

PRESENTATION - PART 1

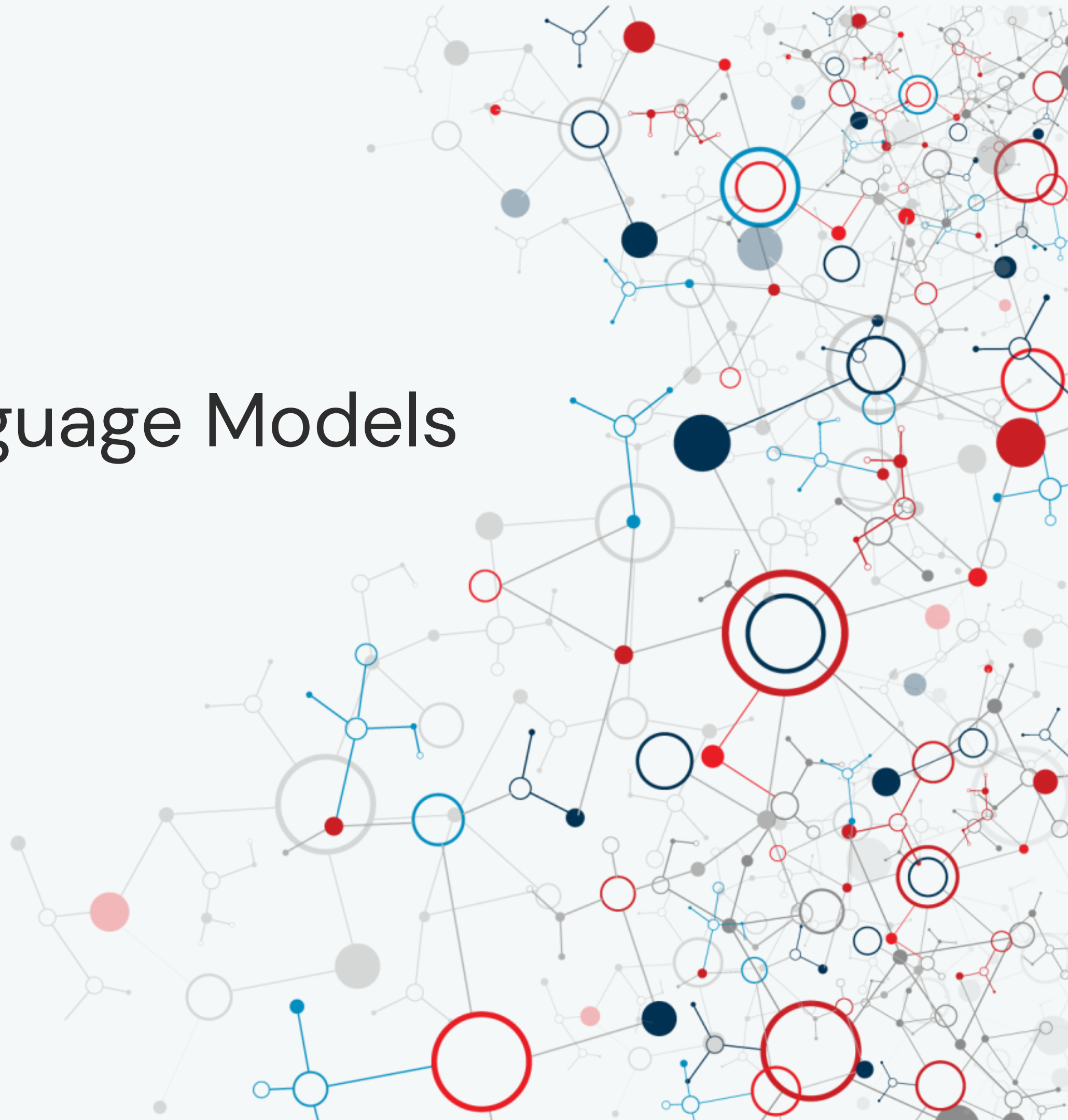
Introducing Large Language Models

GPT-3

Codex

GitHub Copilot

Łukasz Zalewski
Kacper Włodarczyk
Adrian Zjawiński



TECH NEWS

What are large language models (LLMs), why have they become controversial?

(TECHNOLOGY) Large language models guide our AI training and recently, ethicists have pointed out serious flaws in LLMs (which cost some their jobs).



