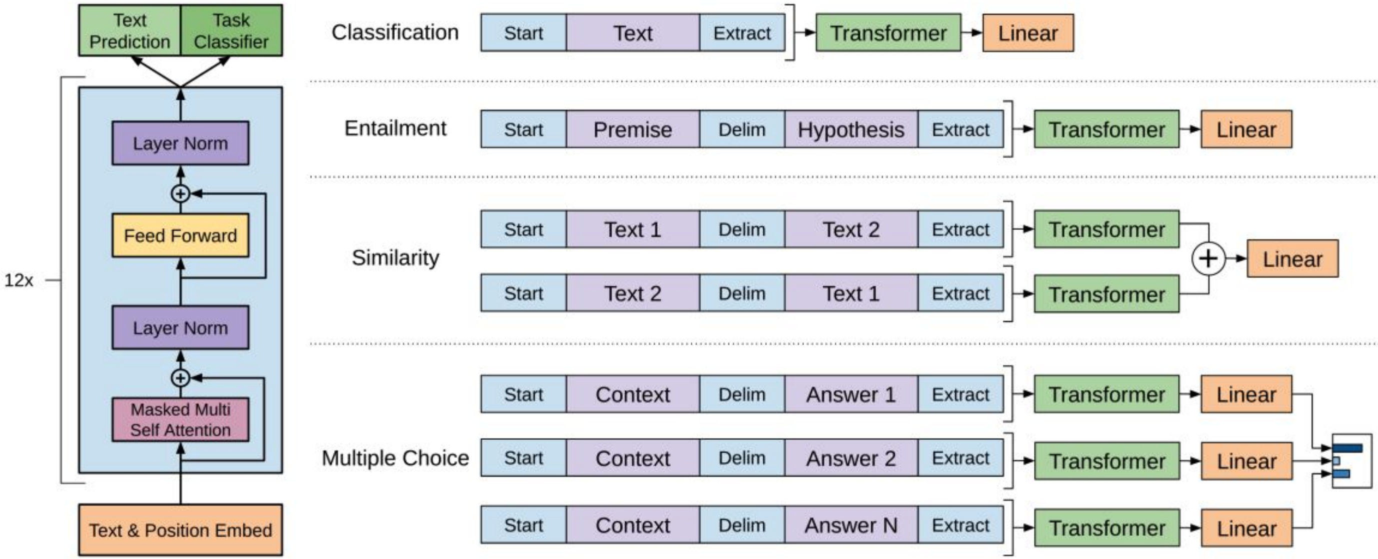


Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

February 14, 2019
24 minute read

Input Formulation (GPT-1)



📝 Input Formulation (GPT-2)

NMT: “Translate to french,” <English text>, <French text>.

QA: “Answer the question,” <Document>, <Question>, <Answer>.

SUMM: <Document> “TL; DR:” <Summarization>

…

ANTHROP\C

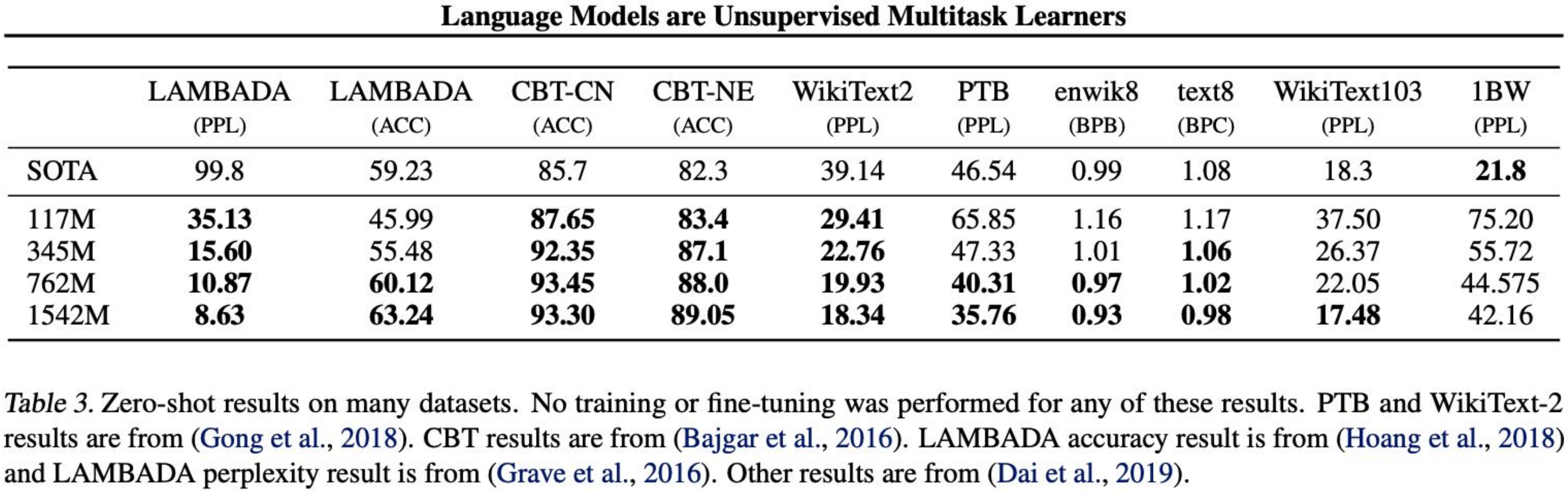
Prompt Engineer and Librarian

SAN FRANCISCO, CA / PRODUCT / FULL-TIME / HYBRID

📊 Experiments

- Zero-shot domain transfer (Language modeling tasks).
  - LAMBADA, CBT …
- Zero-shot NLU tasks.
  - MT, QA, CLS, WSC …
- Misc.
  - Underfit WebText.
  - Model Memorization.

