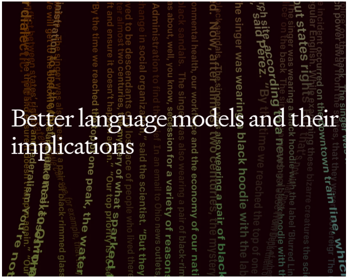
Generative Pre-Trained Transformers

GPT-2 (2019)

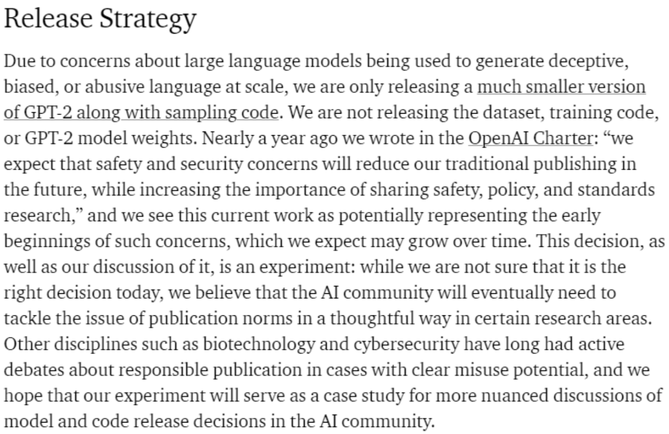


Better language models and their implications

Learning goals

- Get a first idea about prompting
- Understand the implications of such models

STARTING WITH A CONTROVERSY

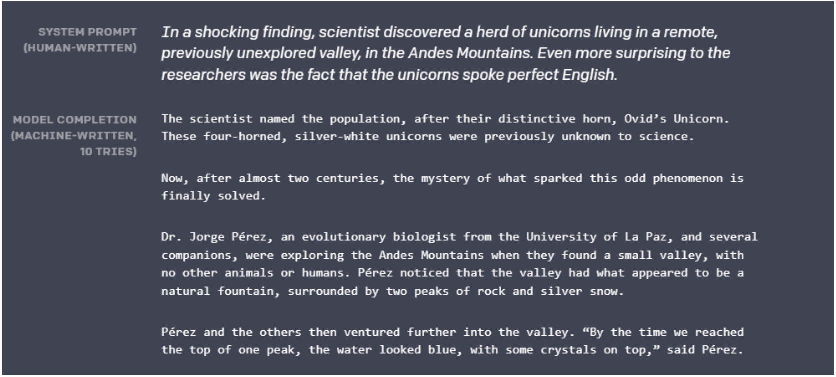


Release Strategy

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: “we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research,” and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time. This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in certain research areas. Other disciplines such as biotechnology and cybersecurity have long had active debates about responsible publication in cases with clear misuse potential, and we hope that our experiment will serve as a case study for more nuanced discussions of model and code release decisions in the AI community.

▸ Source: OpenAI Blog

CAPABILITIES – STORYTELLING



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

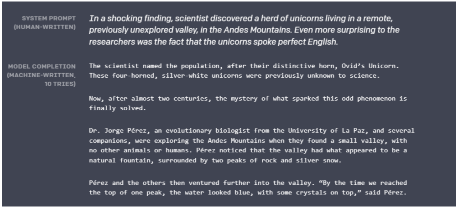
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.

▸ Source: OpenAI Blog

CAPABILITIES – STORYTELLING



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

▸ Source: OpenAI Blog

- **In 2019:** Great achievement as the model is able to continue a made up story *in a coherent way* by making up its own facts

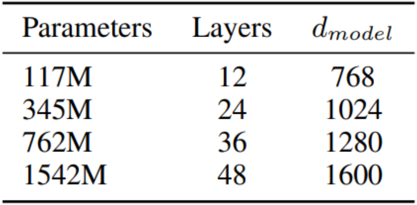
  Sill: Not very consistent (“`four-horned [...] unicorns`“)
- **In 2023:** Nowadays this phenomenon is known as **hallucination** and a lot of research effort is put into mitigating this behaviour

THE ARCHITECTURE

- Transformer decoder pre-trained on AR language modeling
- Custom web scrape (not publicly available) of all outbound links from Reddit