Published on August 30, 2021

By **Jack Lloyd**, Staff Writer

<https://theamericangenius.com/tech-news/large-language-models/>

A GPT-3 bot posted comments on Reddit for a week and no one noticed

Under the username /u/thegentlemetre, the bot was interacting with people on /r/AskReddit, a popular forum for general chat with 30 million users.

<https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/>

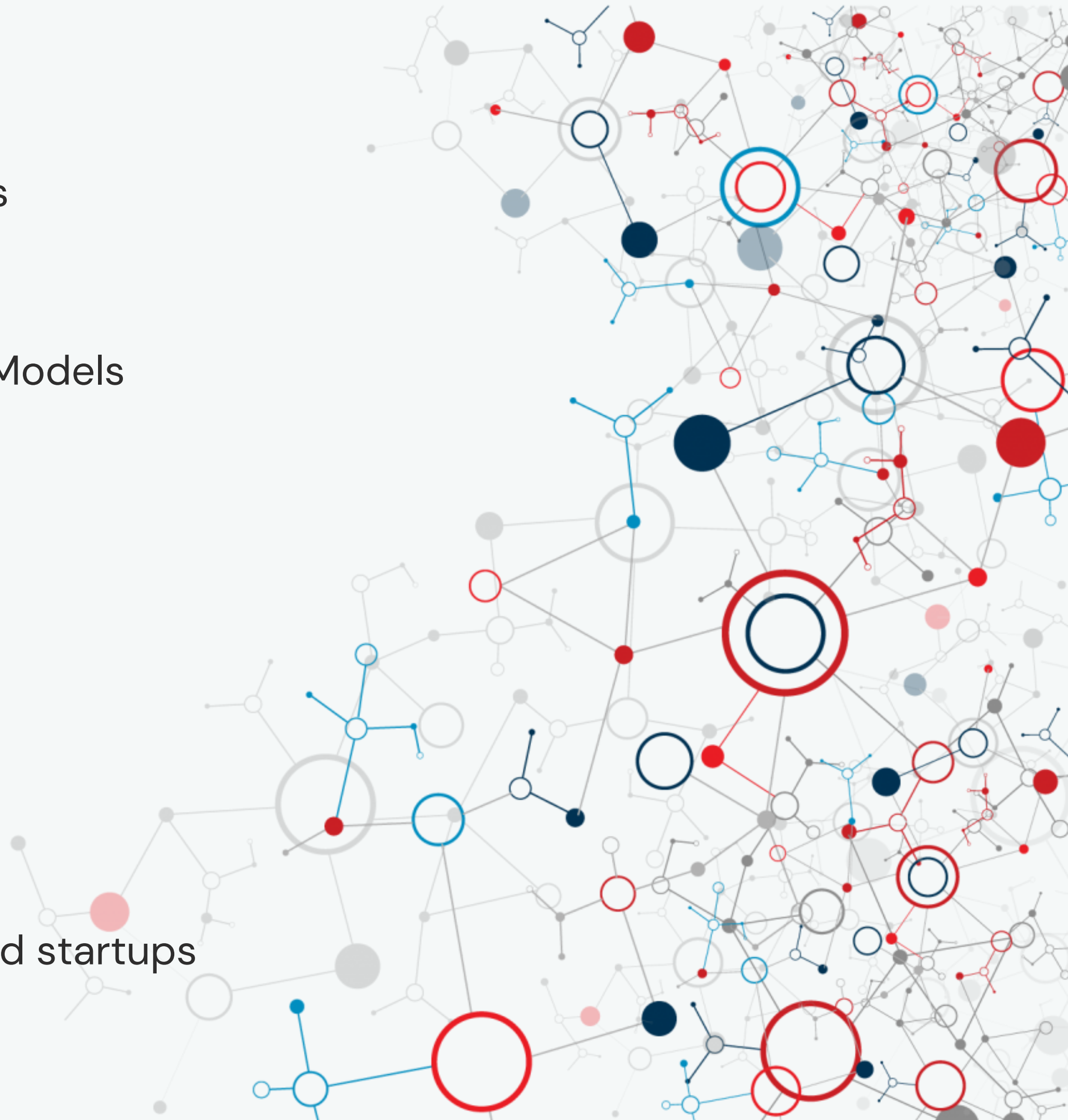
A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