Experiments



Language Models are Unsupervised Multitask Learners

in

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Experiments

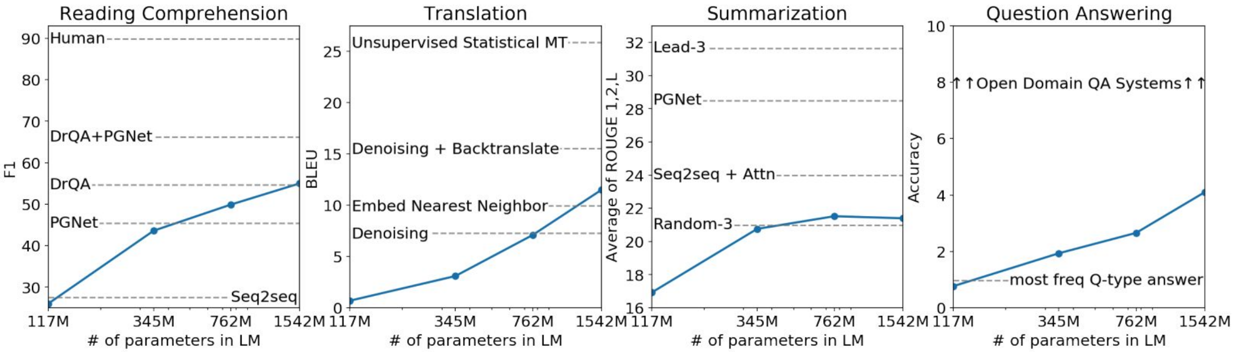


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

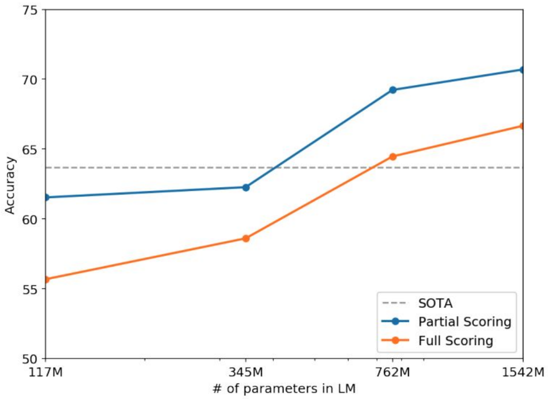
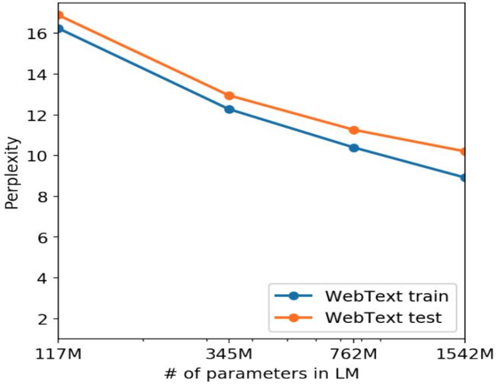


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

📈 Results

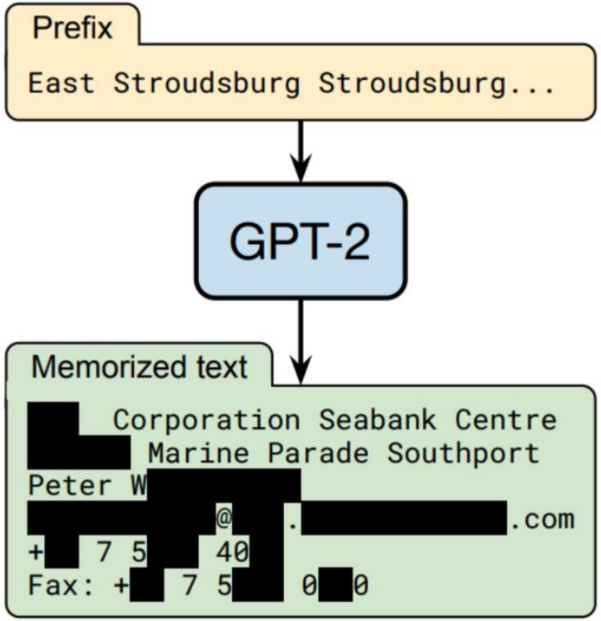
- GPT-2 improved the then existing state-of-the-art for **7 out of 8** language modelling datasets in zero shot setting.
- “Larger the better”
  - Underfitting on WebText.



  - The performance increases in log-linear fashion as model scales.
  - Building even larger language models would reduce the perplexity and make language models better at natural language understanding.

🚨 Limitations

- Limited performance
  - E.g. Summarization
  - Inefficiencies of uni-directional representations.
    - ! given the model size, data, and compute
    - *"The zero-shot performance of GPT-2 is still far from usable.”*
- Memorization / Safety issues
  - Verbatim memorization of private or IP information
  - (Carlini et al. 2020): *"We find that at least 0.1% percent of its text generations contain long verbatim strings in its train set."*



```
East Stroudsburg Stroudsburg...
```

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Thank you !

🤔 Discussion

- Anything I missed? Corrections?
- Discussion points
  - How do we understand and distill what’s learned by the model?
  - …