.UBA fiuba

FACULTAD DE INGENIERÍA

NLP

Grandes modelos de lenguaje: GPT

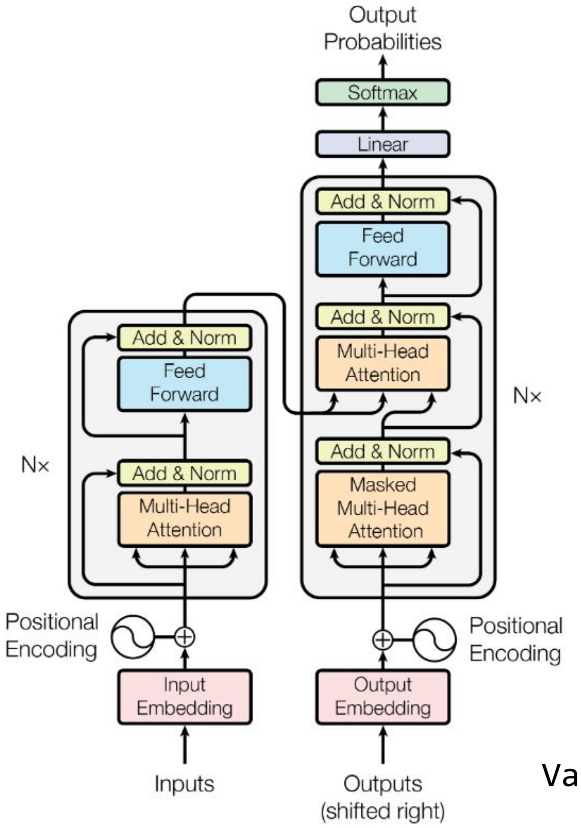
Dr. Rodrigo Cardenas Szigety
rodrigo.cardenas.sz@gmail.com

Programa de la materia

Clase 1: Introducción a NLP, Vectorización de documentos.
Clase 2: Preprocesamiento de texto, librerías de NLP, bots de información.
Clase 3: Word Embeddings, CBOW y SkipGRAM, entrenamiento de embeddings.
Clase 4: Redes recurrentes (RNN), problemas de secuencia y estimación de próxima palabra.
Clase 5: Redes LSTM, análisis de sentimientos.
Clase 6: Modelos Seq2Seq, traductores y bots conversacionales.
Clase 7: Celdas con Attention. Transformers, BERT & ELMo, fine tuning.
Clase 8: Grandes modelos de lenguaje, GPT, cierre del curso.

*Unidades con desafíos a presentar al finalizar el curso.
*Último desafío y cierre del contenido práctico del curso.

GPT: Generative Pre-trained Transformer



Vanilla Transformer

Generativo: es un modelo que permite generar nuevos datos (texto)

Pre-entrenado: no se entrena para una tarea particular, sino que se entrena de forma no-supervisada.

Transformer: arquitectura del modelo.

Vanilla Transformer(big) y BERT(large)

Más de 30k tokens de entrenamiento.

Tokenización BPE
(Byte-Pair encoding)

6 bloques transformers encoder-decoder (24 BERT sólo encoder)

8 attention heads (16 en BERT) REVISAR!!

1024 dimensión de estado oculto

4096 neuronas en capas feed forward

213M parámetros (340M BERT)

Tarea de modelo de lenguaje

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

Corpus de entrenamiento: Datasets de traducción Inglés-Alemán e Inglés-Francés (BERT: BooksCorpus y Wikipedia en inglés)

GPT 1 (2018)

LINK

Algo más de 40k tokens de vocabulario.

