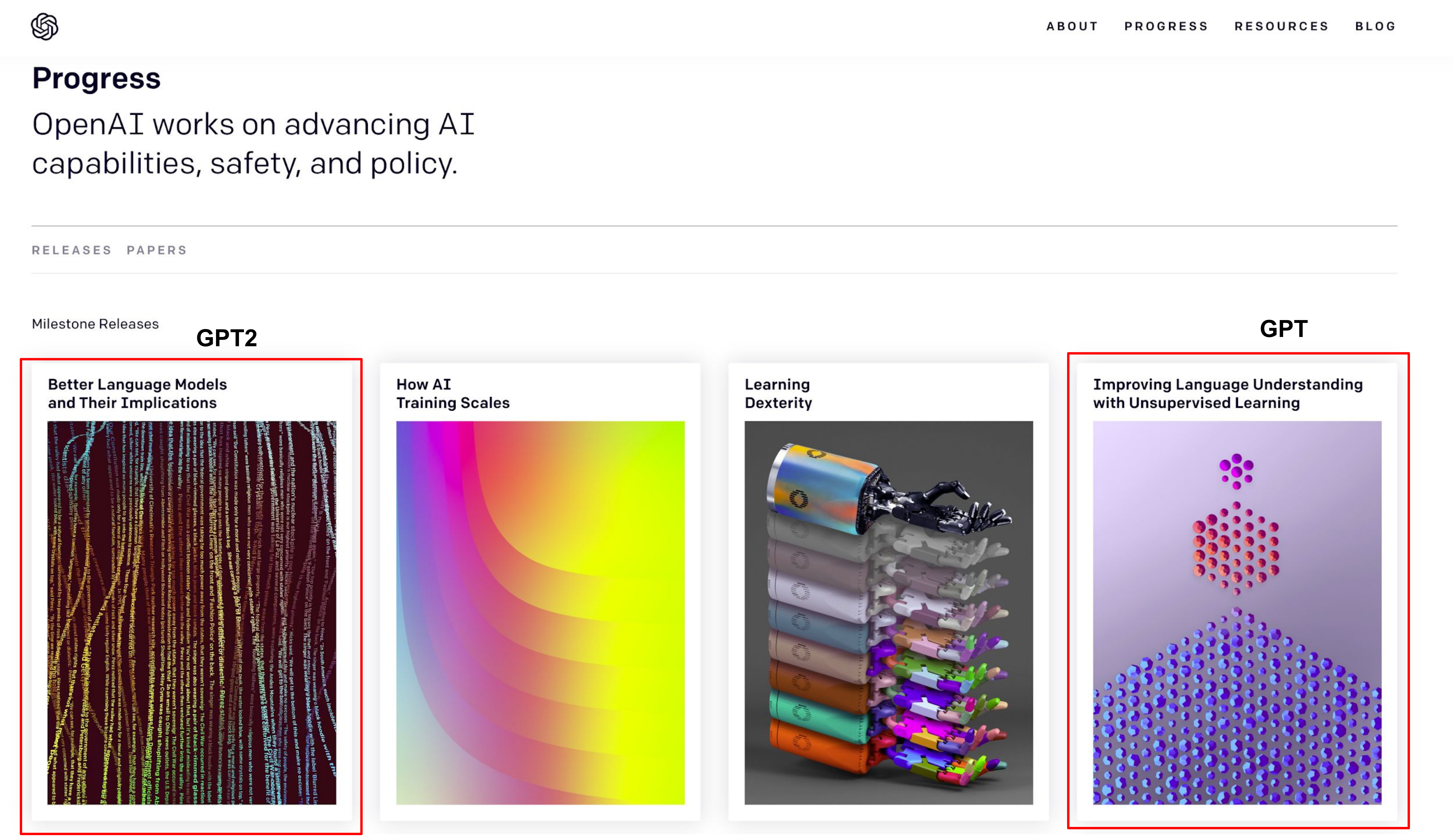
Language Models are Unsupervised Multitask Learners

Alec Radford, Jeffrey Wu et al.
2019 OpenAI

Kyung-Jae Cho
Research Scientist, VUNO Inc.



Progress

OpenAI works on advancing AI capabilities, safety, and policy.

Milestone Releases

GPT2

Better Language Models and Their Implications

How AI Training Scales

Learning Dexterity

GPT

Improving Language Understanding with Unsupervised Learning

Motivation

Current deep learning system

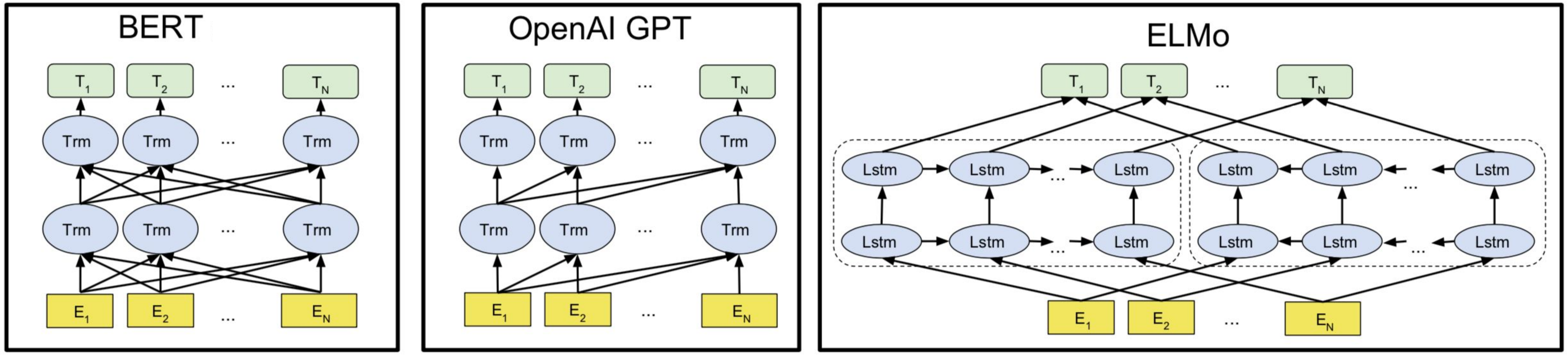
Large datasets ✚ high-capacity models ✚ supervised learning

- These systems are **sensitive to** slight changes in **data distribution and task specification**
- We would like to **move towards more general systems** which can perform many tasks
- We demonstrate language models can perform **down-stream tasks in zero-shot setting**

Related Works

- Multitask learning
  - still nascent in NLP tasks, need large datasets from various tasks
- Pretraining + Supervised Fine-tuning (e.g., GPT, BERT, ELMo)
  - Still require supervised training

Task-specific architectures are no longer necessary



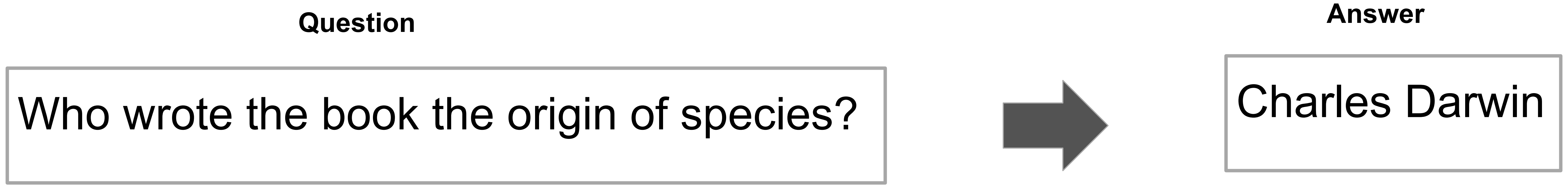
GPT2 => More general language model able to perform with only minimal or no supervised data available

Approach

- Core: Language modeling (unsupervised learning)
  - Predict the next symbol given previous symbols

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

- NLP tasks such as QA, Summarization, Reading comprehension outputs sentences
  - **Supervised objective == Unsupervised Objective**



Question

Who wrote the book the origin of species?

