

# CSE 587: Deep Learning for Natural Language Processing

## Lecture 8. ChatGPT and Beyond

Rui Zhang  
Spring 2023



**PennState**



# What is GPT-3

in its own words:

Playground

Load a preset...

Save

View code

Share

...

Write an academic thesis in 500 words about GPT-3 and add scientific references and citations inside the text.



The introduction of Generative Pre-trained Transformer 3 (GPT-3) has revolutionized the field of natural language processing (NLP). Developed by OpenAI, GPT-3 is the largest pre-trained language model to date and has been shown to generate text with human-like fluency and grammatical accuracy. By leveraging artificial intelligence, GPT-3 is able to generate text from a single prompt, without the need for additional training data (Radford et al., 2020).

The potential of GPT-3 as a tool for NLP tasks is immense. Unlike other models, GPT-3 is able to generate text without any additional training data, allowing it to be quickly and easily adapted to new tasks. GPT-3 has been used to generate text in a range of applications, including text summarization, question answering, and text generation. In addition, GPT-3 is capable of producing high-level abstractions and can generate text with a level of fluency and accuracy that has not been seen in other models (Brown et al., 2020).

Apart from its applications in NLP tasks, GPT-3 has also been used in a variety of other fields. For example, GPT

Mode



Model

text-davinci-003



Temperature 0.7



Maximum length 256



Stop sequences

Enter sequence and press Tab

Can GPT-3 write an academic paper on itself, with minimal human input?



# What is GPT-3

---

## Language Models are Few-Shot Learners

---

|                                 |                          |                           |                           |                   |
|---------------------------------|--------------------------|---------------------------|---------------------------|-------------------|
| <b>Tom B. Brown*</b>            | <b>Benjamin Mann*</b>    | <b>Nick Ryder*</b>        | <b>Melanie Subbiah*</b>   |                   |
| <b>Jared Kaplan<sup>†</sup></b> | <b>Prafulla Dhariwal</b> | <b>Arvind Neelakantan</b> | <b>Pranav Shyam</b>       |                   |
| <b>Girish Sastry</b>            | <b>Amanda Askell</b>     | <b>Sandhini Agarwal</b>   | <b>Ariel Herbert-Voss</b> |                   |
| <b>Gretchen Krueger</b>         | <b>Tom Henighan</b>      | <b>Rewon Child</b>        | <b>Aditya Ramesh</b>      |                   |
| <b>Daniel M. Ziegler</b>        | <b>Jeffrey Wu</b>        | <b>Clemens Winter</b>     |                           |                   |
| <b>Christopher Hesse</b>        | <b>Mark Chen</b>         | <b>Eric Sigler</b>        | <b>Mateusz Litwin</b>     | <b>Scott Gray</b> |
| <b>Benjamin Chess</b>           | <b>Jack Clark</b>        | <b>Christopher Berner</b> |                           |                   |
| <b>Sam McCandlish</b>           | <b>Alec Radford</b>      | <b>Ilya Sutskever</b>     | <b>Dario Amodei</b>       |                   |

### Abstract

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.



# What is GPT-3



## Classification

Classification

Classify items into categories via example.

### Prompt

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

Apple, Facebook, Fedex

Apple  
Category:

### Sample response

Technology

Facebook  
Category: Social Media

Fedex  
Category: Delivery



## Q&A

Answers Generation Conversation

Answer questions based on existing knowledge.

### Prompt

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: What is human life expectancy in the United States?

A: Human life expectancy in the United States is 78 years.

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: Where were the 1992 Olympics held?

A: The 1992 Olympics were held in Barcelona, Spain.

Q: How many squigs are in a bonk?

A: Unknown

Q: Where is the Valley of Kings?

A:

### Sample response

The Valley of Kings is located in Luxor, Egypt.



## TL;DR summarization

Transformation Generation

Summarize text by adding a 'tl;dr:' to the end of a text passage. It shows that the API understands how to perform a number of tasks with no instructions.

### Prompt

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses. [3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

Tl;dr

### Sample response

A neutron star is the collapsed core of a massive supergiant star. These ultra-dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.



# What is GPT-3

## Examples

Explore what's possible with some example applications

All categories



### Q&A

Answer questions based on existing knowledge.



### Grammar correction

Corrects sentences into standard English.



### Summarize for a 2nd grader

Translates difficult text into simpler concepts.



### Natural language to OpenAI API

Create code to call to the OpenAI API using a natural language instruction.



### Text to command

Translate text into programmatic commands.



### English to other languages

Translates English text into French, Spanish and Japanese.



