

# Large language models

Can predicting the  
next word do  
something useful?

# GPT-2 (2019)

"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.

## Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life .. for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?

A: 54

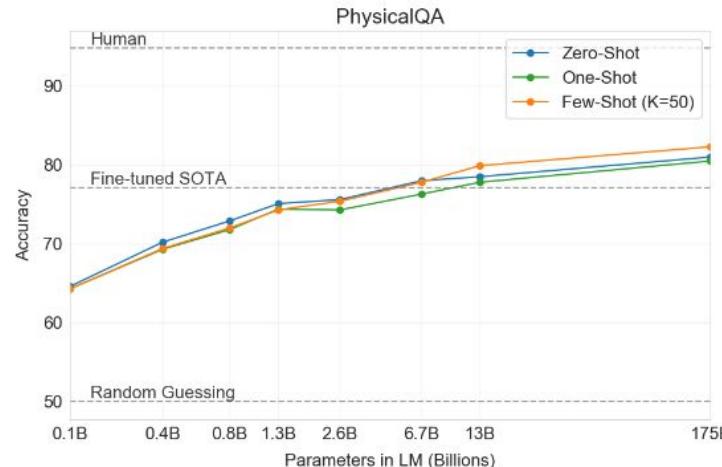
Q: where does she live?

A:

**Model answer:** Stockholm

**Turker answers:** Sweden, Sweden, in Sweden, Sweden

# GPT-3 (2020)



**Figure 3.6:** GPT-3 results on PIQA in the zero-shot, one-shot, and few-shot settings. The largest model achieves a score on the development set in all three conditions that exceeds the best recorded score on the task.

## a. Shape, Material, and Purpose

[Goal] Make an outdoor pillow

[Sol1] Blow into a **tin can** and tie with rubber band ✗  
[Sol2] Blow into a **trash bag** and tie with rubber band ✓

[Goal] To make a hard shelled taco,

[Sol1] put seasoned beef, cheese, and lettuce **onto** the hard ✗ shell.

[Sol2] put seasoned beef, cheese, and lettuce **into** the hard ✓ shell.

[Goal] How do I find something I lost on the carpet?

[Sol1] Put a **solid seal** on the end of your vacuum and turn it ✗ on.  
[Sol2] Put a **hair net** on the end of your vacuum and turn it on. ✓

# More history

- Instruction tuning
  - Turns “next word predictor” into a chatbot
- GPT-3.5 (ChatGPT) - Nov. 2022
- GPT-4 - Mar 2023
- Vision-language models, reasoning, ...

# Prompting

## (a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. X

## (b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) *The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4.* ✓

## (c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 X

## (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

# Prompting



Andrej Karpathy ✅ @karpathy · Dec 7



Don't think of LLMs as entities but as simulators. For example, when exploring a topic, don't ask:



"What do you think about xyz"?

There is no "you". Next time try:

"What would be a good group of people to explore xyz? What would they say?"

The LLM can channel/simulate many perspectives but it hasn't "thought about" xyz for a while and over time and formed its own opinions in the way we're used to. If you force it via the use of "you", it will give you something by adopting a personality embedding vector implied by the statistics of its finetuning data and then simulate that. It's fine to do, but there is a lot less mystique to it than I find people naively attribute to "asking an AI".



1.1K



3.3K



27K



3.5M



# API demo

See docs for your favourite LM, e.g.:

<https://ai.google.dev/gemini-api/docs/quickstart>

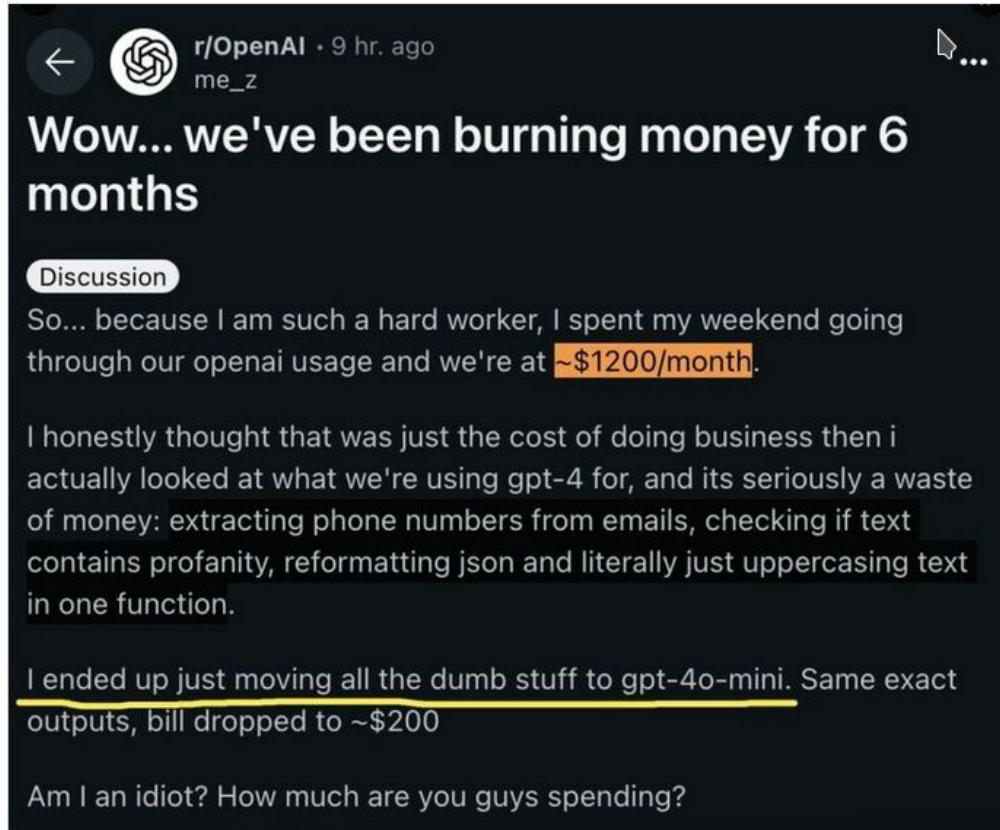
<https://platform.openai.com/docs/quickstart>

