

ECE 696B: Spring 2025
Trustworthy Machine Learning

GPT-2

Presented by: Ashley Tittelbaugh

Overview

Language Models are Unsupervised Multitask Learners

Alec Radford ^{* 1} Jeffrey Wu ^{* 1} Rewon Child ¹ David Luan ¹ Dario Amodei ^{** 1} Ilya Sutskever ^{** 1}

- Zero-shot performance
- Published by OpenAI
- GPT-2
- Cited **15,480 times**
(as of Jan 27, 2025)

Previous Work and Shortcomings

- GPT-1/BERT: Generalized Pre-training and Fine-tuning
 - **Generalized Pre-training:**
 - Large Data-set
 - No given task
 - **Fine-tuning:**
 - Modifies all parameters for a specific task
 - Structure inputs
- What are some issues/problems with this model ?
 - Architectures, Algorithms and/or Training is Task Specific
 - Issue 1: Requires pre-known knowledge and sufficient data
 - Issue 2: Diverse inputs -> low performance

New Idea

- Main Idea: Use a **large dataset** to perform **unsupervised learning** to result in competitive **zero-shot results**

$$p(x) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{i-1}) \quad (1)$$

p(output, task)

- Task interpretation from language

"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 côté? -Quel autre côté?**", 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.

Relevant Changes and Updates

Previous

Dataset

News Articles, Wikipedia
Common Crawl

Small, Task Specific

Large but not quality

Input Encodings

Byte-Pair Unicode
encodings

Large vocab
Duplicate information

Dog. Dog? Dog!

GPT-2

Dataset

WebText – Large and
quality

Input Encodings

Byte-Pair encodings at
byte level
Restricted combinations

WebText

- Reddit outgoing links with 3 or more karma
 - “Quality Control”
- 45 million outgoing links
- De-duplication
- 40GB of text
- NO Wikipedia documents
 - Worries about recursive sourcing

Transformer Architecture

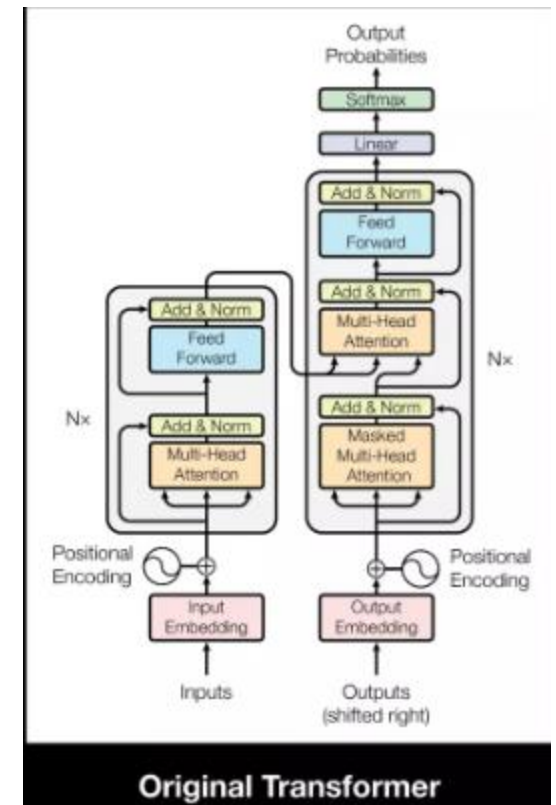
- Based on the GPT-1 transformer architecture
- Layer normalization moved to the input of each block
- Additional LayerNorm added after the self-attention block
- Vocabularly 50,257

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

GPT-1

BERT-large

GPT-2



Aside: Perplexity Scores

“The best model is one that predicts an unseen test set”

- **Lower score** => less “perplexed” the model is with unseen inputs
- Inverse probability of the test set normalized by the number of words

$$\begin{aligned} PP(W) &= P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}} \end{aligned}$$