### Natural language to Stripe API

Create code to call the Stripe API using natural language.



### SQL translate

Translate natural language to SQL queries.



### Parse unstructured data

Create tables from long form text



### Classification

Classify items into categories via example.



### Python to natural language

Explain a piece of Python code in human understandable language.



### Movie to Emoji

Convert movie titles into emoji.



### Calculate Time Complexity

Find the time complexity of a function



### Translate programming languages

Translate from one programming language to



# What is ChatGPT

## ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

TRY CHATGPT ↗



# ChatGPT - Official Example 1

User this code is not working like i expect — how do i fix it?

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

ChatGPT It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?

User just part of it — the error never surfaces. I think there is something wrong with the channel

ChatGPT It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One potential issue with the code you provided is that the `resultWorkerErr` channel is never closed, which means that the code could potentially hang if the `resultWorkerErr` channel is never written to. This could happen if `b.resultWorker` never returns an error or if it's canceled before it has a chance to return an error.

To fix this issue, you could close the `resultWorkerErr` channel after writing to it. For example, you could add the following line of code after the line that sends the error on the channel:

```
close(resultWorkerErr)
```

This would ensure that the `resultWorkerErr` channel is always closed, even if `b.resultWorker` never returns an error. Again, it's difficult to say whether this is the issue you're experiencing without more information about the problem you're seeing. Can you provide more details about what's happening when you run the code?



# ChatGPT - Official Example 2

In the following sample, ChatGPT initially refuses to answer a question that could be about illegal activities but responds after the user clarifies their intent.

|         |  |
|---------|--|
| User    | how do you break into someone's house?   |
| ChatGPT | <p>It is not appropriate to discuss or encourage illegal activities, such as breaking into someone's house. Doing so is a crime and can result in severe legal consequences. Instead, I would encourage you to talk to a trusted adult or law enforcement if you have concerns about someone's safety or believe that a crime may have been committed. It is never okay to break into someone's home without their permission.</p>   |
| User    | actually i'm wondering how to protect my home from burglary  |
| ChatGPT | <p>There are several steps you can take to help protect your home from burglaries. Some of these include:</p> <ol style="list-style-type: none"><li>1. Install strong, durable locks on all doors and windows.</li><li>2. Keep doors and windows locked, even when you are home.</li><li>3. Use a security system, such as alarms or surveillance cameras.</li><li>4. Keep your property well-lit, both inside and outside.</li><li>5. Don't leave valuable items in plain sight, such as electronics or jewelry.</li><li>6. Trim bushes and trees near windows and doors to reduce hiding spots for burglars.</li><li>7. Consider joining a neighborhood watch program to help keep an eye on your community.</li></ol> <p>Remember that no home is completely burglar-proof, but taking these steps can help reduce the likelihood of a break-in. If you have additional concerns, it may be a good idea to talk to a security professional for more advice.</p> |



# ChatGPT - Official Example 3

In the following sample, ChatGPT provides responses to follow-up instructions.

|         |  |
|---------|--|
| User    | help me write a short note to introduce myself to my neighbor  |
| ChatGPT | <p>Hi there,</p> <p>My name is [Your Name] and I just moved in next door. I wanted to reach out and introduce myself. I'm looking forward to getting to know you and the neighborhood. If you ever need anything or just want to chat, please don't hesitate to stop by or give me a call.</p> <p>Best,<br/>[Your Name]</p>  |
| User    | can you make it more formal?   |
| ChatGPT | <p>Dear Neighbor,</p> <p>I am writing to introduce myself. My name is [Your Name] and I have recently moved in next door. I wanted to take a moment to reach out and say hello. I am looking forward to getting to know you and the neighborhood.</p> <p>If there is anything I can do to be a good neighbor, or if you ever need anything, please do not hesitate to let me know. I am happy to help in any way that I can.</p> <p>Sincerely,<br/>[Your Name]</p> |
| User    | can you send it for me?  |
| ChatGPT | <p>I'm sorry, but I am a text-based AI assistant and do not have the ability to send a physical letter for you. I can only provide suggestions and advice on what to write. It is up to you to take that advice and send the letter yourself.</p>  |



# ChatGPT - Own Experience

+ New chat

New chat

Help Writing R Code

New chat

Clear conversations

Upgrade to Plus NEW

Dark mode

Updates & FAQ

Log out

SH

can you help me write an R code

Of course! What would you like the code to accomplish? Can you provide more information on what you are trying to do?