12 bloques transformers de decoder (sin encoder)

12 attention heads

768 dimensión de estado oculto

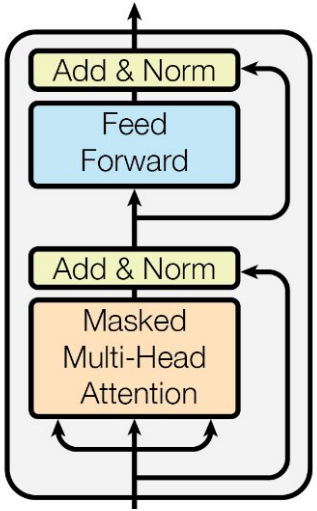
3072 neuronas en capas feed forward

512 tokens de contexto

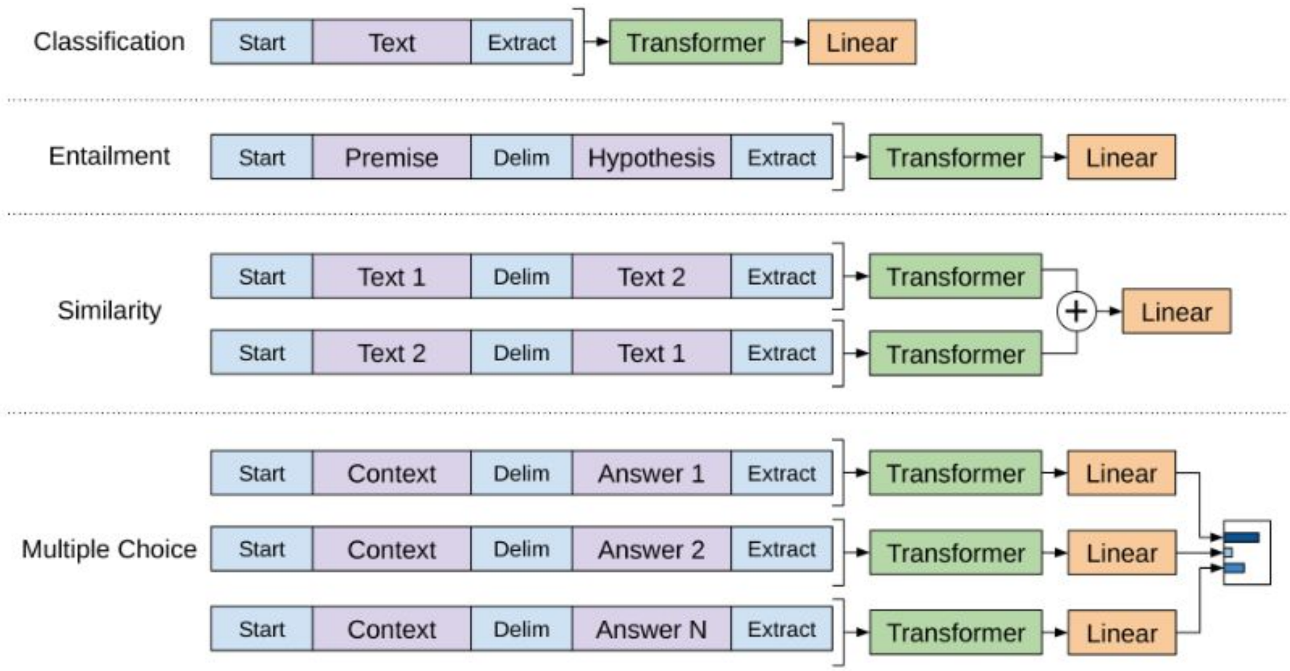
117M parámetros

Se aprende el embedding posicional

Corpus de entrenamiento: BookCorpus (literatura de varios géneros)



GPT 1: el valor del pre-entrenamiento



GPT 2 (2019)

LINK

Más de 50k tokens de vocabulario. (¡ahora múltiples idiomas!)

