Language Models are Unsupervised Multitask Learners – SHORT SUMMARY

Current NLP tasks rely on supervised learning with specialized datasets. This work demonstrates that large language models, when trained on massive, diverse text corpora like WebText, can exhibit capabilities in these tasks without explicit task-specific instruction (zero-shot learning). The model capacity plays a crucial role in this success, with larger models achieving significantly better performance across various tasks. Notably, their largest model, a 1.5-billion parameter Transformer (GPT-2), achieves state-of-the-art results on 7 out of 8 benchmark datasets in a zero-shot setting. However, further scaling appears necessary to fully exploit the potential of WebText. These findings suggest a promising new direction for developing NLP systems that can learn complex tasks by leveraging naturally occurring demonstrations within large text collections.

Introduction

Current Natural Language Processing systems excel at specific tasks but lack the generalizability of human language understanding. This limitation likely stems from their training on isolated datasets designed for a single task and domain. To achieve more versatile NLP systems capable of handling diverse tasks without handcrafted training data for each, a shift towards training on a broader range of domains and tasks is necessary.

Multitask Learning as a Promising Direction: Multitask learning, a framework where a model learns from multiple tasks simultaneously, presents a promising path to improve generalization. Benchmarks like GLUE and decaNLP have already been proposed to facilitate research in this area.

However, current implementations of multitask learning in NLP are in their early stages. Existing approaches likely require a substantial number of training examples per task, mirroring the data requirements of single-task learning. Scaling current techniques to meet this demand might be impractical. Therefore, exploring alternative setups for multitask learning in NLP is crucial to overcome these limitations and achieve robust, general-purpose language models.

Language Models are Unsupervised Multitask Learners

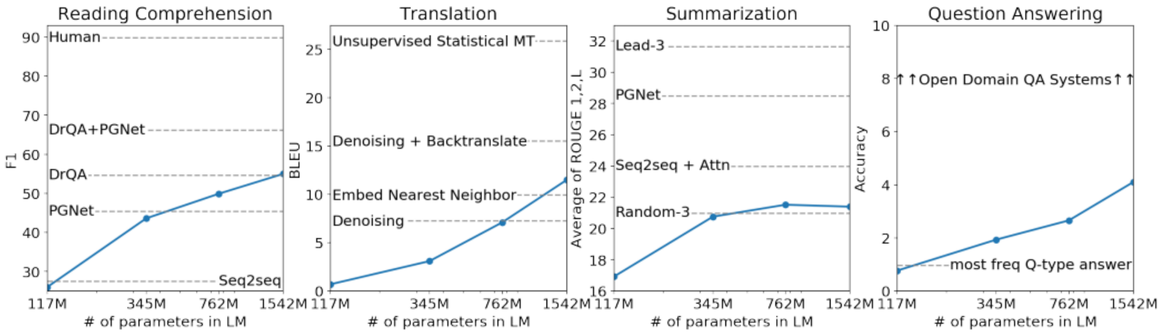


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural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

Evolution of Transfer Learning in NLP

State-of-the-art language processing systems leverage a combination of pre-training and supervised fine-tuning. This approach has undergone a continuous evolution towards more flexible transfer mechanisms. Early methods involved pre-training word vectors for input to task-specific architectures. This later progressed to transferring contextual representations learned by recurrent networks. Recent advancements suggest that dedicated task-specific architectures are no longer necessary. Transferring multiple self-attention appears sufficient.

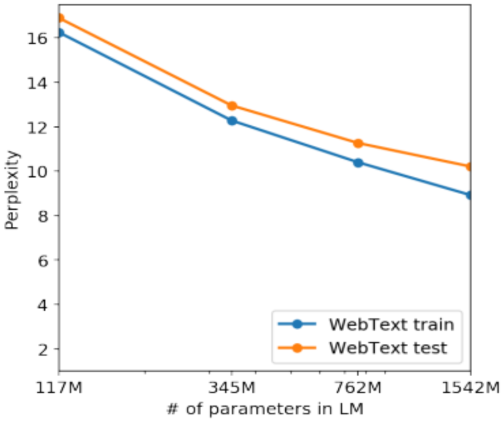


Figure 4. The performance of LMs trained on WebText as a function of model size.