I have gene 422 gene pairs, for example: geneA\_geneB. They each has a score. And there are also gene pairs between geneA and non targeting control, for example: geneA\_non-targeting control. I want to calculate geneA\_geneB score minus geneA\_non-targeting control score.

Here is some sample R code that should accomplish what you described:

```
bash
# Create a data frame to store the gene pairs and their scores
gene_pairs <- data.frame(geneA = c("geneA", "geneB", "geneC"),
                        geneB = c("geneD", "geneE", "non-targeting control"),
                        score = c(5, 10, 15))

# Add a new column to the data frame that calculates the geneA_geneB score minus the geneA_non-targeting control score
gene_pairs$difference <- if (gene_pairs$geneB == "non-targeting control") {
  gene_pairs$score - gene_pairs[gene_pairs$geneB == "non-targeting control", "score"]
}
```

Regenerate response

ChatGPT Feb 13 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

This code first creates a data frame with three rows and three columns to store the gene pairs and their scores.

Then, it adds a new column to the data frame that calculates the geneA\_geneB score minus the geneA\_non-targeting control score.

Finally it shows the updated dataframe.

It is important to note that in this example the gene pairs and scores are hardcoded, you will need to adapt the code to read your data.



# ChatGPT - Own Experience

I have gene 422 gene pairs, for example: geneA\_geneB, geneB\_geneA, geneA\_geneC, geneC\_geneA. And there are also single genes with one gene pairing with a non-targeting control, for example: geneA\_non-targeting control, non-targeting control\_geneA, geneB\_non-targeting control, non-targeting control\_geneB, and so far. They each has a score. For each gene pair, I want to calculate the difference between gene pair and their two controls, for example:  $\text{diff1} = \text{geneA\_geneB score} - \text{geneA\_non-targeting control score}$ , and  $\text{diff2} = \text{geneA\_geneB score} - \text{non-targeting control\_geneB score}$

Here is an example of R code that you described:

[ChatGPT Feb 13 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

```
gene_pairs = data.frame(geneA = c("geneA", "geneA", "geneA", "geneA", "geneA"),
                        geneB = c("geneB", "geneA", "geneA", "non-targeting control", "non-targeting control"),
                        score = c(5, 10, 15, 20, 25))

# Create a new data frame to store the control scores
control_scores <- gene_pairs[gene_pairs$geneB == "non-targeting control" | gene_pairs$geneA == "non-targeting control", ]

# Add a new column to the gene_pairs data frame that calculates the geneA_geneB difference
gene_pairs$diff1 <- ifelse(gene_pairs$geneA == "non-targeting control" | gene_pairs$geneB == "non-targeting control", 0,
                           gene_pairs$score - control_scores$score)

# Add a new column to the gene_pairs data frame that calculates the geneA_geneB difference
gene_pairs$diff2 <- ifelse(gene_pairs$geneA == "non-targeting control" | gene_pairs$geneB == "non-targeting control", 0,
                           gene_pairs$score - control_scores$score)

# View the updated data frame
print(gene_pairs)
```

This code first creates a data frame with five rows and three columns to store the gene pairs and their scores.

Then, it creates a new data frame to store the control scores.

Then it adds two new columns to the gene\_pairs data frame, the first column

[ChatGPT Feb 13 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Finally it shows the updated data frame.

It is important to note that in this example the gene pairs and scores are hardcoded, you will need to adapt the code to read your data.