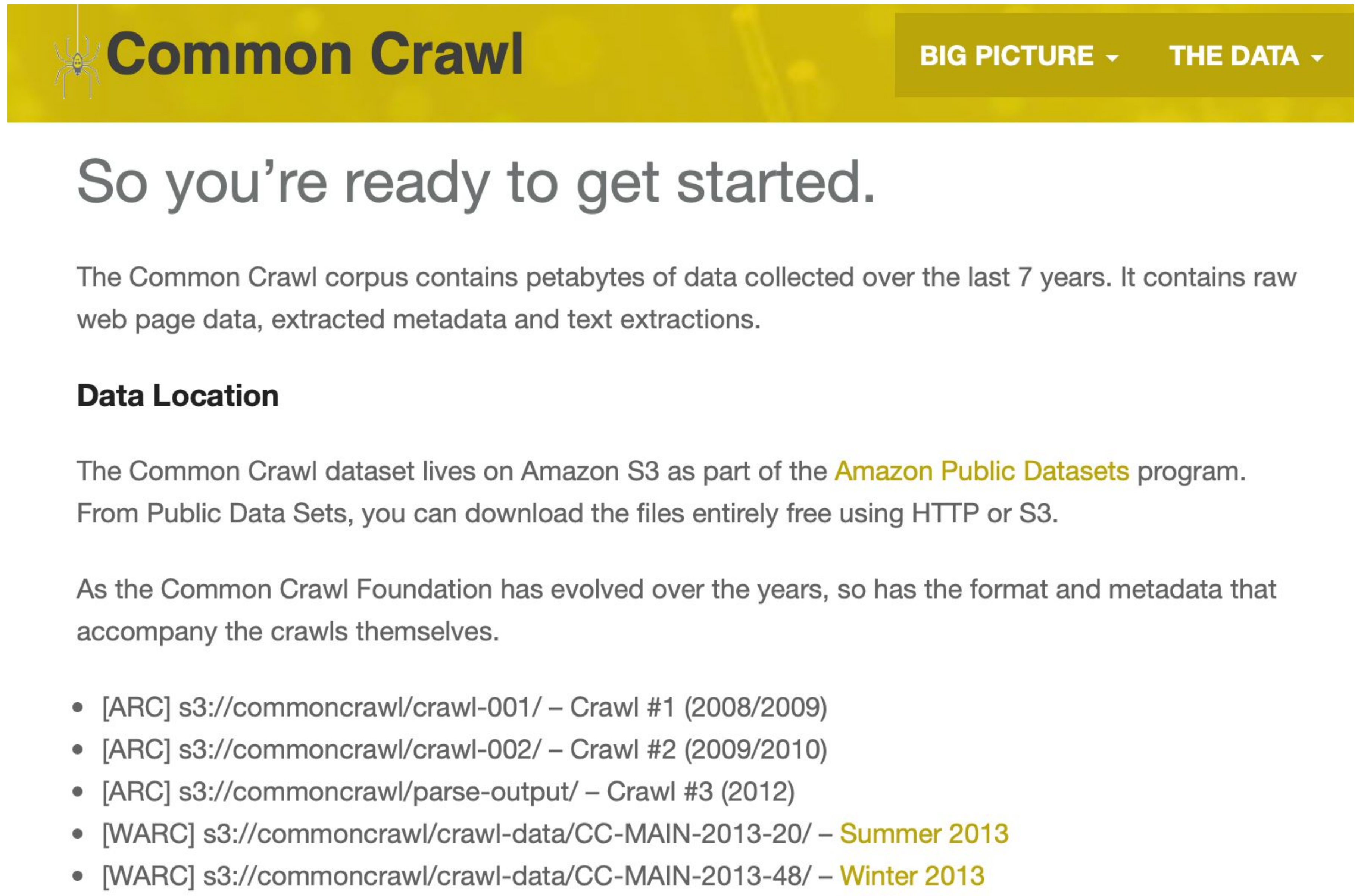
Answer

Charles Darwin

- **LM with sufficient capacity + diversity of dataset**
  - It will begin to **naturally learn to infer** and perform NLP tasks

Training Dataset

- Our approach motivates **building as large and diverse dataset possible**
- **Scraping content from the Internet**
  – Used only pages which have been curated/filtered by humans
  – Outbound links from Reddit which received at least 3 karma
- After de-duplication and some heuristic based cleaning contains slightly over **8 million documents for a total of 40GB of text**



So you're ready to get started.

The Common Crawl corpus contains petabytes of data collected over the last 7 years. It contains raw web page data, extracted metadata and text extractions.

Data Location

The Common Crawl dataset lives on Amazon S3 as part of the Amazon Public Datasets program. From Public Data Sets, you can download the files entirely free using HTTP or S3.

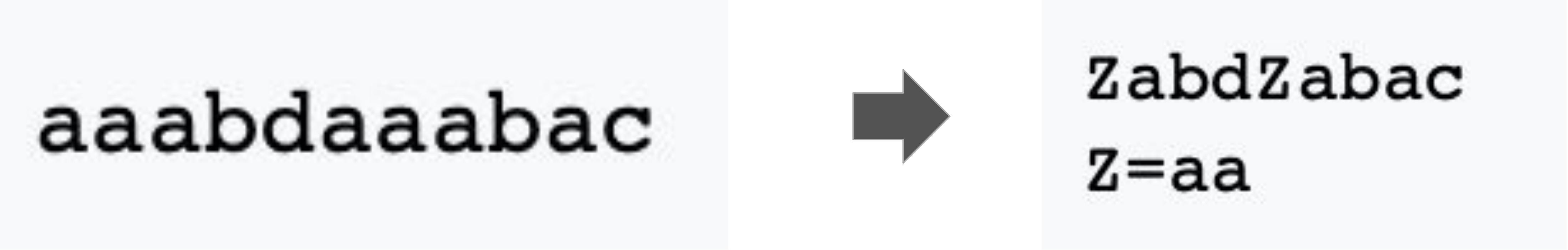
As the Common Crawl Foundation has evolved over the years, so has the format and metadata that accompany the crawls themselves.

- [ARC] s3://commoncrawl/crawl-001/ – Crawl #1 (2008/2009)
- [ARC] s3://commoncrawl/crawl-002/ – Crawl #2 (2009/2010)
- [ARC] s3://commoncrawl/parse-output/ – Crawl #3 (2012)
- [WARC] s3://commoncrawl/crawl-data/CC-MAIN-2013-20/ – Summer 2013
- [WARC] s3://commoncrawl/crawl-data/CC-MAIN-2013-48/ – Winter 2013

Input Representation

- Current large scale LMs include pre-processing steps (e.g., lower-casing, tokenization, out-of-vocabulary tokens)
  - **Restrict the space of model-able strings**, (low generalization and performable tasks)