48 bloques transformers de decoder

48 attention heads

1600 dimensión de estado oculto

6400 neuronas en capas feed forward

1024 tokens de contexto

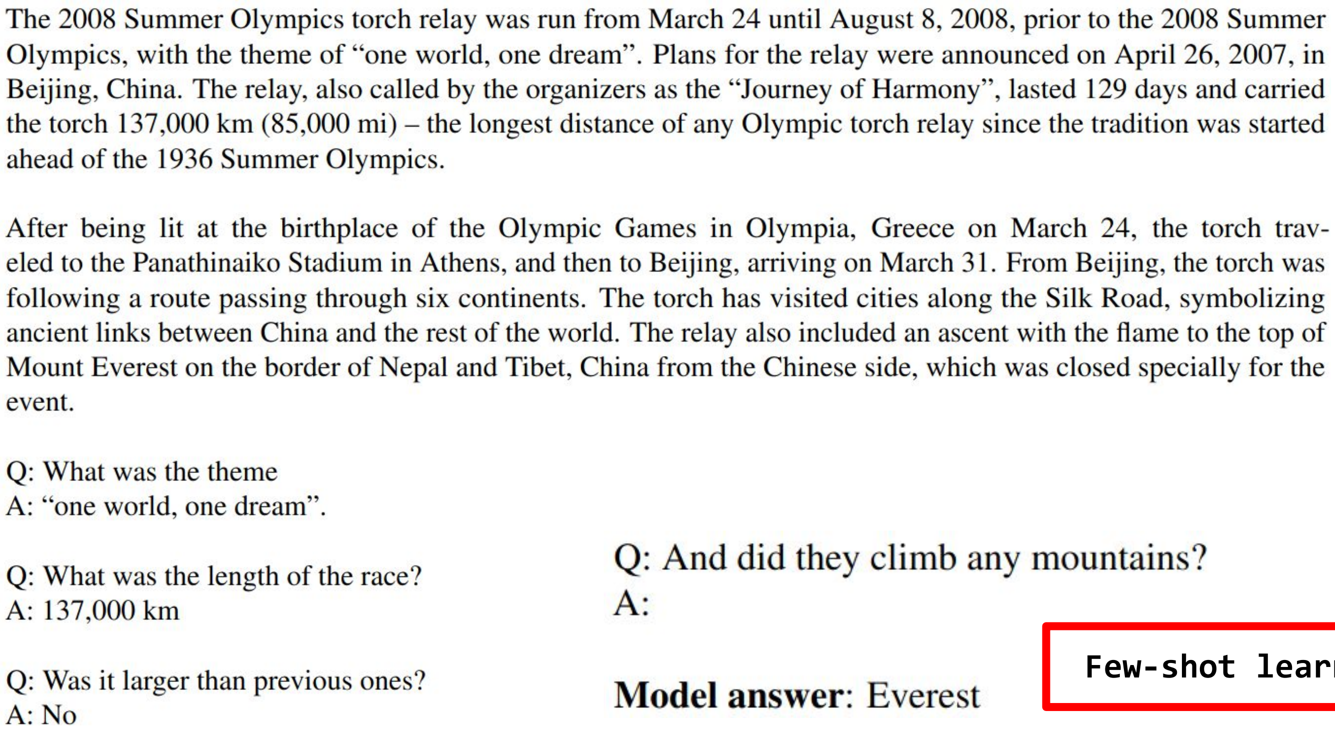
1.54B parámetros

Corpus de entrenamiento: WebText (OpenAI): webs de links salientes de Reddit con karma >= 3 (+ limpieza, + deduplicación - wikipedia) 40GB de texto (8M documentos)

GPT 2: condicionamiento para tareas

$$p(output|input) \qquad p(output|input, task)$$

(answer the question, document, question, answer)



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: What was the theme
A: “one world, one dream”.