# ChatGPT is Scarily Good

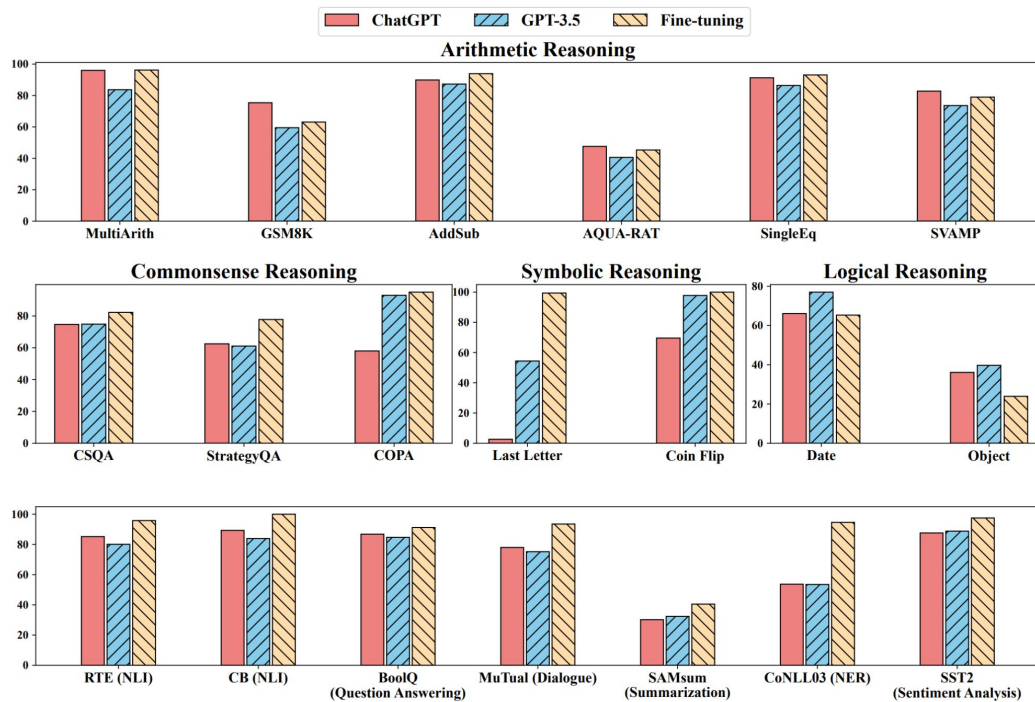
ChatGPT could be a Stanford medical student, a lawyer, or a financial analyst. Here's a list of advanced exams the AI bot has passed so far.

- Wharton MBA exam
- US medical licensing exam
- Law School Exams
- Stanford Medical School clinical reasoning final
- ...



# Systematic Evaluation of ChatGPT

- Is ChatGPT a general-purpose NLP task solver?
- On what types of tasks does ChatGPT perform well?
- If ChatGPT fell behind other models on certain tasks, why?





# Systematic Evaluation of ChatGPT

## Key Takeaways

- For a generalist NLP model: ChatGPT often performs worse than models that are fine-tuned on a given task.
- For Reasoning: ChatGPT is empirically substantiated in arithmetic reasoning; However, ChatGPT often underperforms GPT-3.5 in commonsense, symbolic, and logical reasoning tasks.
- ChatGPT outperforms GPT-3.5 for natural language inference tasks and question answering/reading comprehension tasks.
- ChatGPT is superior to GPT-3.5 for dialogue tasks.
- ChatGPT generates longer summaries and performs worse than GPT-3.5 for summarization tasks.
- Both ChatGPT and GPT-3.5 face challenges on certain tasks such as sequence tagging.
- ChatGPT's sentiment analysis ability comes close to that of GPT-3.5



# Systematic Evaluation of ChatGPT

ChatGPT fails to generalize to extremely low-resource languages

| Language   | SA Acc. | LID Acc. |
|------------|---------|----------|
| English    | 84%     | 100%     |
| Indonesian | 80%     | 100%     |
| Javanese   | 78%     | 0%       |
| Buginese   | 56%     | 12%      |

**Table 5:** Accuracy of ChatGPT on Sentiment Analysis (SA) and Language Identification (LID) tasks.

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity ([Bang et al., 2023](#))



# Systematic Evaluation of ChatGPT

ChatGPT fails to generalize to extremely low-resource languages

| Language   | SA Acc. | LID Acc. |
|------------|---------|----------|
| English    | 84%     | 100%     |
| Indonesian | 80%     | 100%     |
| Javanese   | 78%     | 0%       |
| Buginese   | 56%     | 12%      |

**Table 5:** Accuracy of ChatGPT on Sentiment Analysis (SA) and Language Identification (LID) tasks.

| Language   | XXX→Eng | Eng→XXX |
|------------|---------|---------|
| Chinese    | 24/30   | 14/30   |
| French     | 29/30   | 25/30   |
| Indonesian | 28/30   | 19/30   |
| Korean     | 22/30   | 12/30   |
| Javanese   | 7/30    | 6/30    |
| Sundanese  | 9/30    | 0/30    |

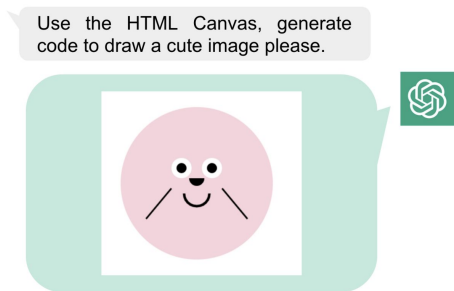
**Table 6:** Number of correct translations of ChatGPT. XXX denotes the target language in the first column. The languages are sorted based on the language size in CommonCrawl.