<https://platform.claude.com/docs/en/get-started>

# I can use this for many things (with many caveats)

- Converting documents to structured data
- Generating training data
- Improving search

# When using LLM is not a good idea



A screenshot of a Reddit post from the r/OpenAI subreddit. The post was made 9 hours ago by user me\_z. The title is "Wow... we've been burning money for 6 months". The post content discusses the high cost of OpenAI usage, mentioning a bill of approximately \$1200 per month for extracting phone numbers from emails, checking for profanity, reformatting JSON, and uppercasing text. The user then switched to gpt-4o-mini, which resulted in a bill drop to about \$200. The post ends with a question about the user's spending.

← r/OpenAI • 9 hr. ago  
me\_z

...

## Wow... we've been burning money for 6 months

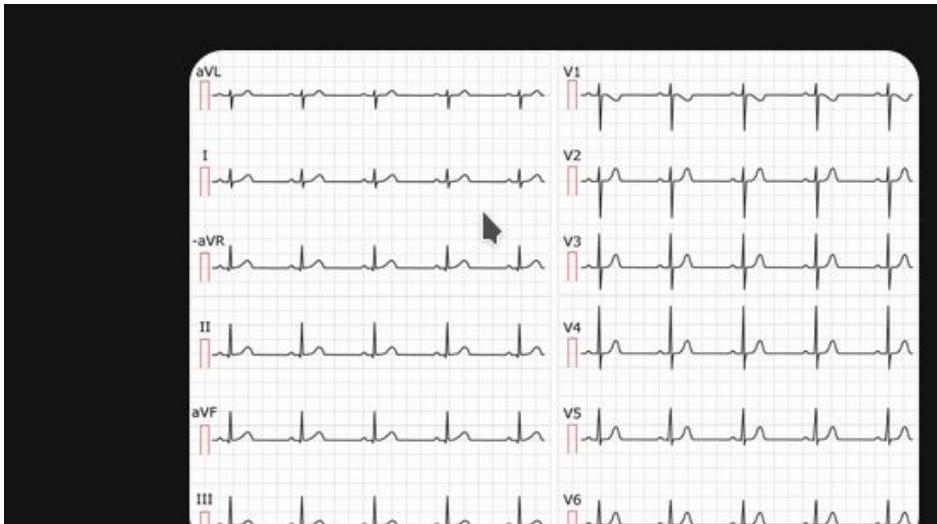
**Discussion**

So... because I am such a hard worker, I spent my weekend going through our openai usage and we're at **~\$1200/month**.

I honestly thought that was just the cost of doing business then i actually looked at what we're using gpt-4 for, and its seriously a waste of money: extracting phone numbers from emails, checking if text contains profanity, reformatting json and literally just uppercasing text in one function.

I ended up just moving all the dumb stuff to gpt-4o-mini. Same exact outputs, bill dropped to **~\$200**

Am I an idiot? How much are you guys spending?



Does this ECG show signs of a myocardial infarction?

**Yes, this ECG shows signs of pathology.** While it does not show the massive "tombstone" ST-elevation seen in some heart attacks, the distinct ST-segment depression in the anterior leads is a serious finding consistent with heart ischemia or a specific type of heart attack (Posterior or NSTEMI).