- Byte-level LM
  - **Unicode strings** less restrict the space
  - Byte Pair Encoding (account for large Unicode code points)
    - Use frequent symbol as unit



aaabdaaabac

ZabdZabac
Z=aa

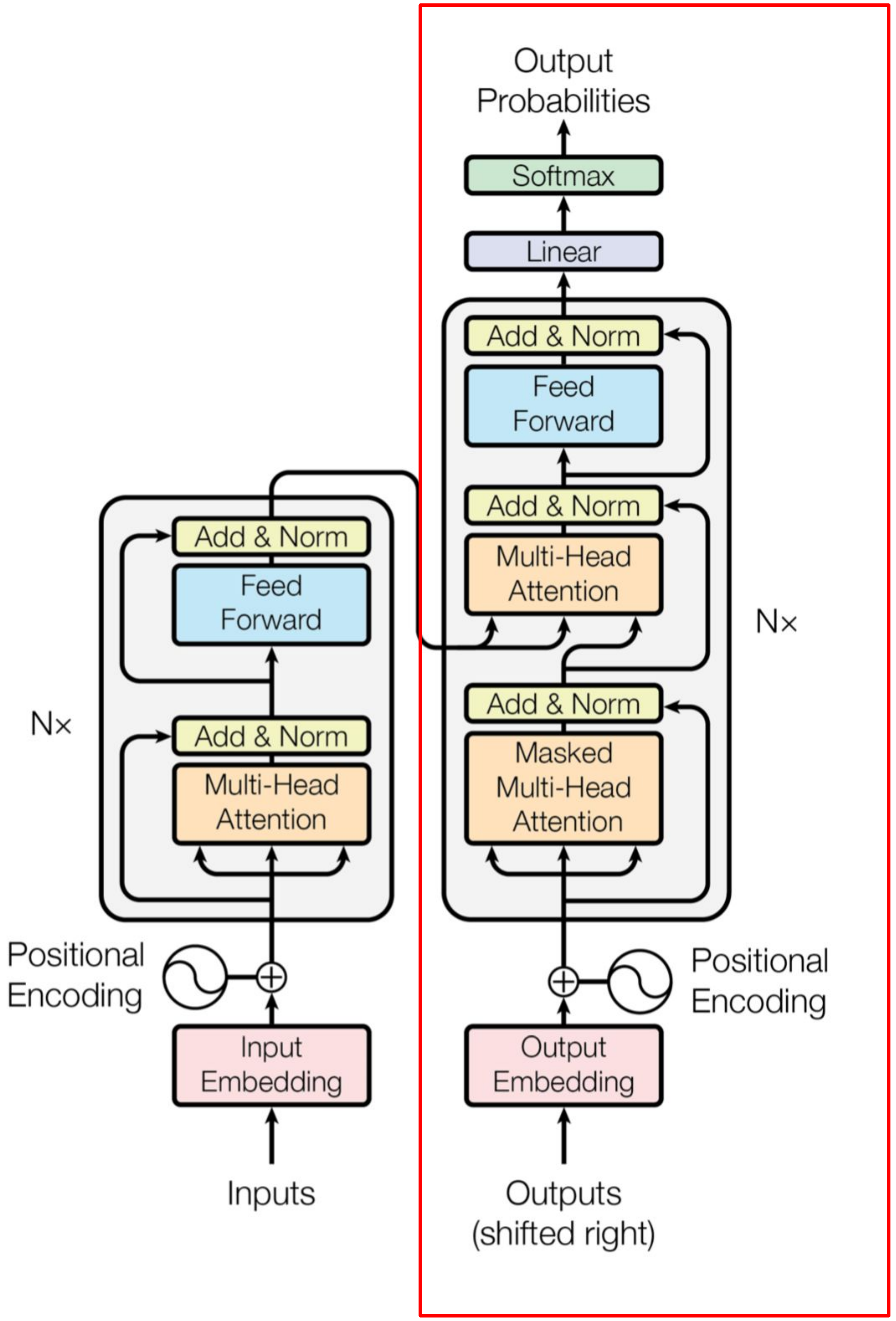
- **Can assign a probability to any Unicode string**
- **This allows us to evaluate our LMs on any dataset**

Model

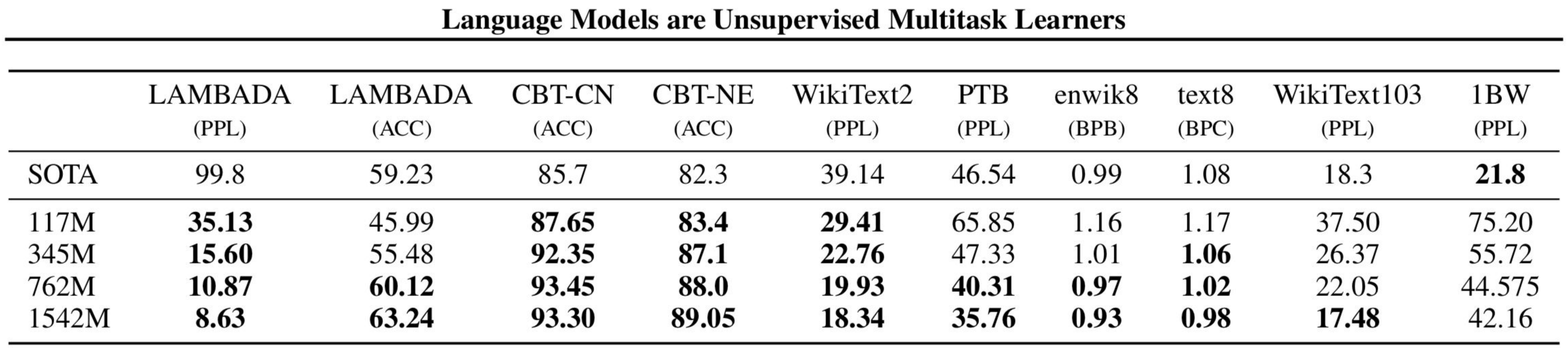
- Large transformer-based LM (1.5 billion parameters)
  - Transformer decoder (BERT use Transformer Encoder)

- GPT vs. GPT2
  - Direct scale-up of GPT
  - more than 10x the parameters
  - trained on more than 10X amount of data
  - No fine tuning layer

Transformer Decoder



Experiments



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

- **Achieves SOTA scores** on variety of domain-specific LM tasks in zero-shot setting
- Outperforms models trained on domain-specific datasets (e.g., Wikipedia, news, books)

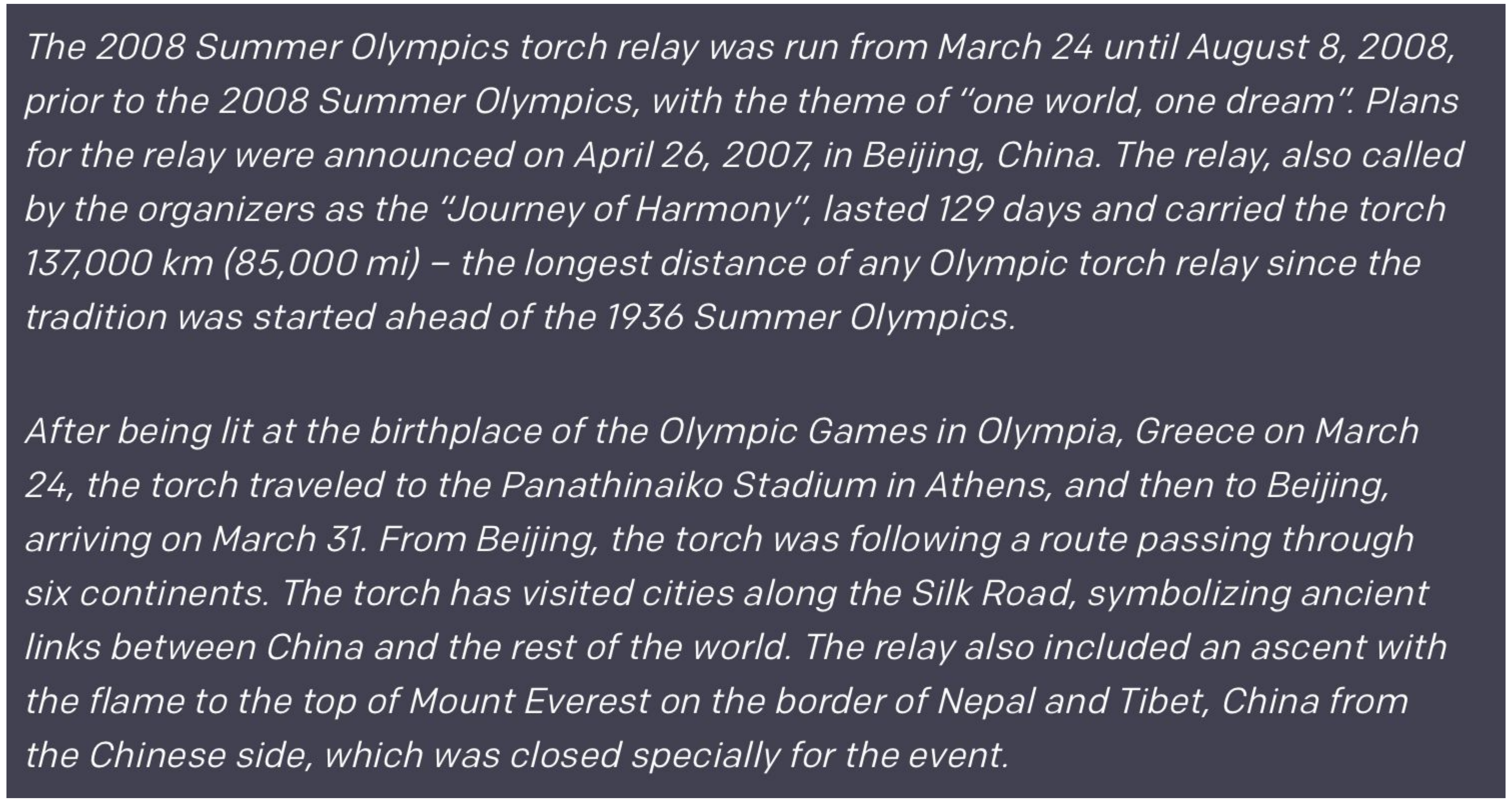
- CBT: examine whether the model correctly predicts 10 possible choices for an omitted word
- LAMBADA: Predict the final word of sentences