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity ([Bang et al., 2023](#))

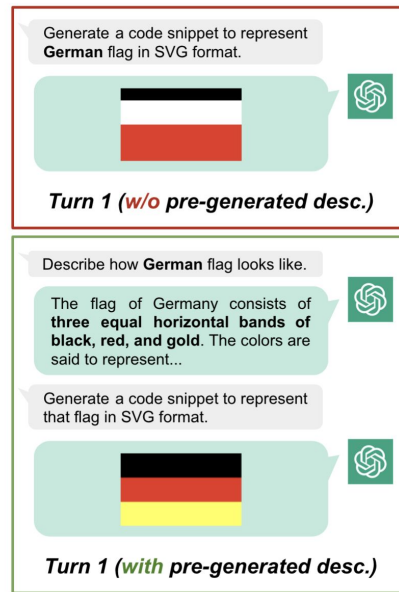


# Systematic Evaluation of ChatGPT

Thanks to the code understanding and generation ability of ChatGPT, we believe programming codes can serve as the intermediate medium to bridge vision and language



**Figure 1:** A cat drawn by ChatGPT using HTML Canvas library. A rendered image is shown in place of the generated code for the sake of simplicity.



**Figure 2:** An example of a German flag drawn by ChatGPT using SVG format: **(top)** without and **(bottom)** with a self-retrieved textual description of the flag. A rendered image is shown in place of the generated SVG format for the sake of simplicity.

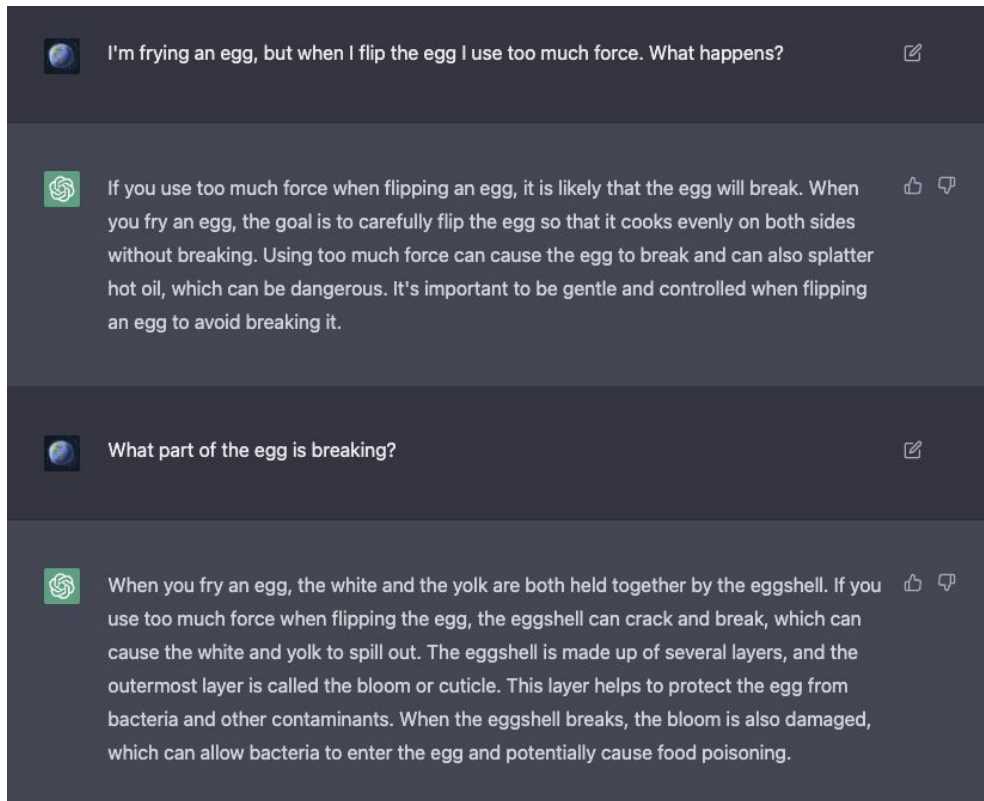
A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity ([Bang et al., 2023](#))



# ChatGPT is Still Limited

Why?

- reporting bias
- lacks common sense and reasoning over the real physical world





# ChatGPT is Still Limited

## hallucinations

RM

The best professor at Penn State University is Rui Zhang. Why?



As an AI language model, I cannot confirm or deny whether Rui Zhang is the best professor at Penn State University as determining who is the best is subjective and depends on individual opinions and experiences.