- Comes from information theory $PP(W) = 2^{\text{cross entropy}}$

$$\text{Perplexity} = e^{H(P,Q)} = e^{-\frac{1}{N} \sum_{i=1}^N \log q(w_i)}$$

Results: Language Modeling


Language Models are Unsupervised Multitask Learners

	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

- **Language Modeling:**
 - **Primary task** that they are trained for
 - Modify Preprocessing from normal testing parameter
 - De-tokenization
 - Remove unfamiliar characters
 - De-standardize text
 - **“De preprocessing” Introduce extra error**
 - Significant improvement in small datasets
 - **Penn Tree Bank** and Wiki Text each have 1 to 2 million tokens
 - Underperforms the 1 Billion Word benchmark
 - Largest data set
 - Could indicate underfitting
 - Most aggressive preprocessing

Aside: Cloze Tasks

most	up	grow	necks	Africa
feet	leaves	bus	long	tallest



Giraffes are the _____ living animal in the world. They can _____ up to about 5 metres tall. That is about as tall as a doubledecker _____!

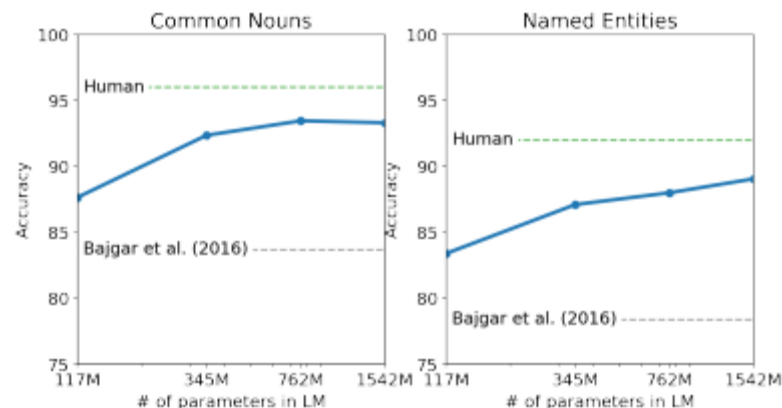
Giraffes live in _____. Their long _____ help them to eat the _____ in the tallest part of the trees. They like the leaves on the acacia trees _____ of all.

- Random parentage of words are masked
- LLM are asked to “Fill in the blanks” from their vocabulary

Results: Children's Book Test

Language Models are Unsupervised Multitask Learners										
	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

- **Children's Book test:**
- **Cloze task** – random words are masked from various children's books
- Performance of LLM on different categories of words
- Performance increases with model size
- Closes the majority of the gap between previous state of the art and human
- One book – The Jungle Book – included in WebText
 - removed for testing

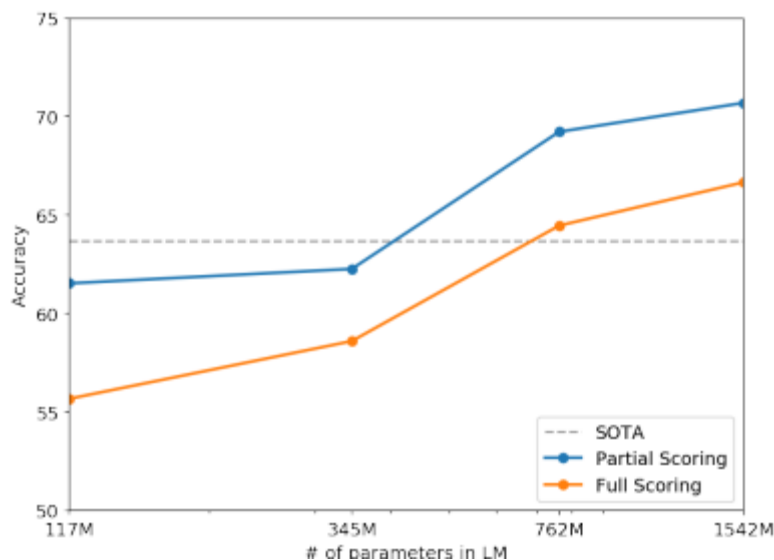


Results: Long Term Dependencies

Language Models are Unsupervised Multitask Learners										
	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