Q: What was the length of the race?
A: 137,000 km

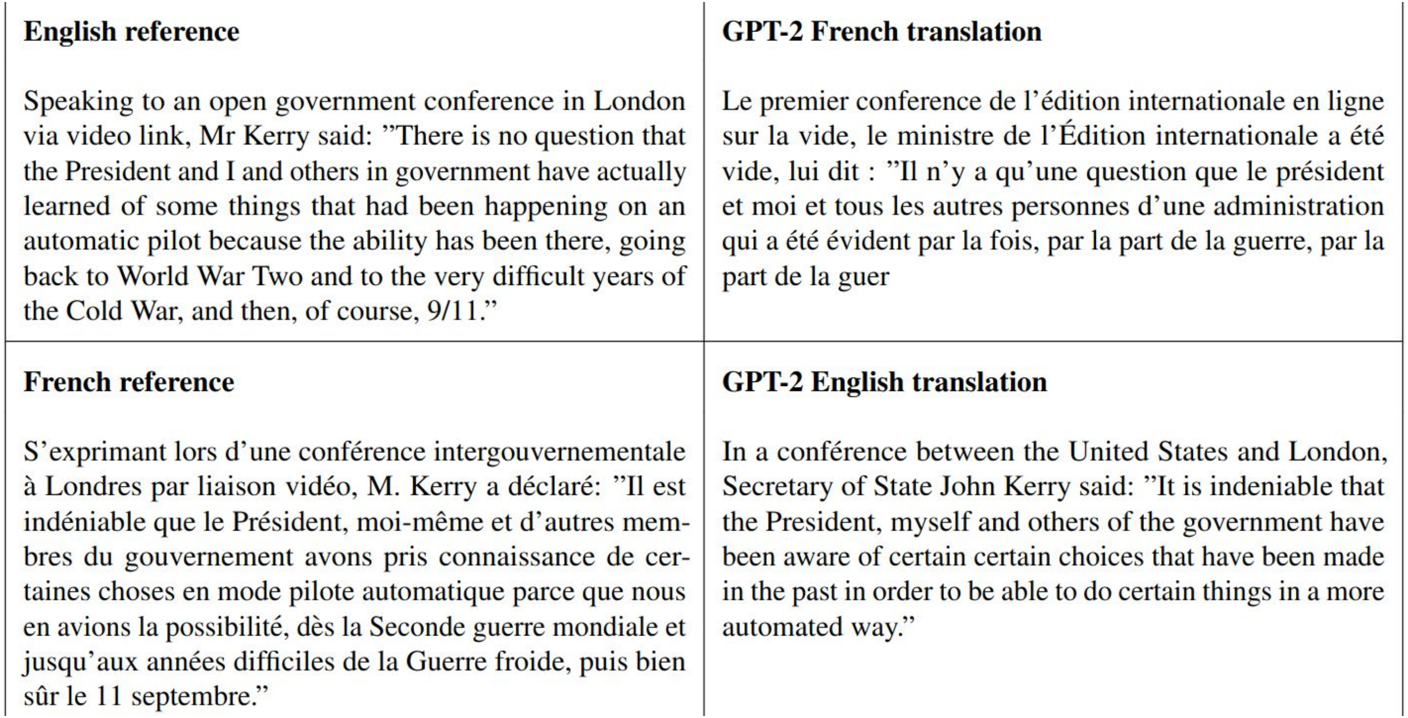
Q: Was it larger than previous ones?
A: No

Q: And did they climb any mountains?
A:

Model answer: Everest

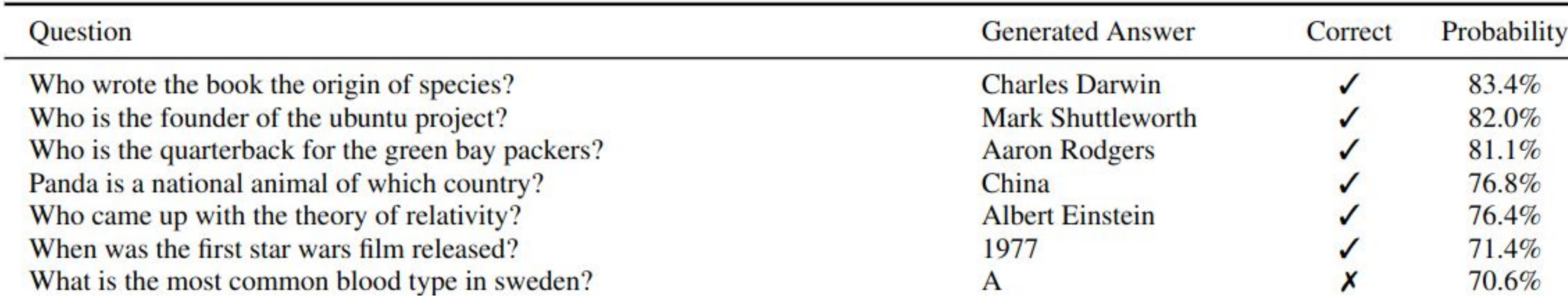
Few-shot learning

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



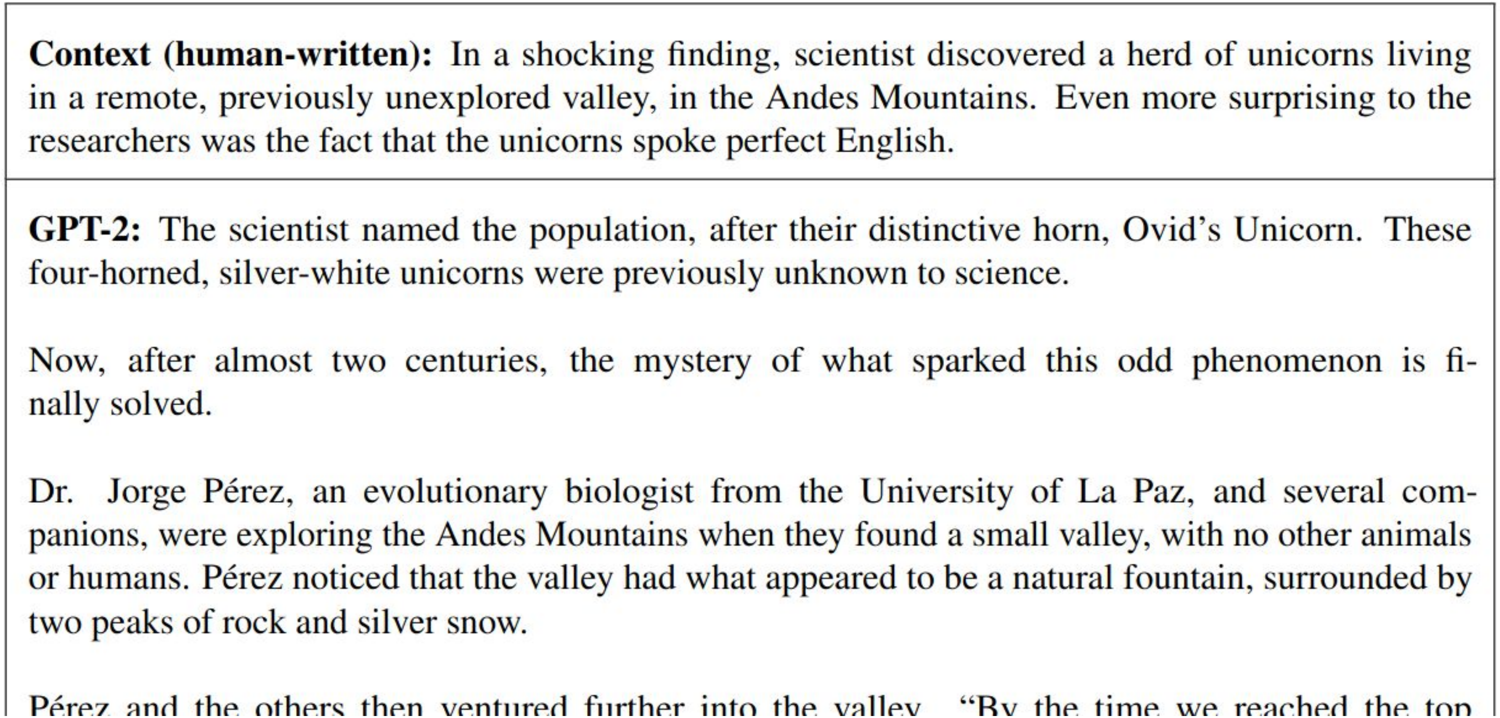
English reference	GPT-2 French translation
Speaking to an open government conference in London via video link, Mr Kerry said: ”There is no question that the President and I and others in government have actually learned of some things that had been happening on an automatic pilot because the ability has been there, going back to World War Two and to the very difficult years of the Cold War, and then, of course, 9/11.”	Le premier conference de l’édition internationale en ligne sur la vide, le ministre de l’Édition internationale a été vide, lui dit : ”Il n’y a qu’une question que le président et moi et tous les autres personnes d’une administration qui a été évident par la fois, par la part de la guerre, par la part de la guer
French reference	**GPT-2 English translation**
S’exprimant lors d’une conférence intergouvernementale à Londres par liaison vidéo, M. Kerry a déclaré: ”Il est indéniable que le Président, moi-même et d’autres membres du gouvernement avons pris connaissance de certaines choses en mode pilote automatique parce que nous en avions la possibilité, dès la Seconde guerre mondiale et jusqu’aux années difficiles de la Guerre froide, puis bien sûr le 11 septembre.”	In a conférence between the United States and London, Secretary of State John Kerry said: ”It is indeniable that the President, myself and others of the government have been aware of certain certain choices that have been made in the past in order to be able to do certain things in a more automated way.”