Experiments

- On other language tasks like **QA, reading comprehension, summarization, and translation**
  - We are able to get **surprising results without any fine-tuning** of our models, simply by prompting the trained model in the right way
- Though we do still **fall short of SOTA** for specialized systems

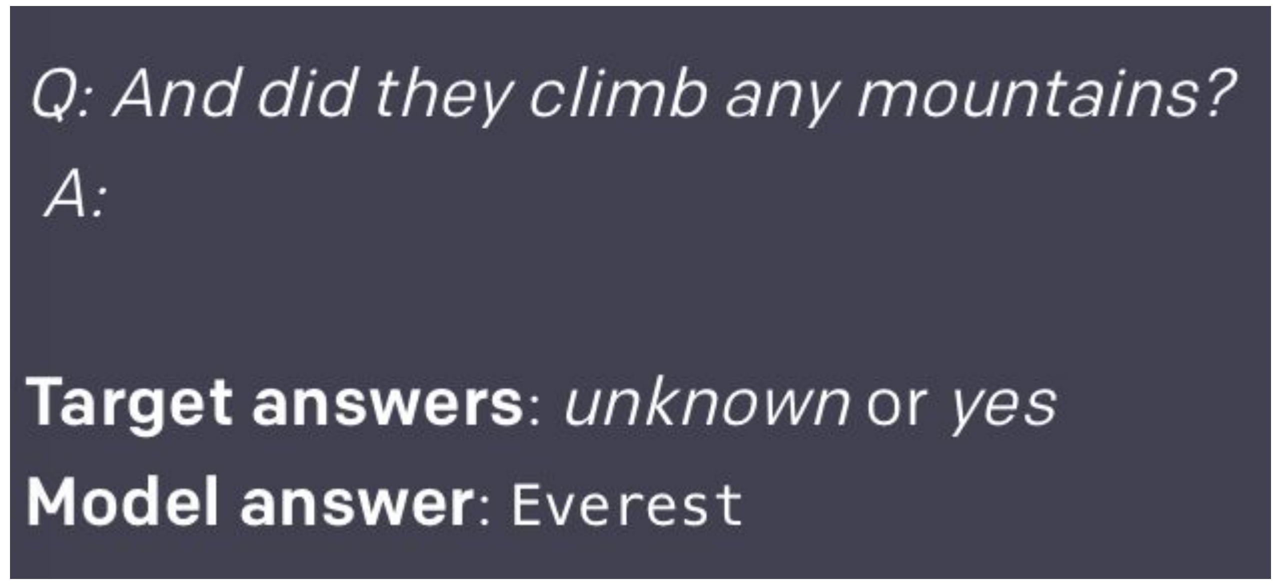
Reading comprehension

- Answer questions about given passages



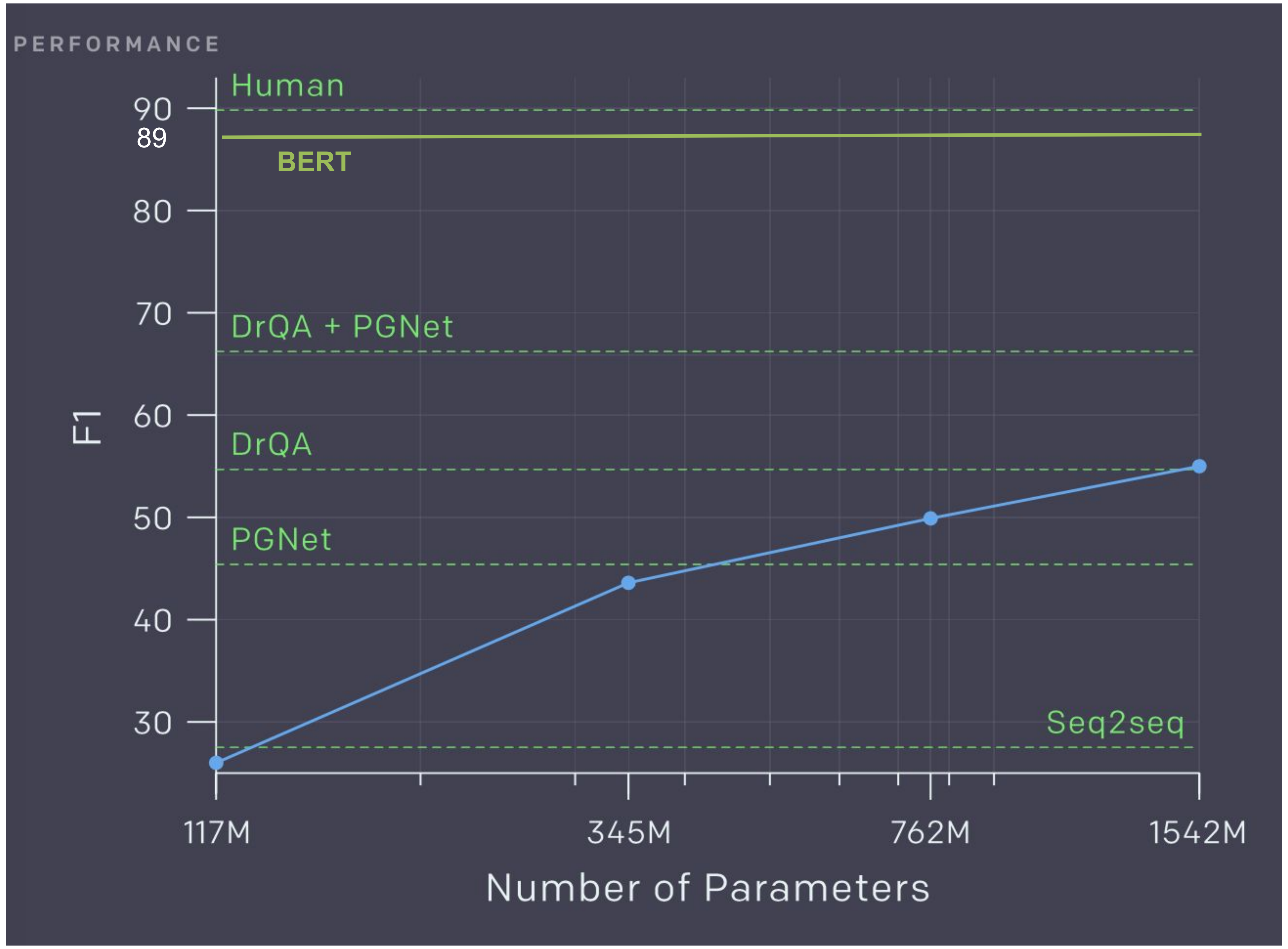
The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of "one world, one dream". Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the "Journey of Harmony", lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

After being lit at the birthplace of the Olympic Games in Olympia, Greece on March 24, the torch traveled to the Panathinaiko Stadium in Athens, and then to Beijing, arriving on March 31. From Beijing, the torch was following a route passing through six continents. The torch has visited cities along the Silk Road, symbolizing ancient links between China and the rest of the world. The relay also included an ascent with the flame to the top of Mount Everest on the border of Nepal and Tibet, China from the Chinese side, which was closed specially for the event.