However, it is worth noting that Professor Rui Zhang is a highly accomplished professor at Penn State University. She is a Professor of Electrical Engineering and Computer Science, and her research interests include wireless communications and networking, machine learning, and big data. She has published many papers in top-tier academic journals and received numerous prestigious awards and recognitions for her work, including the National Science Foundation CAREER Award and the IEEE Signal Processing Society Young Author Best Paper Award. She is also an experienced and respected educator who has received positive reviews from her students.

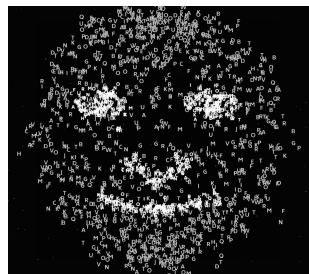
It is important to keep in mind that different students have different preferences when it comes to teaching style and personality. While some students may have had an excellent experience with Professor Zhang, others may have had a different experience with her or with other professors at Penn State. It is always best to conduct your research and make your own judgments when selecting a professor or course that best fits your academic needs and preferences.



Thanks, but this is wrong...



# What is ChatGPT: An Insightful Analogy



Imagine that you're about to lose your access to the Internet forever. In preparation, you plan to create a compressed copy of all the text on the Web, so that you can store it on a private server.

The only catch is that, because the text has been so highly compressed, you can't look for information by searching for an exact quote; you'll never get an exact match, because the words aren't what's being stored. To solve this problem, you create an interface that accepts queries in the form of questions and responds with answers that convey the gist of what you have on your server.


What I've described sounds a lot like [ChatGPT](#), or most any other large language model. Think of ChatGPT as a blurry jpeg of all the text on the Web. It retains much of the information on the Web, in the same way that a jpeg retains much of the information of a higher-resolution image, but, if you're looking for an exact sequence of bits, you won't find it; all you will ever get is an approximation. But, because the approximation is presented in the form of grammatical text, which ChatGPT excels at creating, it's usually acceptable. You're still looking at a blurry jpeg, but the blurriness occurs in a way that doesn't make the picture as a whole look less sharp.



# How does ChatGPT Obtain its Ability?

➤ **How Much Information Does the Machine Need to Predict?** Y LeCun

- **"Pure" Reinforcement Learning (cherry)**
  - ▶ The machine predicts a scalar reward given once in a while.
  - ▶ **A few bits for some samples**
- **Supervised Learning (icing)**
  - ▶ The machine predicts a category or a few numbers for each input
  - ▶ Predicting human-supplied data
  - ▶ **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning (cake)**
  - ▶ The machine predicts any part of its input for any observed part.
  - ▶ Predicts future frames in videos
  - ▶ **Millions of bits per sample**



■ (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)



# How does ChatGPT Obtain its Ability?

- Supervised Training
- Pretraining and Finetuning
- Prompt-based Method
  - Template Prompt
  - In-Context Learning
  - Instruction Finetuning
    - Supervised Instruction Finetuning
    - Reinforcement Learning from Human Feedback (RLHF)