Zero-shot learning



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%

One-shot learning



Context (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.

GPT-2: 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

GPT 3 (2020)

LINK

Más de 50k tokens de vocabulario.

96 bloques transformers de decoder

96 attention heads

12288 dimensión de estado oculto

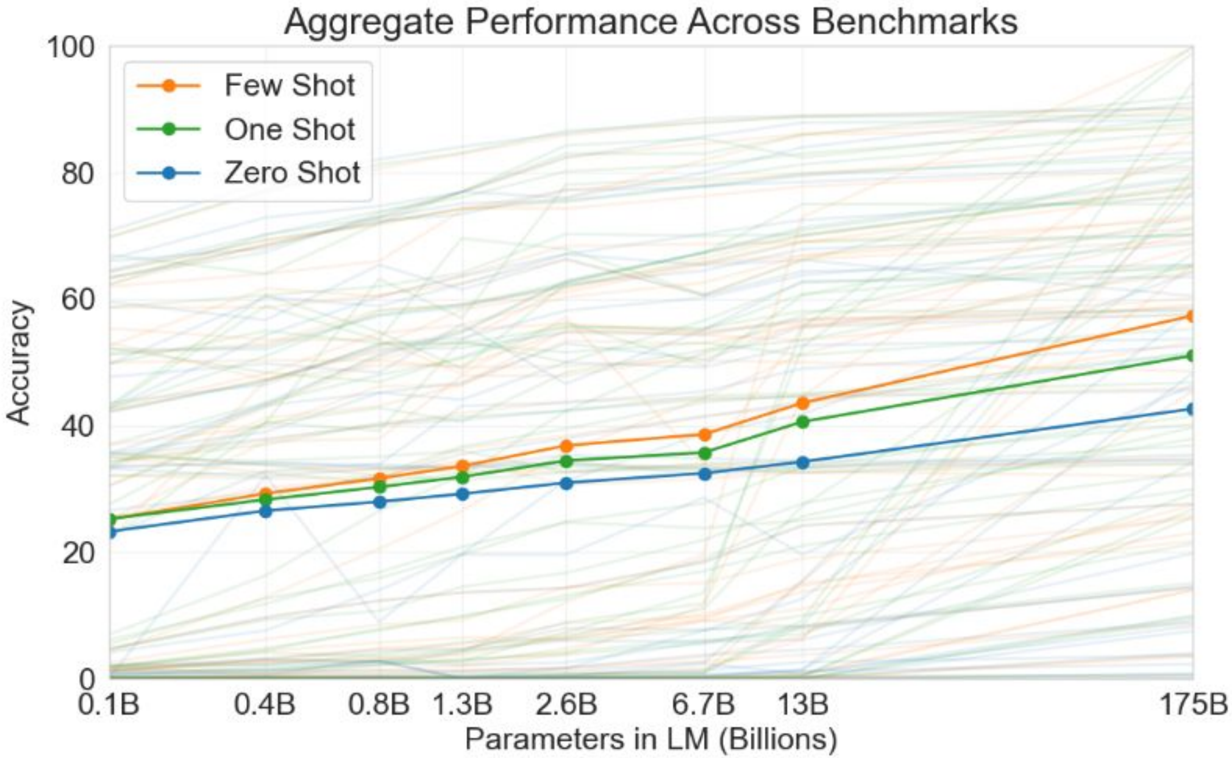
49152 neuronas en capas feed forward

2048 tokens de contexto

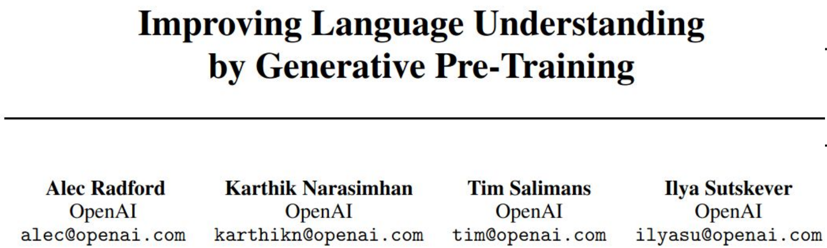
175B parámetros

Corpus de entrenamiento: CommonCrawl (filtrado, 570GB de texto comprimido). WebText expandido. Wikipedia en inglés. 2 corpora de libros.