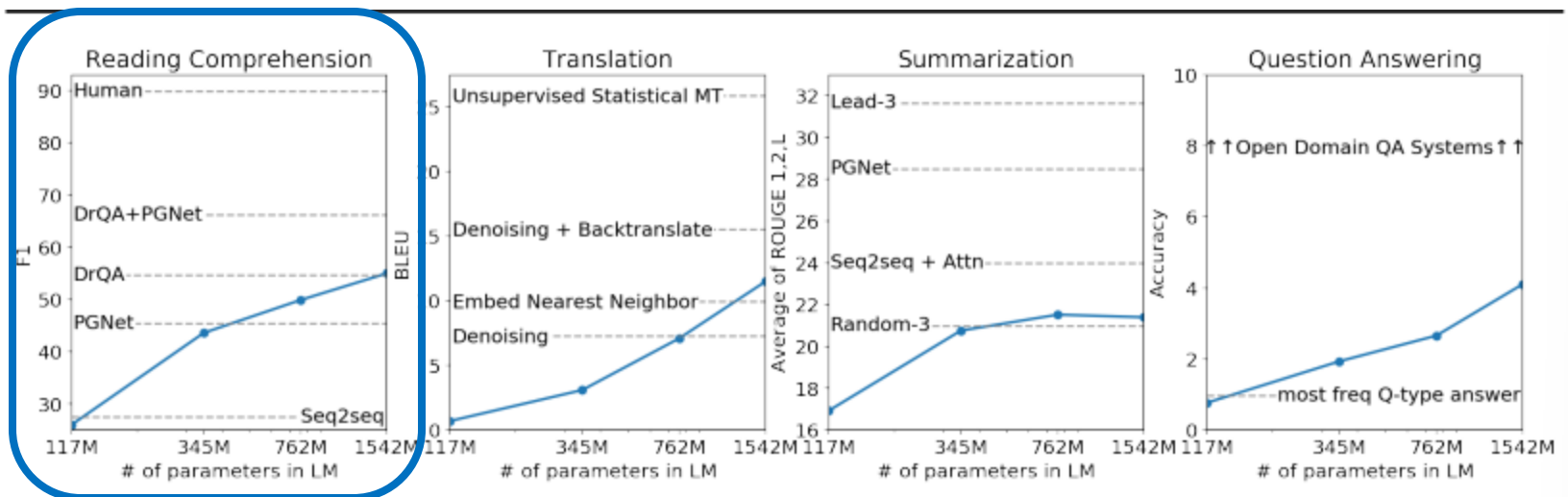
- **Language Modeling Broadened to Account for Discourse Aspects**
 - Task is to predict final word of a sentence that requires at least 50 tokens for humans to predict correctly
 - GPT-2 improved on state of the art
 - Mistakes were often a valid continuation of the sentence but not the end of the sentence

Results: Common Sense reasoning



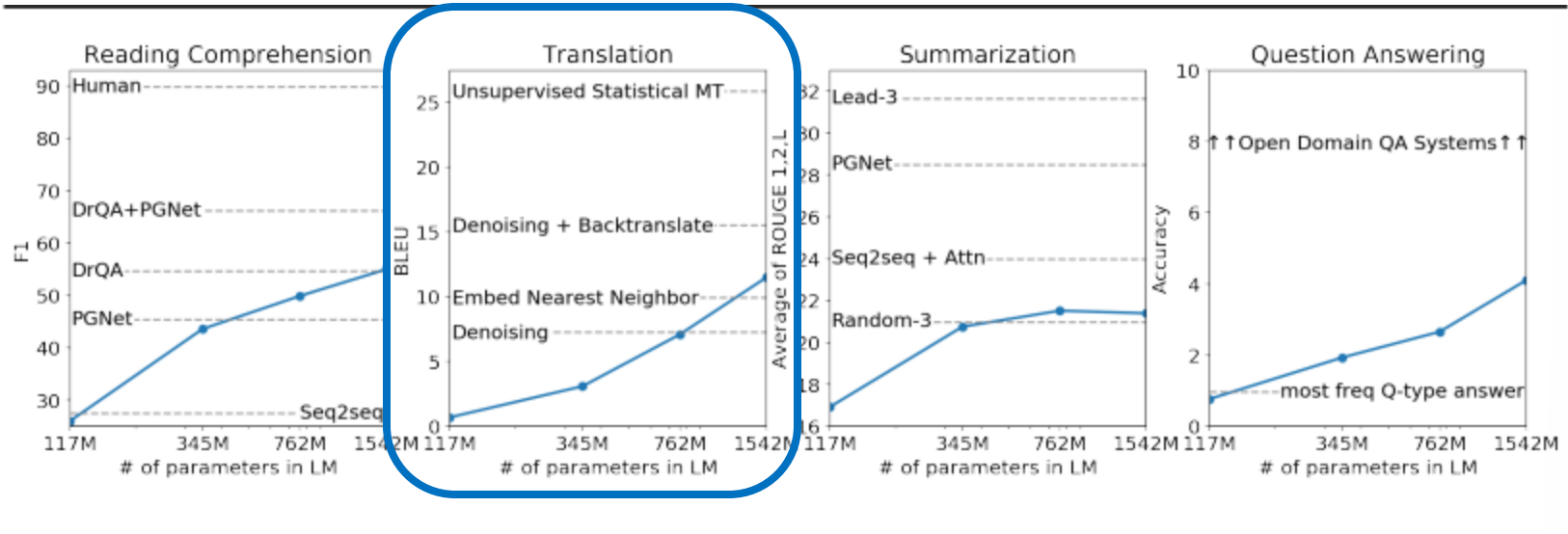
- **Winograd Schema Challenges**
 - Measures common sense reasoning by the ability to resolve ambiguities in text
 - Very small dataset