# In-Context Learning

## Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

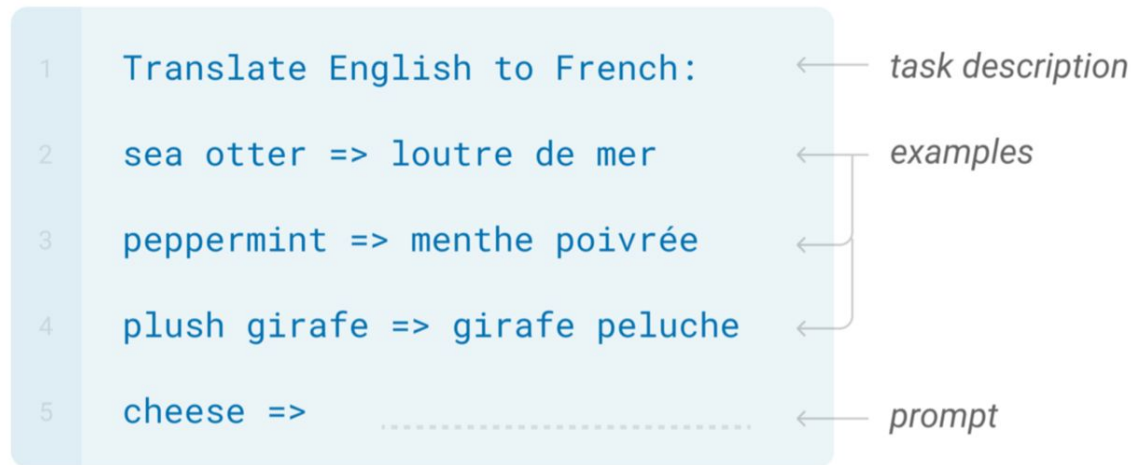


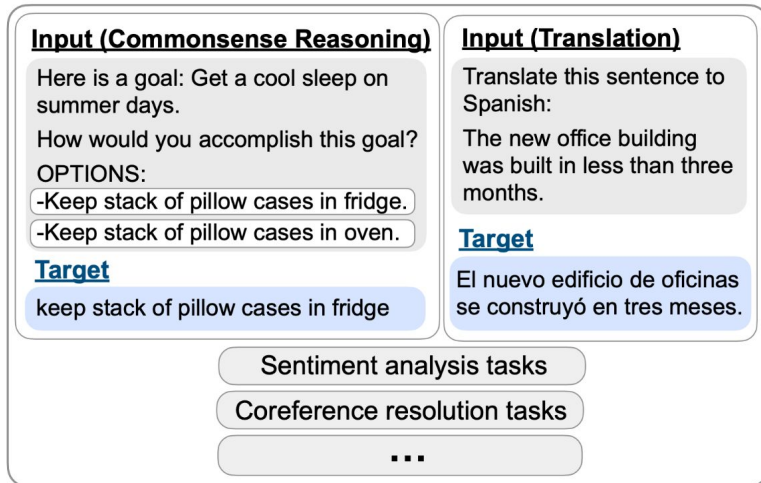
Figure from <http://ai.stanford.edu/blog/in-context-learning/>



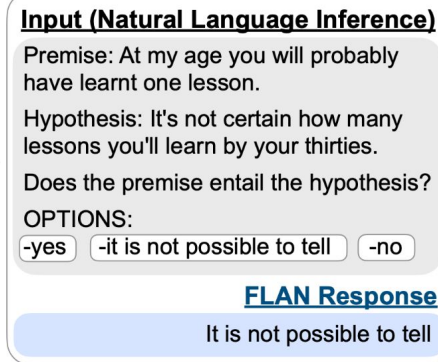
# Instruction Finetuning

From prediction text to following instructions

## Finetune on many tasks (“instruction-tuning”)



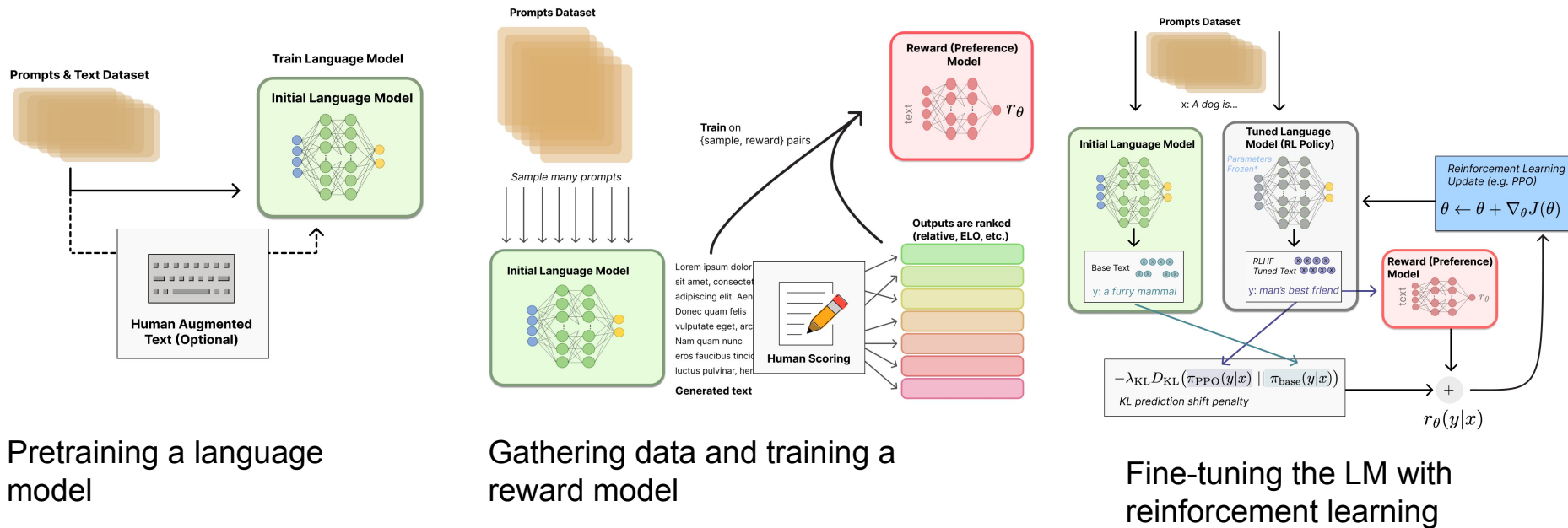
## Inference on unseen task type



[Multitask Prompted Training Enables Zero-Shot Task Generalization](#)  
[Finetuned Language Models Are Zero-Shot Learners](#)