While these methods require supervised training for specific tasks, alternative approaches utilizing language models have shown promise in scenarios with limited or no supervised data. Such models have demonstrated capabilities in tasks like commonsense reasoning. This work builds upon these advancements, pushing the boundaries of transfer learning towards more generalizable methods. They introduce a zero-shot approach, where language models perform downstream tasks without requiring any parameter or architecture adjustments. And showcase the potential of this approach by demonstrating the ability of these models to handle a wide range of tasks in this zero-shot setting. Results are promising, achieving competitive and even state-of-the-art performance depending on the specific task.

Model: Recent architectures like Transformers with self-attention have significantly improved the ability of models to capture complex relationships within text, making them better suited for tasks that require understanding these relationships.

Unsupervised Multitask Learning: Language models, by nature, learn to predict the next element in a sequence. This unsupervised objective aligns with many NLP tasks where the goal is also predicting specific outputs within a sequence. This overlap suggests that language models could potentially learn these tasks without explicit supervision. However, previous concerns about density estimation as a training objective remain. The key challenge lies in efficiently optimizing this unsupervised objective for real-world applications. While initial experiments show promise with large models, learning is demonstrably slower compared to supervised approaches.

Input Representation: Ideally, a language model should handle any possible string. Current large-scale models typically involve pre-processing steps like tokenization and handling unknown words, which limit the set of representable strings. Processing raw text as a sequence of bytes offers an elegant solution, but current byte-level models haven't achieved performance comparable to word-level models on large datasets. Their own experiments confirm this performance gap with byte-level models trained on WebText.

BPE for Efficient Byte-Level Language Modeling

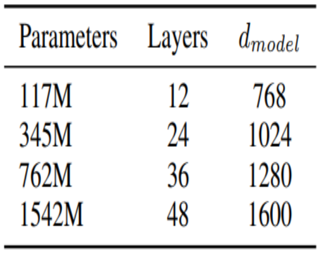
Byte Pair Encoding (BPE) offers a practical balance between character-level and word-level language modeling. It leverages frequent symbol sequences as words while handling infrequent ones as characters. While traditional BPE implementations operate on Unicode code points, this approach necessitates a very large base vocabulary (over 130,000 symbols) for comprehensive Unicode coverage.

This work proposes a byte-level variant of BPE that addresses this limitation. A byte-level BPE only requires a base vocabulary of 256, significantly reducing memory footprint. However, applying standard BPE directly to byte sequences leads to suboptimal merges due to its greedy frequency-based approach. We observed frequent merging of common words with variations (e.g., "dog", "dog!"), wasting valuable vocabulary slots and model capacity.

To address this, they introduce a modification that prevents BPE from merging across character categories within a byte sequence. This ensures efficient vocabulary usage while minimizing word fragmentation across multiple tokens. An exception is made for spaces, significantly improving compression without major word splitting.

Advantages of Byte-Level BPE Representation

This byte-level BPE representation offers a unique advantage by combining the established effectiveness of word-level language models with the adaptability of byte-level approaches. Crucially, it enables the model to assign a probability to any valid Unicode string. This characteristic grants us the flexibility to evaluate language models on any dataset without constraints imposed by pre-processing steps, tokenization schemes, or vocabulary size limitations.



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.

Transformer-Based Architecture with Refinements

Their language models leverage a Transformer architecture, similar to the OpenAI GPT model with several key modifications:

- **Layer Normalization Placement:** They adopt pre-activation residual network style by placing layer normalization before the input of each sub-block within the Transformer. An additional layer normalization is added after the final self-attention block.
- **Initialization Strategy:** A modified initialization approach is used to account for the accumulation of weights along the residual pathway as the model depth increases. This involves scaling the weights of residual layers by a factor of $1/\sqrt{N}$ during initialization, where N represents the total number of residual layers.
- **Vocabulary and Training Parameters:** The model vocabulary is expanded to 50,257 tokens compared to the original GPT model. Additionally, the context size is increased from 512 to 1024 tokens, and a larger batch size of 512 is employed during training.