Results: Reading Comprehension



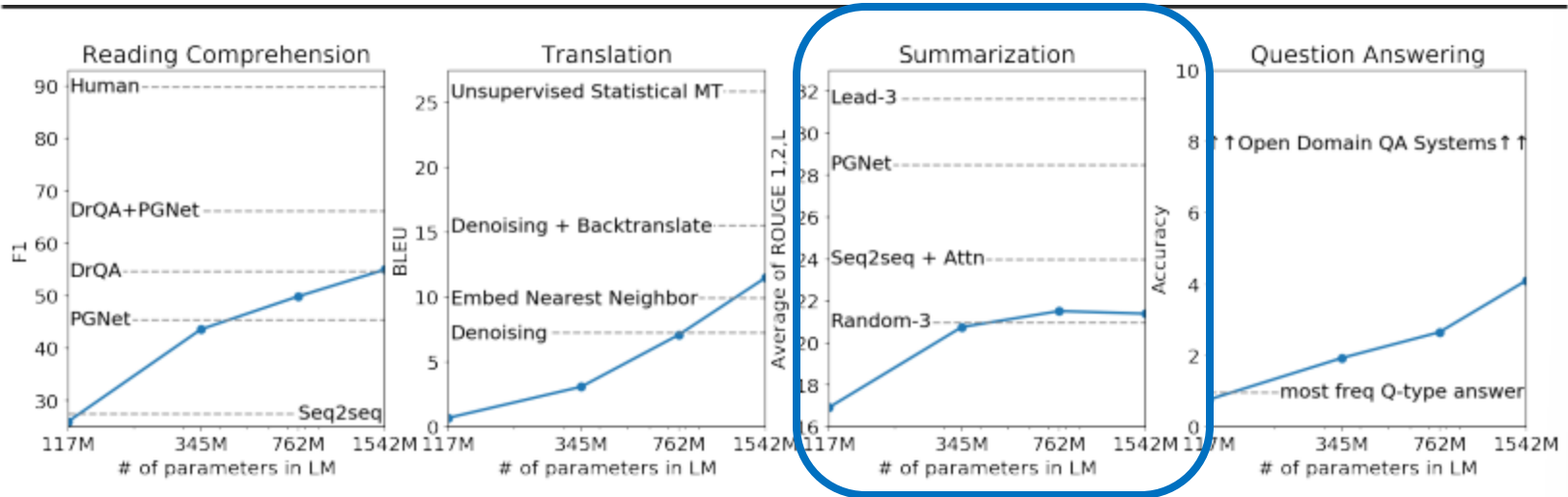
- **Conversation Question Answering dataset (CoQA)**
 - 7 distinct domains that contain dialogue between a question asker and a question answerer
 - Measures both reading comprehension ability and ability to retain context from previous conversation
- GPT-2 seemed to use simple retrieval based characteristics rather than a more complex understanding