→ 8M documents / 40GB of text

- Byte-level BPE for tokenization
- Increased context size: 512 → 1024



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

Source: Radford et al., 2019

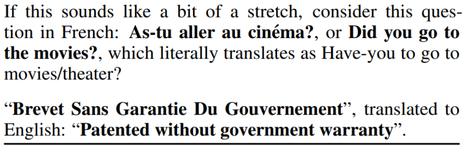
CAPABILITIES – LANGUAGE MODELING

- Authors state that “*language provides a flexible way to specify tasks, inputs, and outputs all as a sequence of symbols*“:

  `(translate to french, english text, french text)`

- Model benefits from seeing “natural“ occurrences of task demonstrations during pre-training on large-scale web corpora

  → Related rationale to T5 (*literal task descriptions*), but slightly different (*acquired on the fly already during pre-training*).



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?

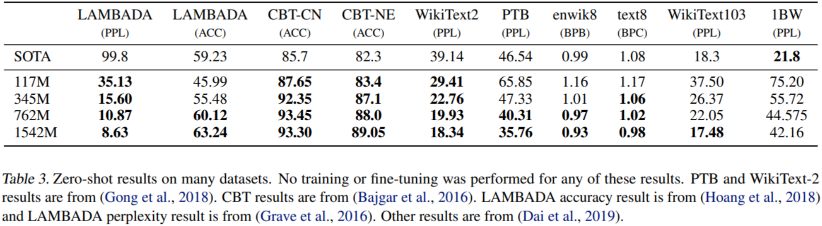
“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

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

▸ Source: Radford et al., 2019

CAPABILITIES – LANGUAGE MODELING

- Language modeling performance measured via perplexity / bits-per-character (lower is better) on held-out corpora
- Other data sets (accuracy):
  - LAMBADA ▸ Paperno et al., 2016: Sentence completion
  - Children's Book Test (CBT) ▸ Hill et al., 2015: Cloze-style task; predict correct one (among 10 alternatives for omitted word)



	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

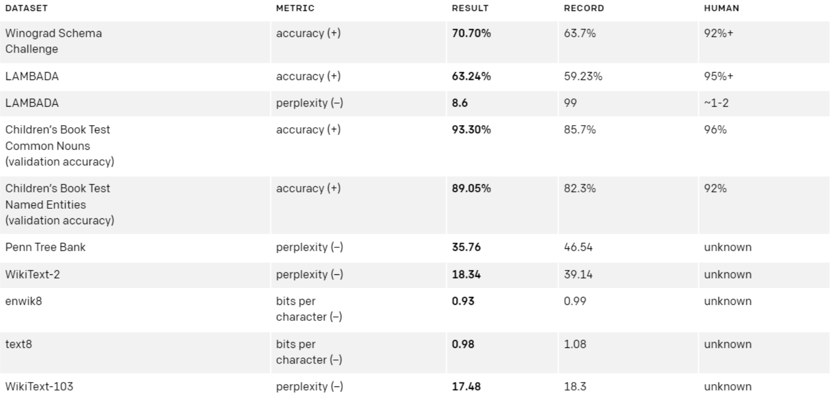
Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from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

▸ Source: Radford et al., 2019

CAPABILITIES – ZERO-SHOT

- *Zero-shot* refers to solving a task which the model was not previously explicitly trained on, just providing it with a description of the task but no demonstrations, i.e. input/output pairs.
- *Few-shot* would be a relaxation of this setting, allowing for also providing the models with demonstrations (but still no training/gradient updates).
- Radford et al. show GPT-2's zero-shot capabilities on a range of different tasks, paving the way for the developments that came with GPT-3.