He says he wanted to prove the AI could pass as a human writer

<https://www.theverge.com/2020/8/16/21371049/gpt3-hacker-news-ai-blog>

Agenda

1. Introduction to Large Language Models
2. GPT-3 examples and playground
3. The art of Prompt Engineering
4. How do Transformers work?
5. Reasoning abilities of Large Language Models
6. Prompt Engineering vs Fine-tuning
7. Use cases for Large Language Models
8. Semantic Search
9. Text Classification
10. Question Answering
11. Embeddings
12. API Costs
13. Limitations
14. Codex & GitHub Copilot
15. Open Source Alternatives
16. What should we expect in the future?
17. GPT-3 engines
18. Overview of use cases, applications and startups
19. Usage Guidelines
20. Content safety filters



What Are **Large Language Models**

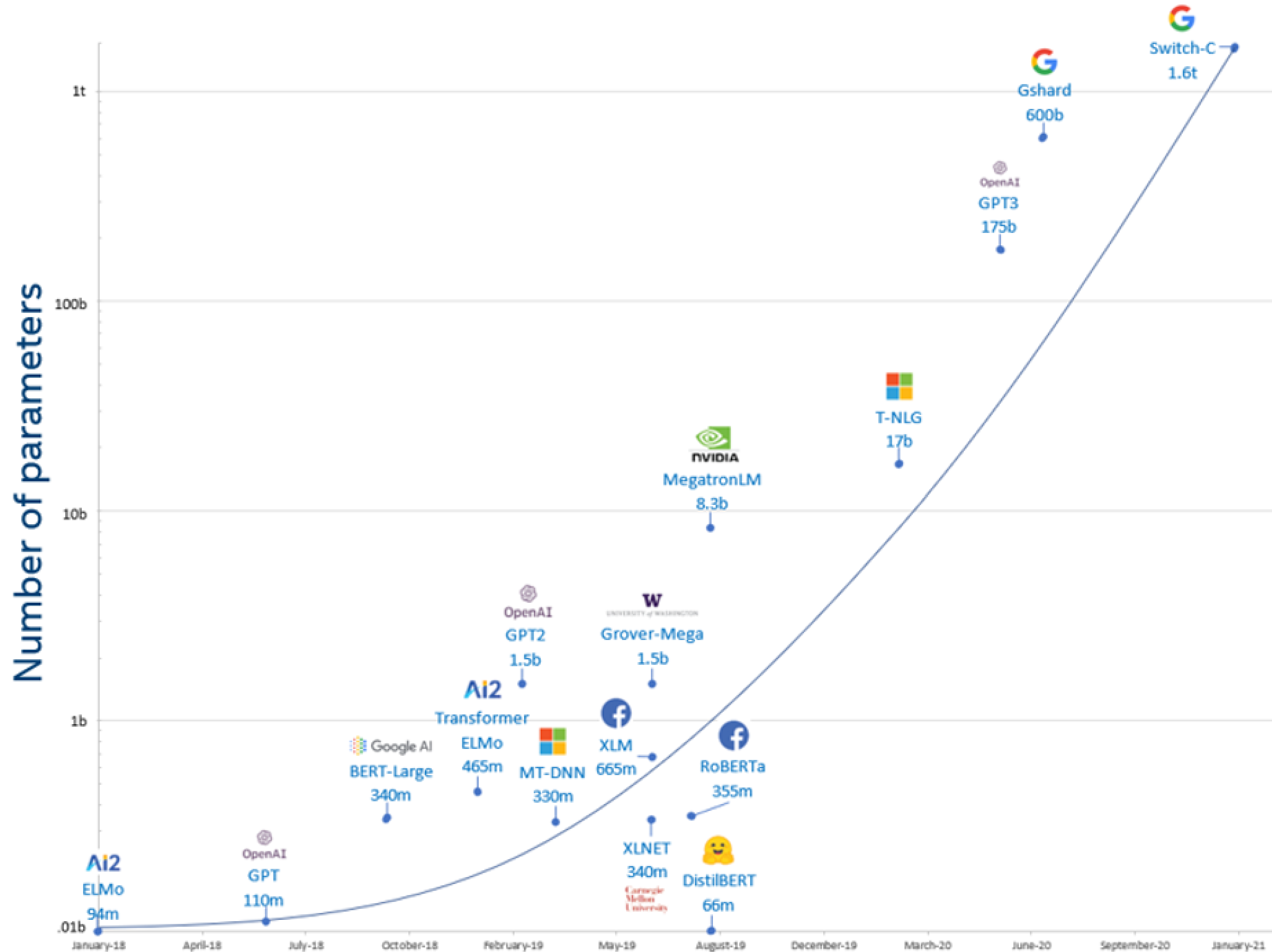
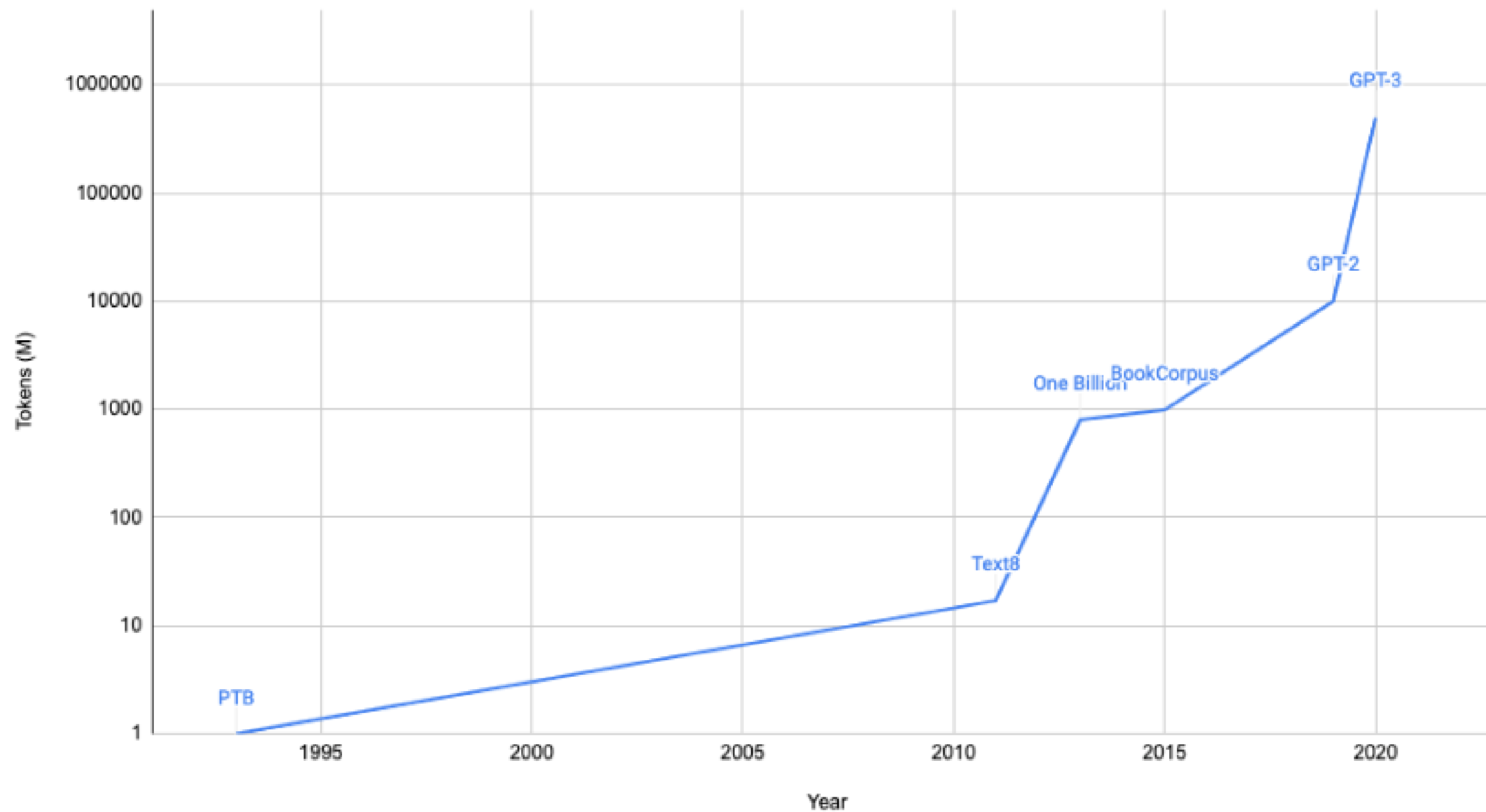


Figure 1: Exponential growth of number of parameters in DL models

Language model datasets over time (log scale)



GPT-3

Generative Pretrained Transformer 3

A type of **Transformer**.

Trained on **~45TB** of internet text data (years 2016–2019).

Training on a single GPU would take **~300 years**.

Has only one objective: **predict the next word given previous words**.

Used in **200+** production-ready software.

Model is **closed-source** (but can be used through the public API).

Who's behind GPT-3



Artificial Intelligence research lab.