Results: Translation



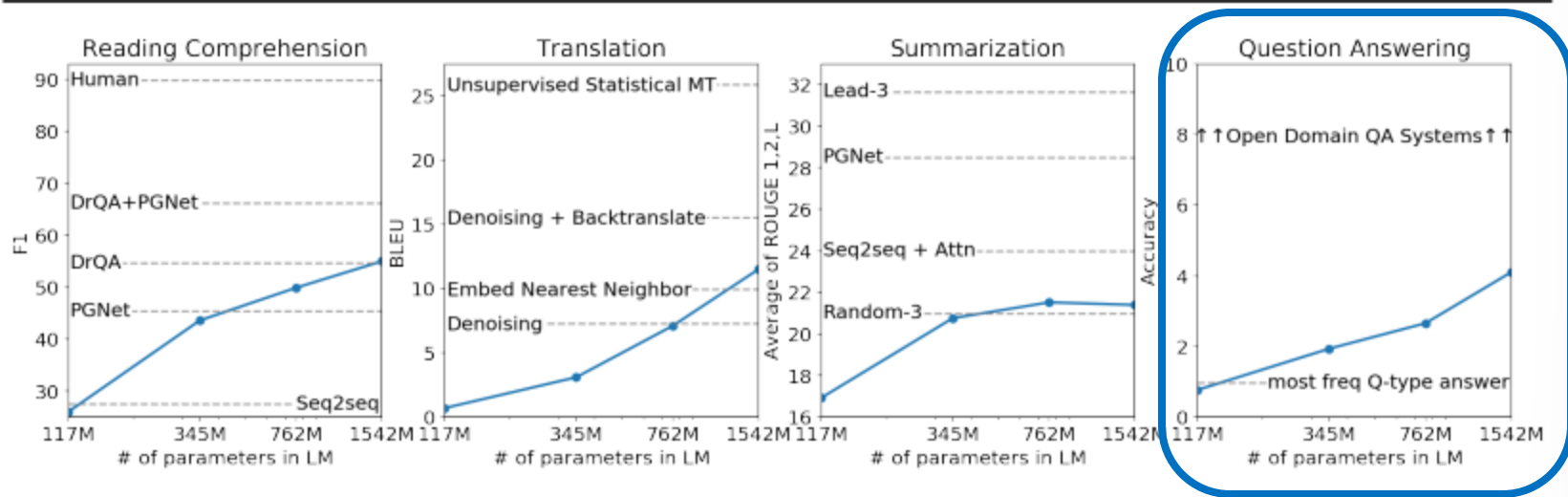
- **Previously seen WMT**
- GPT-2 only slightly better than word by word replacement from English to French **5 BLEU**
- GPT-2 does much better French to English **11.5 BLEU**
 - Due to large English database
- Surprising result!
 - Had removed all non-English documents from WebText

Results: Summarization



- **CNN and daily mail dataset**
- Added TL;DR: token for task inference
- Outputs resembled length and verbiage of summeries but often focused on specific details
- Only slightly better than picking 3 sentences at random

Results: Question answering



- **Natural Questions dataset**
 - Usually used with neural systems
- Test what information is contained in a language model
- GPT-2 is 5.3 times more accurate than smallest model
- Greatly underperforms open domain question answering systems
- Hybridize information retrieval with extractive document question answering

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%

Memorization vs Generalization

Does such a large dataset just contain most of the answers?

	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.

- GPT-2 contains equivalent or less overlapping data than standard training data
- Can answer questions without having the answer in the dataset

Key Takeaways

- With a large and quality dataset and model, GPT-2 Zero-shot results often preform at or above baseline
- Language models can learn tasks from language
- There are still some tasks where GPT-2 doesn't preform very well

What's Next?

- Does not argue for complete switch to unsupervised models
- Rather promotes further research into these multitask models with large unsupervised training as a first step.
- Results show GPT-2 **underfits** WebText
 - There are benefits to be gained with even larger models