CAPABILITIES – ZERO-SHOT



DATASET	METRIC	RESULT	RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+
LAMBADA	accuracy (+)	**63.24%**	59.23%	95%+
LAMBADA	perplexity (–)	**8.6**	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	**93.30%**	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	**89.05%**	82.3%	92%
Penn Tree Bank	perplexity (–)	**35.76**	46.54	unknown
WikiText-2	perplexity (–)	**18.34**	39.14	unknown
enwik8	bits per character (–)	**0.93**	0.99	unknown
text8	bits per character (–)	**0.98**	1.08	unknown
WikiText-103	perplexity (–)	**17.48**	18.3	unknown

▸ Source: Radford et al., 2019

(Note: The numbers correspond to the last row of the Table on slide 6)

CAPABILITIES – FACTUAL KNOWLEDGE

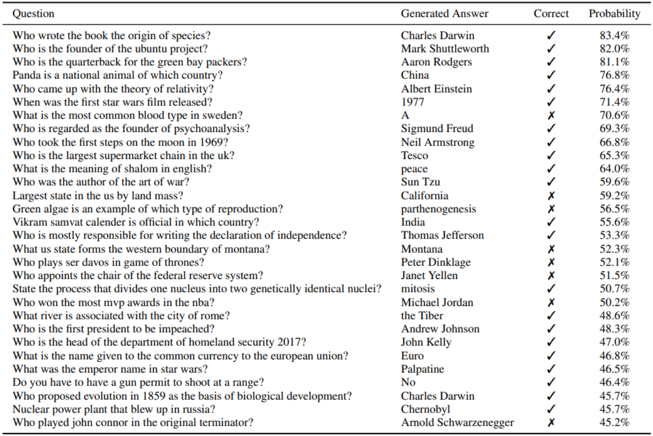
Natural Questions ▸ Kwiatkowski et al., 2019

- Testing the *factual knowledge* that is present in the pre-trained model
- Smallest model (117M): $< 1\%$
- GPT-2 (1542M): 4.1$\%$

  $\rightarrow$ Model capacity as a major factor
- Calibration: High accuracy (63.1%) on the 1% most confident answers

CAPABILITIES – FACTUAL KNOWLEDGE

The 30 most confident answers:

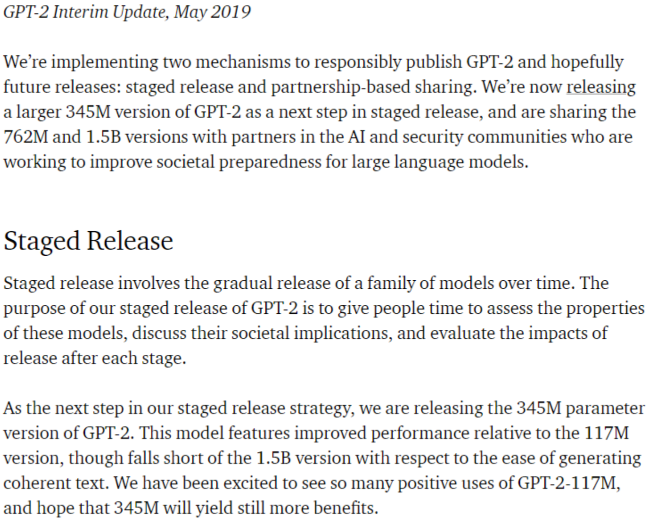


Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	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.

▸ Source: Radford et al., 2019

IN THE MEANTIME



GPT-2 Interim Update, May 2019

We're implementing two mechanisms to responsibly publish GPT-2 and hopefully future releases: staged release and partnership-based sharing. We're now releasing a larger 345M version of GPT-2 as a next step in staged release, and are sharing the 762M and 1.5B versions with partners in the AI and security communities who are working to improve societal preparedness for large language models.

Staged Release

Staged release involves the gradual release of a family of models over time. The purpose of our staged release of GPT-2 is to give people time to assess the properties of these models, discuss their societal implications, and evaluate the impacts of release after each stage.

As the next step in our staged release strategy, we are releasing the 345M parameter version of GPT-2. This model features improved performance relative to the 117M version, though falls short of the 1.5B version with respect to the ease of generating coherent text. We have been excited to see so many positive uses of GPT-2-117M, and hope that 345M will yield still more benefits.

▸ Source: OpenAI Blog

IN THE MEANTIME

- Available on huggingface: `https://huggingface.co/gpt2`
- GPT-3 built on the foundation laid by GPT-2
- (also ChatGPT happened)
- Prompting models has become more and more common (cf. next chapter)
- Few-/Zero-Shot capabilites of models have become more important (cf. next chapter)
- Models of over 200× the size of GPT-2 have been trained
- Transformer still the backbone of (nearly) all of them