GPT 3: modelos más grandes aprenden más efectivamente FS, 1S y 0S



Los equipos (e infraestructura) también son large



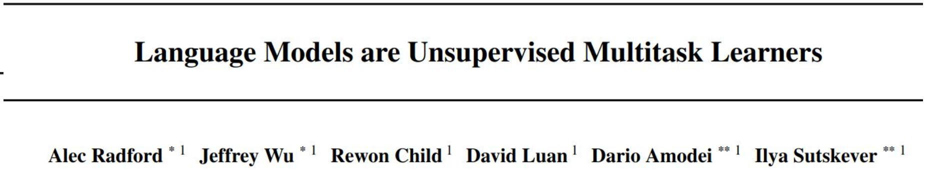
Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

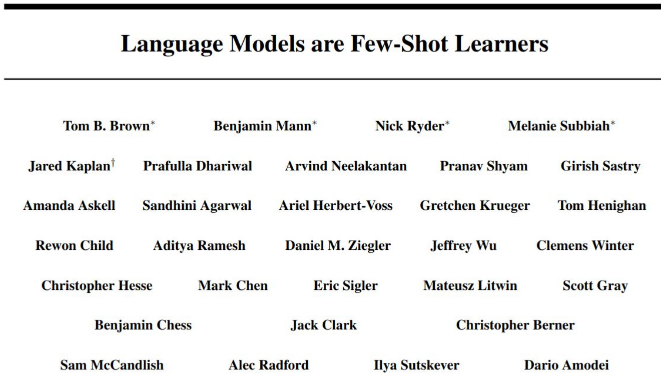
Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com



Language Models are Unsupervised Multitask Learners

Alec Radford[* 1] Jeffrey Wu[* 1] Rewon Child[1] David Luan[1] Dario Amodei[** 1] Ilya Sutskever[** 1]



Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

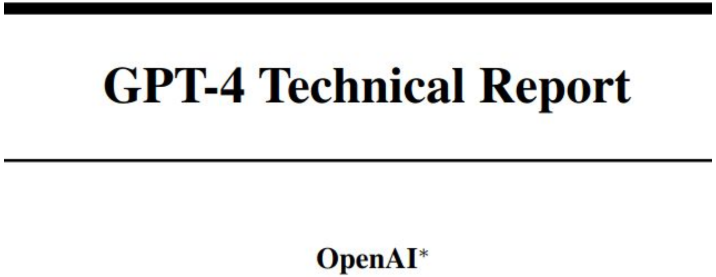
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei



GPT-4 Technical Report

OpenAI[*]