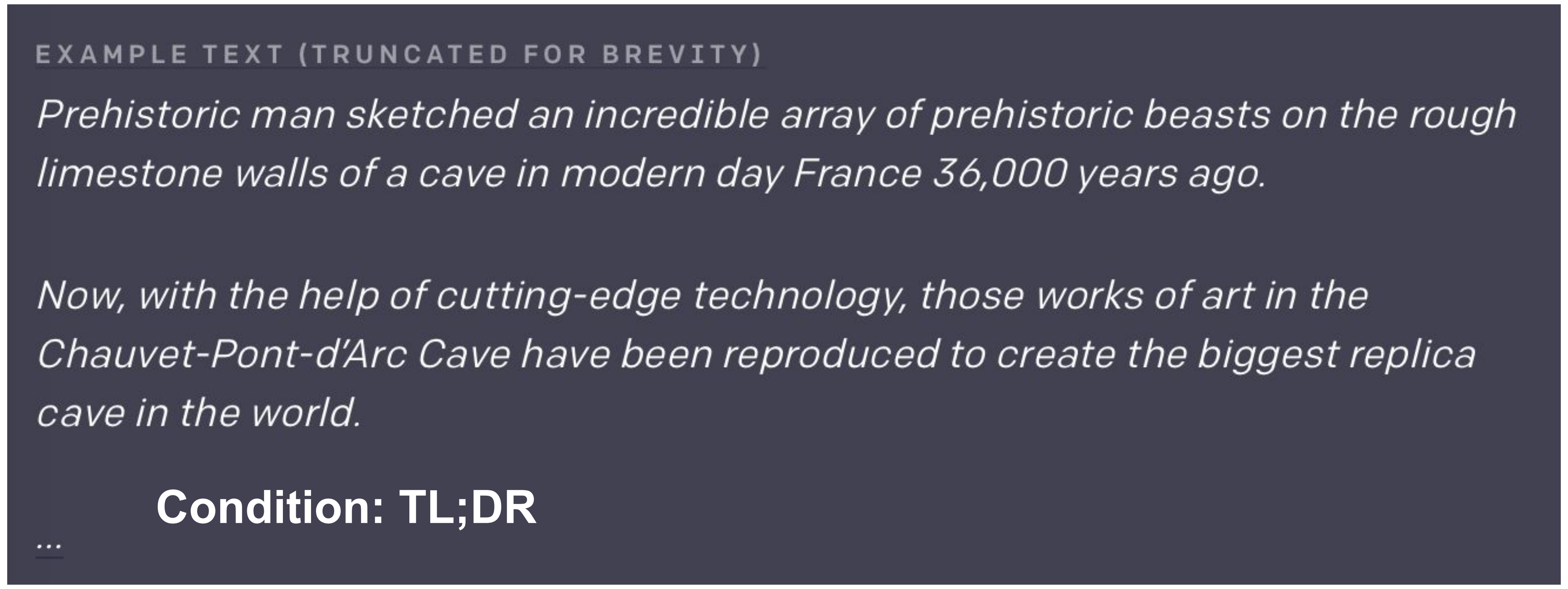
Q: And did they climb any mountains?
A:

Target answers: *unknown* or *yes*
Model answer: Everest



Summarization

- Summarize news articles



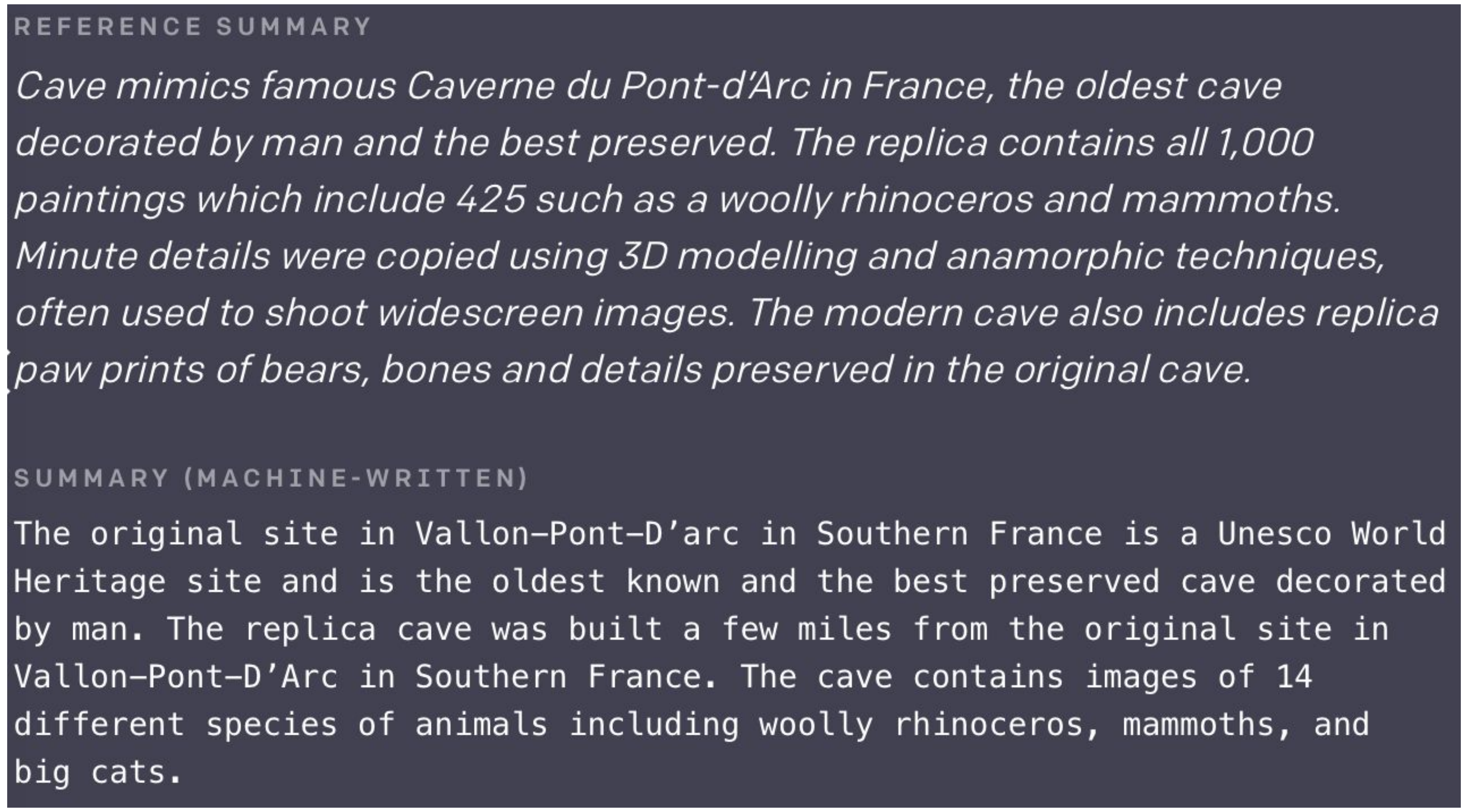
EXAMPLE TEXT (TRUNCATED FOR BREVITY)

Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago.

Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.