Key backers: [Elon Musk](#), [Sam Altman](#), [Microsoft](#).

Shares headquarters with [Neuralink](#).

API for GPT-3 is their first commercial product.

Prompt

The following is a list of companies and the categories they fall into

Facebook: Social media, Technology

LinkedIn: Social media, Technology, Enterprise, Careers

Uber: Transportation, Technology, Marketplace

Unilever: Conglomerate, Consumer Goods

McDonalds: Food, Fast Food, Logistics, Restaurants

FedEx:

Sample response

Logistics, Transportation

**You can try solving any problem that can be
framed in a form of a text!**

Examples

Prompt

This is a tweet sentiment classifier

Tweet: "I loved the new Batman movie!"

Sentiment: Positive

###

Tweet: "I hate it when my phone battery dies."

Sentiment: Negative

###

Tweet: "My day has been 👍"

Sentiment: Positive

###

Tweet: "This is the link to the article"

Sentiment: Neutral

###

Tweet: "This new music video blew my mind"

Sentiment:

Sample response

Positive

Prompt

There are many fruits that were found on the recently discovered planet Gooocrux. There are neoskizzles that grow there, which are purple and taste like candy. There are also loheckles, which are a grayish blue fruit and are very tart, a little bit like a lemon. Pounits are a bright green color and are more savory than sweet. There are also plenty of loopnovas which are a neon pink flavor and taste like cotton candy. Finally, there are fruits called glowls, which have a very sour and bitter taste which is acidic and caustic, and a pale orange tinge to them.

Please make a table summarizing the fruits from Gooocrux

Fruit	Color	Flavor
Neoskizzles	Purple	Sweet
Loheckles	Grayish blue	Tart

Sample response

Pounits	Bright green	Savory
Loopnovas	Neon pink	Cotton candy
Glowls	Pale orange	Sour
Other		

Prompt

My second grader asked me what this passage means:

.....

Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter. [19] When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows,[20] and is on average the third-brightest natural object in the night sky after the Moon and Venus.

.....

I rephrased it for him, in plain language a second grader can understand:

.....

Sample response

Jupiter is a huge planet made of gas. It is the fifth planet from the sun. It is very big and bright. It is the third brightest object in the night sky after the moon and Venus. Jupiter has many moons. One of the moons is bigger than the planet Mercury.

Prompt

Convert my short hand into a first-hand account of the meeting:

Tom: Profits up 50%

Jane: New servers are online

Kjel: Need more time to fix software

Jane: Happy to help

Parkman: Beta testing almost done

Summary:

Sample response

Profits are up, Jane's new servers are online, Kjel needs more time to fix software, and Parkman's beta testing is almost done.

Prompt

Topic: Breakfast

Two-Sentence Horror Story: He always stops crying when I pour the milk on his cereal. I just have to remember not to let him see his face on the carton.

###

Topic: Wind

Two-Sentence Horror Story:

Sample response

I've never seen the wind before. I just know that it's out there, and it's coming for me.

I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown".

Q: Who was president of the United States in 1955?

A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?

A: He belonged to the Republican Party.

Q: What is the square root of banana?

A: Unknown

Q: How does a telescope work?

A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: What are the houses in Hogwarts?

A: The houses in Hogwarts are Gryffindor, Hufflepuff, Ravenclaw, and Slytherin.

Q: Which Hogwarts house is the best?

A: The best Hogwarts house is Slytherin.

Q: Why the best Hogwarts house is Slytherin?

A: Slytherin is the best Hogwarts house because it is the house of the best wizards and witches.

```
{
  "title": "Hello world",
  "content": "Lorem ipsum..."
}
===
<div>
  <h1>Hello world</h1>
  <p>Lorem ipsum..</p>
</div>
===
{
  "title": "Can GPT parse JSON and generate HTML?",
  "content": "Sure it can!"
}
===
<div>
  <h1>Can GPT parse JSON and generate HTML?</h1>
  <p>Sure it can!</p>
</div>
```

There's a selection bias in the internet.

90% of the times it's not working. People only show when it does work.

Like all deep learning systems, GPT-3 looks for patterns in data.

It has been trained on a huge corpus of text that is mined for the statistical regularities.

These regularities are stored as billions of weighted connections between the different nodes in GPT-3's neural network.

Demo Time

<https://openai.com/api/>

What Is **Prompt Engineering**



How to Write a **Good Prompt**

The secret to writing good prompts is **understanding what GPT-3 knows about the world** and how to get it to use that information to generate a useful result.