# <1M parameter NN is much better

ECG measurements — Core AI

Heart rate	Axis	P Wave	PR Interval	QRS Duration	QT Interval	QTcFra	RR Interval	PP Interval
54 BPM	Normal	118 ms	174 ms	99 ms	455 ms	440 ms	1099 ms	1099 ms

Visualization      Original

Low confidence (20%)  
I aVR V1 V4

Low confidence (11%)  
II aVL V2 V5

Mid confidence (29%)  
III aVF V3 V6

Low confidence (24%)  
aVR V1 V4

① Understanding the AI's decision

Patient outcome

Queen of Hearts™

⚠ STEMI equivalent

Immediate invasive strategy or equivalent management is recommended.

Presentation last edited by Steve Smith, Oct 14, 1:26 PM

Hide explainability

LVsense™

⚠ Reduced LVEF (<40%) detected

Refer to confirmatory echocardiography.

Overall ECG assessment — Core AI

Rhythm  
Sinus bradycardia

# Structured text extraction

rýchlotest SARS-CoV-2: protilátky IgM: POZIT, IgG: POZIT

S-anti-SARS-CoV-2 IgM (ELISA)/1.12.2020/	55,5	NTU	pozitívny
S-anti-SARS-CoV-2 IgG (ELISA)/1.12.2020/	34,7	NTU	pozitívny

09.12.2020- PROTILÁTKY SARS - CoV - 2: IgM: POZIT IgG: POZIT

Rýchlotest na vyšetrenie protilátkov proti SARS2-COV: IgM silno pozit, IgG slabo pozit

16.3.2021 Protilátky SARS-CoV-2: IgM:pozit., IgG: pozit

Rýchlotest na protilátky SARS CoV-12: IgM slabo pozit. IgG negat.

Protilátky COVID 19 rýchlotest z kapilárnej krvi: IgG: POZIT IgM: NEGAT

Rýchlotest na protilátky SARS-CoV2 (z kapilárnej krvi): IgM pozitívne, IgG pozitívne

PROTILÁTKY SARS - CoV - 2:/5.1.21/.:IgM: POZIT, IgG: POZIT

Protilátky na SARS-Cov-2: 29.12. IgM, IgG negat., 5.1.21 IgM,IgG pozit.

Obr. 3.3: Rôzne spôsoby zápisu testov na protilátky v prepúšťacích správach

## Here's 50 confirmed hallucinations in ICLR 2026 submissions

In the table below, we've included a specific human-verified hallucitation our tool flagged in each paper. According to the ICLR's editorial policy, even a single, clear hallucitation is an ethics violation that could lead to the paper's rejection. Given that we've only scanned 300 out of 20,000 submissions, we estimate that we will find 100s of hallucinated papers in the coming days. We are collaborating with the ICLR program chairs, who have been very responsive and thoughtful with their challenging assignment.

Title	Average Review Rating	Paper Link	Citation Check Scan Link	Example of Verified Hallucination	Comment
TamperTok: Forensics-Driven Tokenized Autoregressive Framework for Image Tampering Localization	8.0	<a href="https://app.gptzero.me/documents/4645494f-70eb-40bb-aea7-0007e13f7179/share">TamperTok: Forensics-Driven Tokenized Autoregressive Framework for Image Tampering Localization   OpenReview</a>	<a href="https://app.gptzero.me/documents/4645494f-70eb-40bb-aea7-0007e13f7179/share">https://app.gptzero.me/documents/4645494f-70eb-40bb-aea7-0007e13f7179/share</a>	Chong Zou, Zhipeng Wang, Ziyu Li, Nan Wu, Yuling Cai, Shan Shi, Jiawei Wei, Xia Sun, Jian Wang, and Yizhou Wang. Segment everything everywhere all at once. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023.	This paper exists, but all authors are wrong.
MixtureVitae: Open Web-Scale Pretraining Dataset With High Quality Instruction and Reasoning Data Built from Permissive Text Sources	8.0	<a href="https://app.gptzero.me/documents/bfd10666-ea2d-454c-9ab2-75faa8b84281/share">MixtureVitae: Open Web-Scale Pretraining Dataset With High Quality Instruction and Reasoning Data Built from Permissive Text Sources   OpenReview</a>	<a href="https://app.gptzero.me/documents/bfd10666-ea2d-454c-9ab2-75faa8b84281/share">https://app.gptzero.me/documents/bfd10666-ea2d-454c-9ab2-75faa8b84281/share</a>	Dan Hendrycks, Collin Burns, Steven Basart, Andy Critch, Jerry Li, Dawn Ippolito, Aina Lapedriza, Florian Tramer, Rylan Macfarlane, Eric Jiang, et al. Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.	The paper and first 3 authors match. The last 7 authors are not on the paper, and some of them do not exist

# Importance of evaluation

- Super easy to build demoable system
- Super hard to improve
- Do not be afraid to get your hands dirty and annotate data
  - For text data, I recommend:
    - Exporting to .csv
    - Opening in Excel
    - Grabbing 4 cans of your favourite beverage, put on good music
    - Annotate

# Local LLM demo