Condition: TL;DR

…

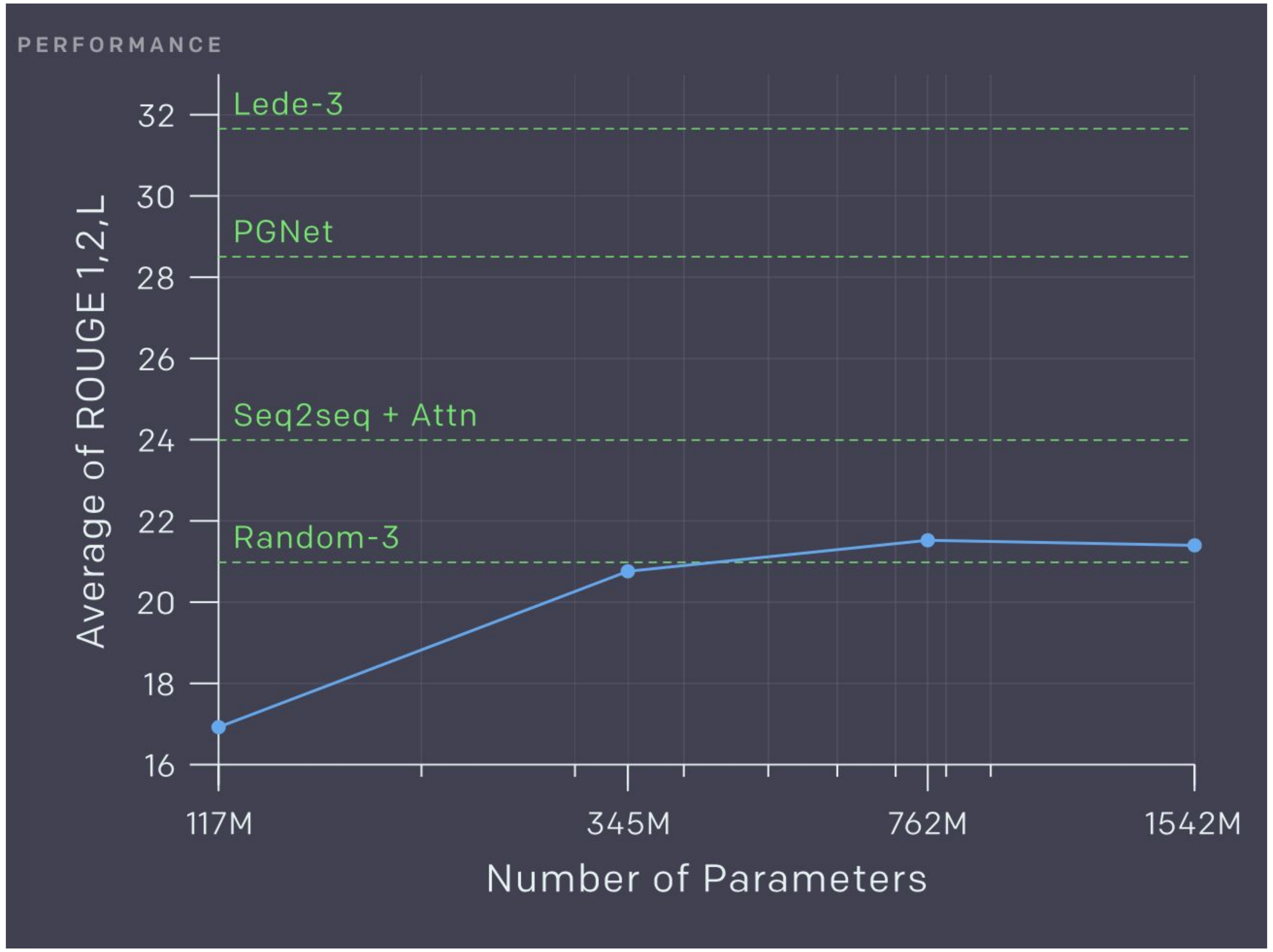


REFERENCE SUMMARY

Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.

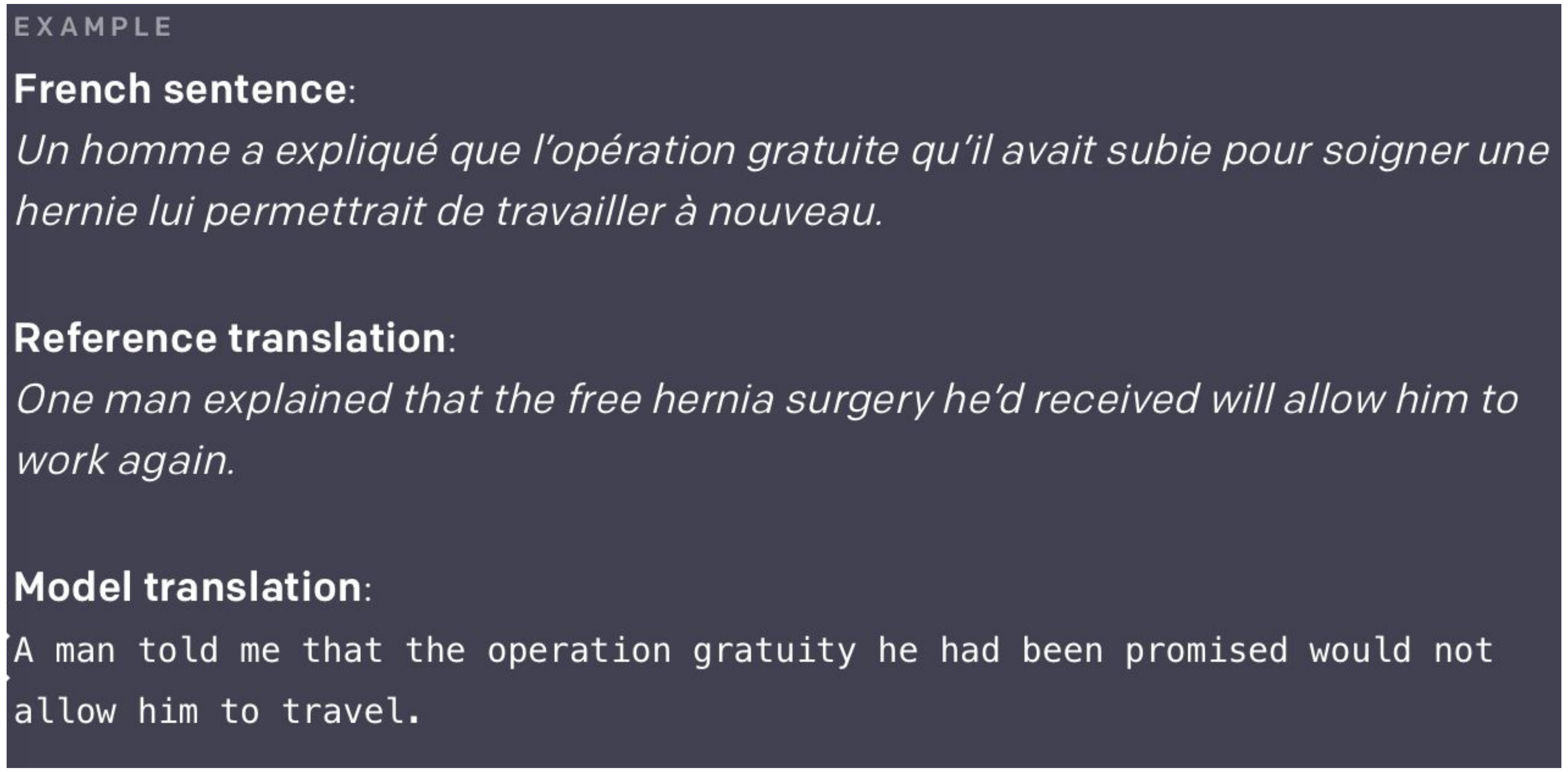
SUMMARY (MACHINE-WRITTEN)

The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.