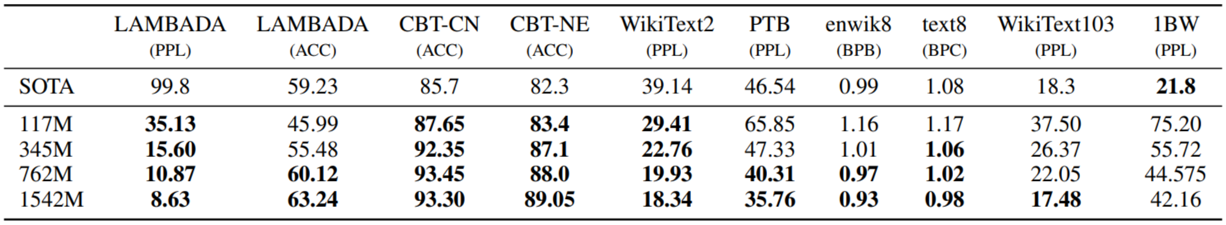
Model Evaluation

Our experiments involve several language modeling benchmarks to assess the capabilities of the proposed models:

- **Model Sizes:** We evaluate a range of model sizes, with the smallest being equivalent to the original GPT and the second-smallest matching the largest model from BERT (Devlin et al., 2018). Our largest model, GPT-2, boasts over ten times more parameters than the original GPT.
- **Training and Optimization:** The learning rate for each model was manually tuned to achieve the best perplexity on a held-out validation set (5%) of WebText data. It's important to note that all models still exhibit underfitting on WebText, with held-out perplexity potentially improving with further training.

Zero-Shot Language Modeling Transfer

As a first step towards zero-shot task transfer, they assess how WebText-trained language models perform on language modeling tasks in new domains without any fine-tuning. Since they used byte-level models eliminate the need for pre-processing or tokenization, they can be evaluated on various language modeling benchmarks:



	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).

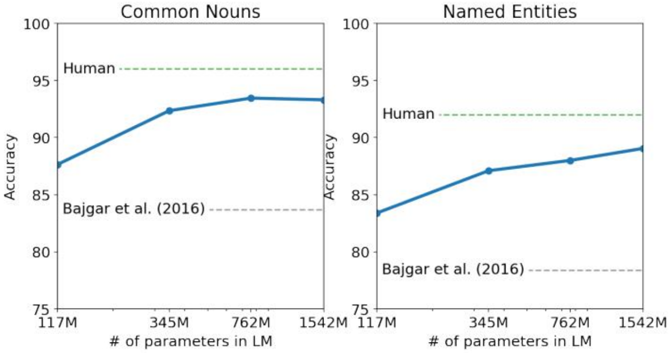


Figure 2. Performance on the Children's Book Test as a function of model capacity. Human performance are from Bajgar et al. (2016), instead of the much lower estimates from the original paper.

- **Children's Book Test (CBT):** This test evaluates a model's ability to handle different word categories (nouns, verbs, etc.) by presenting cloze tasks where the model predicts the missing word from ten possible choices. Here, they report accuracy instead of perplexity. Findings demonstrate a clear correlation between model size and performance on the CBT. Notably, GPT-2 achieves state-of-the-art results of 93.3% accuracy for common nouns and 89.1% for named entities.

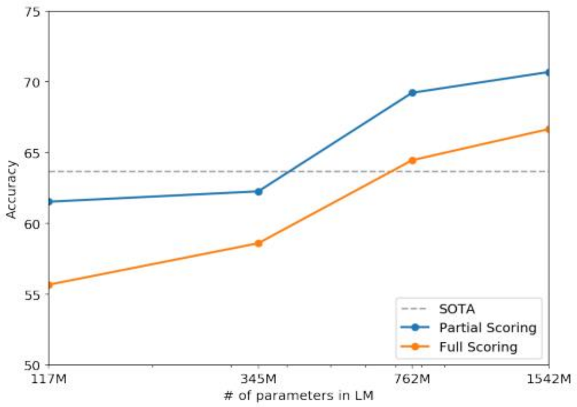
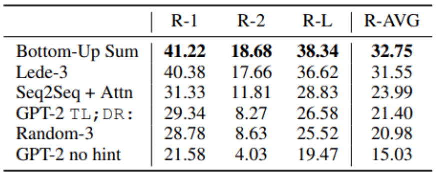


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

- **WikiText-103:** This benchmark measures perplexity on a dataset of English Wikipedia articles. GPT-2 achieves a state-of-the-art perplexity of 8.6, significantly improving on previous results.