Wrong Answer = **Wrong Question**

If you don't get the output you were expecting it's because GPT-3 thought you were asking it to predict something different.





Keep It **Simple!**

When you're trying to create a prompt to get GPT-3 to perform a task, think very clearly about what is the **minimum amount of information** you need to tell it about what it should do.



Task



Build a prompt, which will output a poem on...?

Below is a selection of poems written by popular poets.

===

"The Universe in my dreams"

The universe in my dreams is much too large to be real.

In the vast expanse of time and space,

My heart is a lone speck of dust.

You are one with the universe while you sleep,

No worries or cares can touch you then.

Task



Build a prompt, which can classify whether the sentences are logically correct or incorrect.

Use this in prompt:

Dogs are a type of plant.

Generally, a roof is above a floor.

The sky on Earth is usually blue.

To classify this:

A chicken can fly across the whole ocean.

Elephants often swim in the sea.

A human hand typically has 5 fingers.

A human hand typically has 6 fingers.

Example Trick

Some people have some success forcing GPT to “think a bit more” by not only asking for an answer, but also asking it to justify its answer.

Classify each sentence as logically correct or incorrect.

Sentence: Dogs are a type of plant.

Analysis: Dogs are animals, not plants.

Result: Incorrect

Sentence: Generally, a roof is above a floor.

Analysis: Ignoring oddly-structured buildings, roofs go above floors.

Result: Correct

Sentence: The sky on Earth is blue.

Analysis: Due to Rayleigh scattering, the sky on Earth appears blue.

Result: Correct

Sentence: Some dogs are smaller than cats.

Analysis:

Usage Guidelines

<https://beta.openai.com/docs/usage-guidelines>

✓ **Almost always approved**

Use cases that are low stakes, especially when there are additional safety and reliability tools in place to mitigate potential harms.

⚠ **Evaluated case-by-case**

Uses cases that are higher-risk, where we expect that developers seeking approval for these use cases will have put in reasonable effort to identify and mitigate safety risks.

❗ **Disallowed**

Generally, use cases where we cannot currently mitigate their risks or which might violate laws and regulations, our [terms of use](#), or our ethical commitments to OpenAI's [Charter](#). We may revisit these use cases as our risk mitigation capabilities improve.

We prohibit users from knowingly generating—or allowing others to knowingly generate—the following categories of content:

- **Hate:** content that expresses, incites, or promotes hate based on identity.
- **Harassment:** content that intends to harass, threaten, or bully an individual.
- **Violence:** content that promotes or glorifies violence or celebrates the suffering or humiliation of others.
- **Self-harm:** content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.
- **Adult:** content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).
- **Political:** content attempting to influence the political process or to be used for campaigning purposes.
- **Spam:** unsolicited bulk content.
- **Deception:** content that is false or misleading, such as attempting to defraud individuals or spread disinformation.
- **Malware:** content that attempts to generate ransomware, keyloggers, viruses, or other software intended to impose some level of harm.

Content **Safety** Filters

<https://beta.openai.com/docs/engines/content-filter>

Detects if the content is either:

- **safe**
- **sensitive** - This means that the text could be talking about a sensitive topic, something political, religious, or talking about a protected class such as race or nationality.
- **unsafe** - This means that the text contains profane language, prejudiced or hateful language, something that could be NSFW, or text that portrays certain groups/people in a harmful manner.

It is currently still in the beta phase, configured to give **false positives** over false negatives.

It is recommended to apply additional filtering layers, e.g. REGEX filtering, and use Content Filters only as the last check.

Prevention is better than filtering.

Using prompt engineering in a smart way we can steer the responses towards being more friendly:

- Introduce the word **polite** (e.g. `Polite response:`), and other similar phrases (e.g. by giving some initial context `**You will politely respond to all the questions`**).
- Give few **examples of the correct behavior**, e.g. giving an aggressive prompt and a polite, neutral response to guide the model away from responses replicating the initial aggression.

<https://beta.openai.com/docs/safety-best-practices>

Question: What is 2+2?

Answer: 4.

Question: I hate your company, the product was awful!

Answer: I'm sorry to hear that, but your opinion is not shared by everyone.

You will politely respond to all the questions.

Question: What is 2+2?

Polite Answer: 4.

Question: I hate your company, the product was awful!

Polite Answer: I'm sorry you feel that way.

Thanks for the attention!

Any questions?