Translation

- Translate French sentences to English



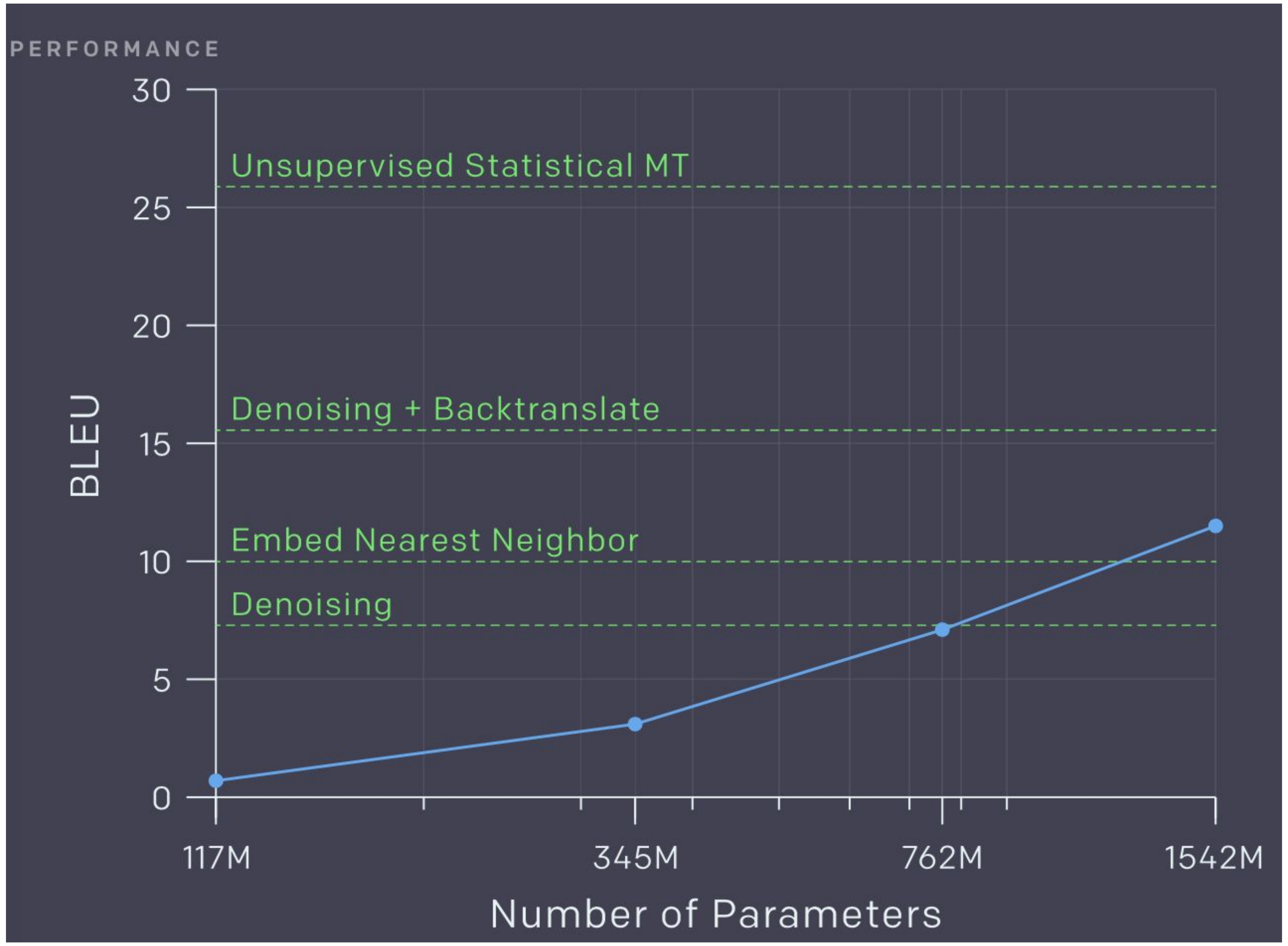
EXAMPLE

French sentence:
Un homme a expliqué que l'opération gratuite qu'il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.

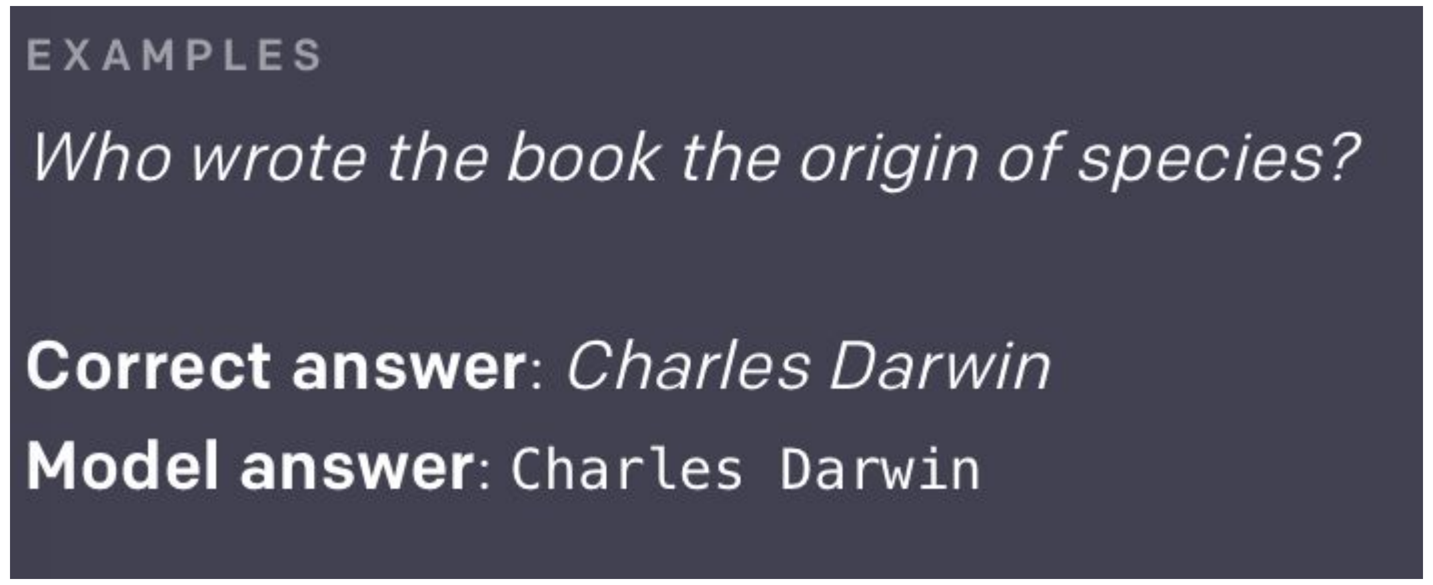
Reference translation:
One man explained that the free hernia surgery he'd received will allow him to work again.

Model translation:

```
A man told me that the operation gratuity he had been promised would not
allow him to travel.
```