Commonsense Reasoning Evaluation



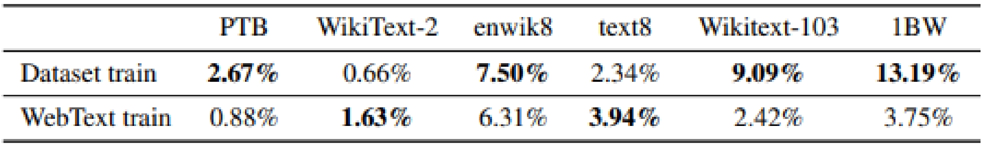
	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL;DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

- **Winograd Schema Challenge:** This benchmark assesses a model's ability to perform commonsense reasoning by resolving ambiguities within text. They follow the problem formulation by Trinh & Le (2018) and present their models' performance using both full and partial scoring techniques. GPT-2 surpasses the previous state-of-the-art by 7%, achieving an accuracy of 70.70%. However, it's crucial to consider the dataset size (273 examples) when interpreting this result.

Key Takeaways:

- Models exhibit a strong correlation between size and performance on various language modeling tasks.
- GPT-2 achieves state-of-the-art results on several benchmarks, demonstrating its effectiveness in zero-shot language modeling transfer.

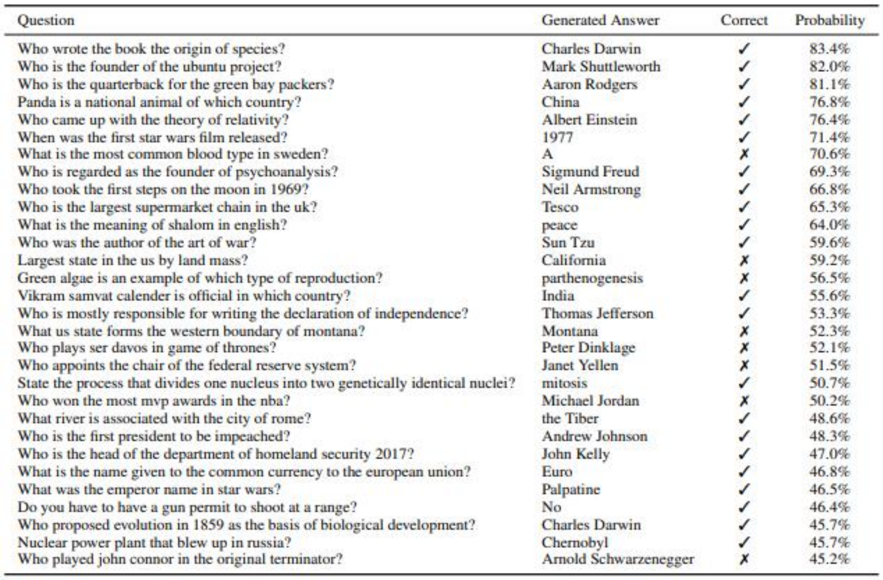


	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

- The models also showcase promising capabilities in commonsense reasoning tasks.

All in all, findings demonstrate that large language models, when trained on extensive and diverse datasets like WebText, exhibit the ability to perform well on tasks from various domains, even without explicit task-specific training (zero-shot setting). Notably, GPT-2 achieves state-of-the-art performance on 7 out of 8 language modeling benchmarks in a zero-shot evaluation.



Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

The versatility of GPT-2's performance across diverse tasks suggests a significant learning capacity within these models. Training on a broad and varied text corpus appears to enable them to acquire capabilities for a surprising range of tasks without requiring explicit supervision. This observation opens new avenues for exploring the potential of unsupervised learning in developing robust and generalizable language models.