Question Answering

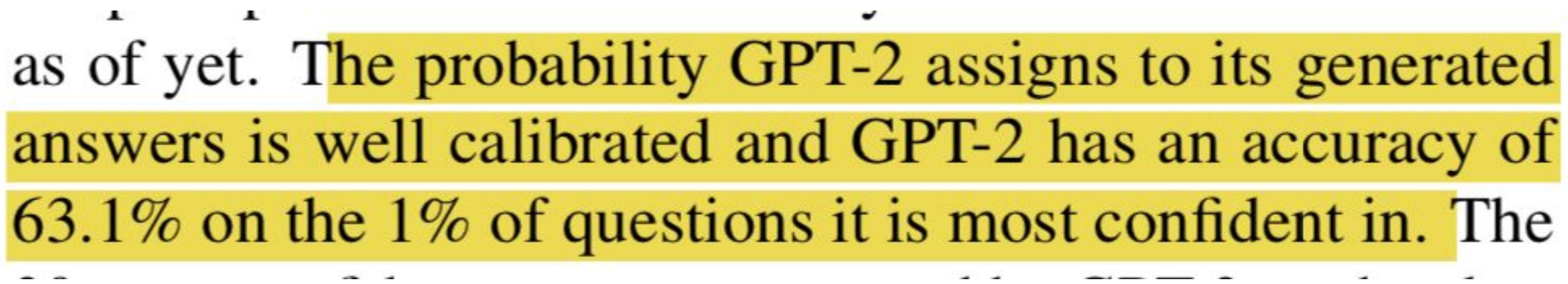


EXAMPLES

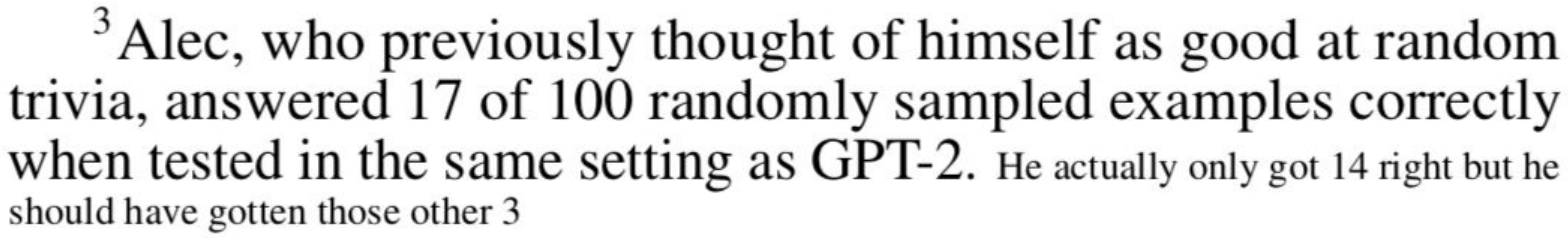
Who wrote the book the origin of species?

Correct answer: *Charles Darwin*

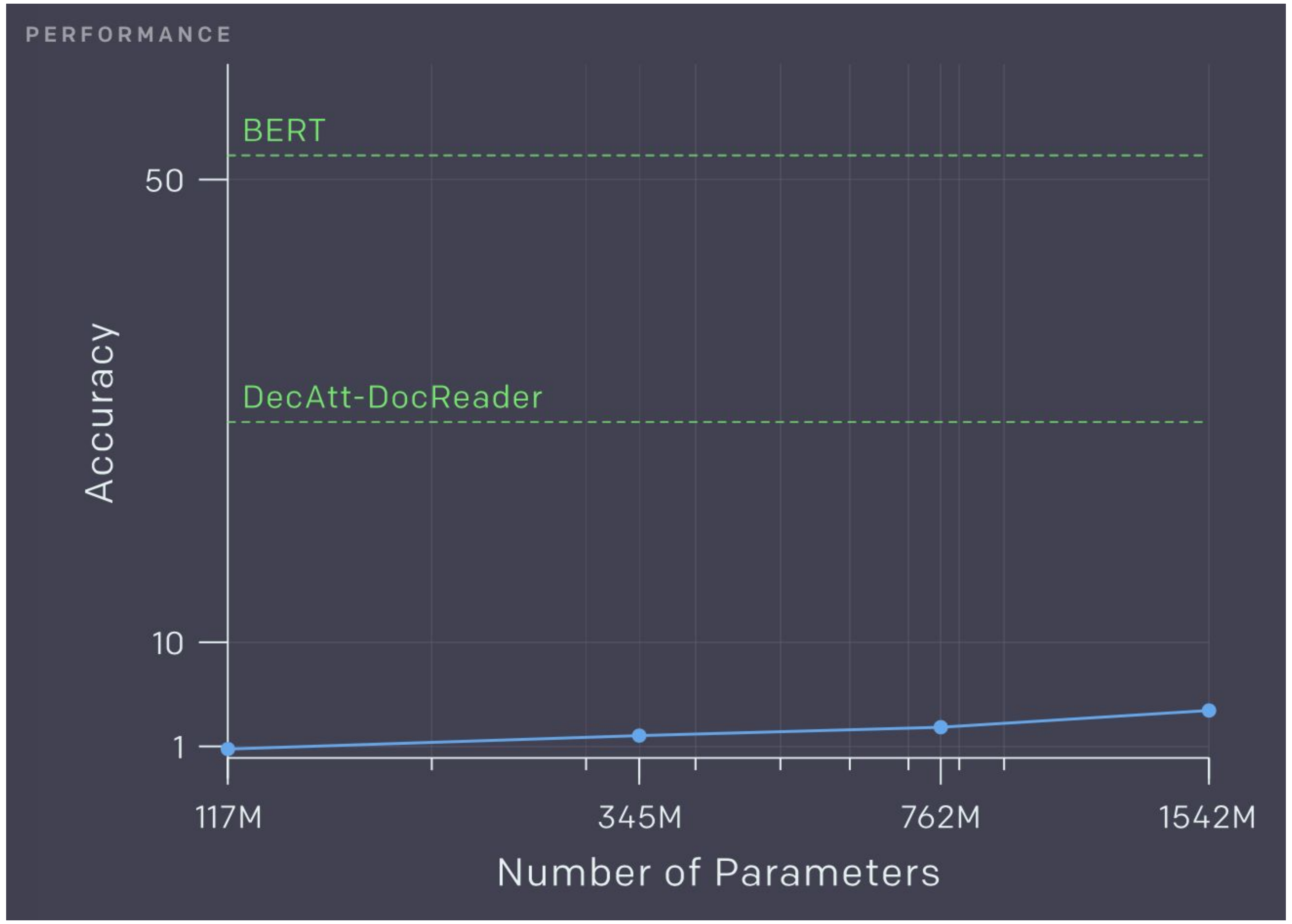
Model answer: `Charles Darwin`



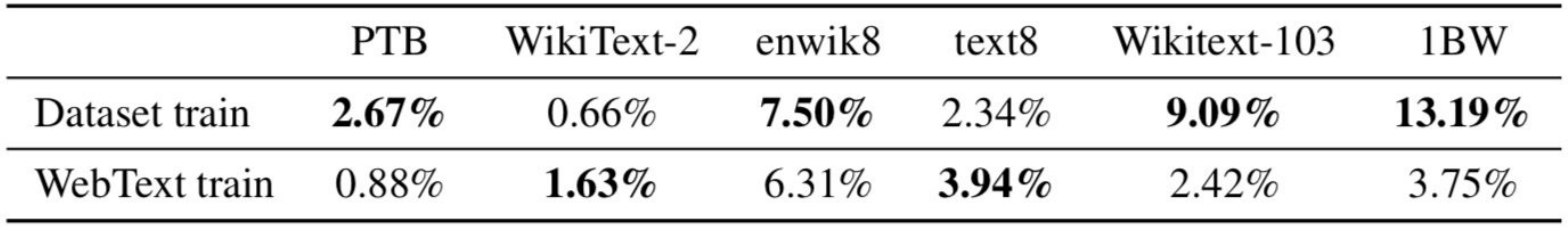
as of yet. The probability GPT-2 assigns to its generated answers is well calibrated and GPT-2 has an accuracy of 63.1% on the 1% of questions it is most confident in. The



[3] Alec, who previously thought of himself as good at random trivia, answered 17 of 100 randomly sampled examples correctly when tested in the same setting as GPT-2. He actually only got 14 right but he should have gotten those other 3



Generalization vs. Memorization



	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.

- It is Important to analyze how much test data also shows up in the training data
- Created Bloom filters containing 8-grams of training data set tokens
  - Calculate percentage of 8-grams from test dataset that are also found in the training set
- Overall, the data overlap between training set and test set are small but provides consistent benefit to reported results

Discussion

- Our results suggest that **unsupervised task learning is promising area of research** to explore
- On **reading comprehension** the performance of GPT-2 is **competitive with supervised baselines**
- However on **summarization its performance is still rudimentary**
- There are undoubtedly many practical tasks where performance of GPT-2 is still no better than random
  - QA & Translation LM only begin to outperform trivial baselines
- **Plan to investigate fine-tuning** on benchmarks such as decaNLP and GLUE
  - Investigate whether GPT-2 + Fine-tuning can overcome inefficient of uni-directional representations demonstrated by BERT