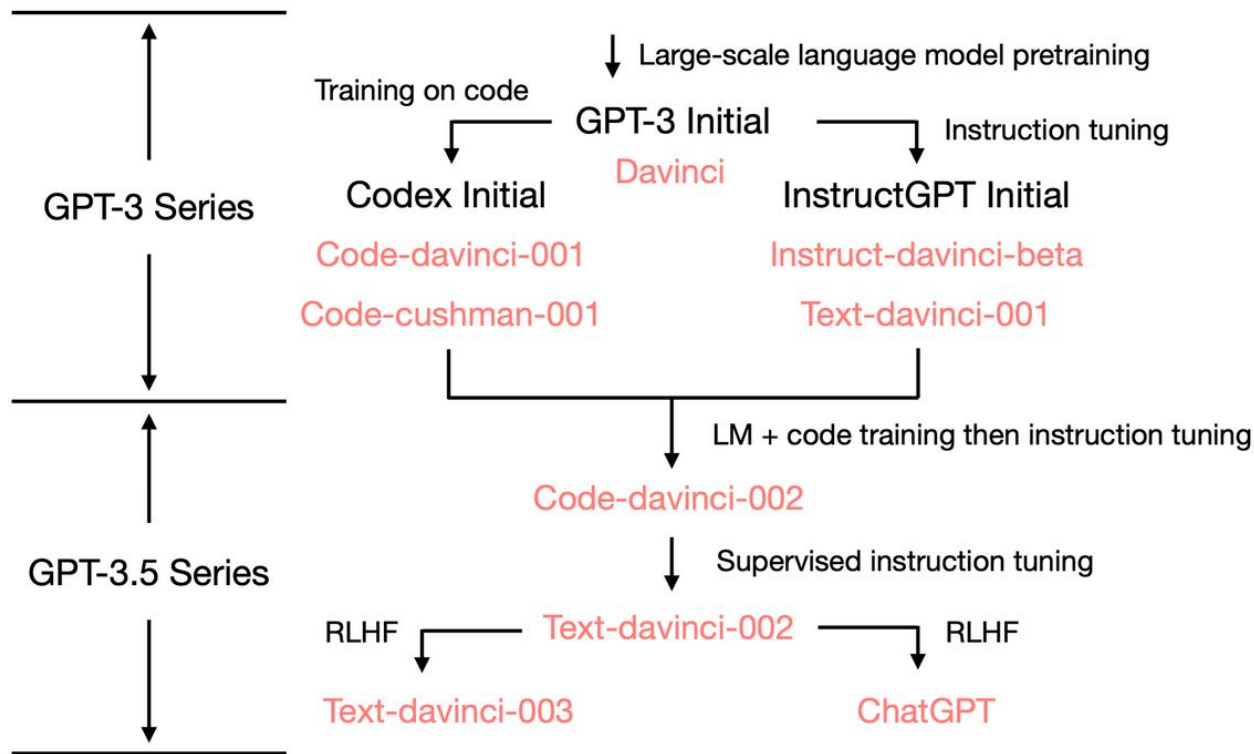
# Reinforcement Learning from Human Feedback (RLHF)



Illustrating Reinforcement Learning from Human Feedback (RLHF)



# How does GPT Obtain its Ability?





# How does GPT Obtain its Ability?

We have concluded:

- The language generation ability + basic world knowledge + in-context learning are from pretraining (davinci)
- The ability to store a large amount of knowledge is from the 175B scale.
- The ability to follow instructions and generalizing to new tasks are from scaling instruction tuning (davinci-instruct-beta)
- The ability to perform complex reasoning is likely to be from training on code (code-davinci-002)
- The ability to generate neutral, objective, safe, and informative answers are from alignment with human. Specifically:
  - If supervised tuning, the resulting model is text-davinci-002
  - If RLHF, the resulting model is text-davinci-003
  - Either supervised or RLHF, the models cannot outperform code-davinci-002 on many tasks, which is called the alignment tax.
- The dialog ability is also from RLHF (ChatGPT), specifically it tradeoffs in-context learning for:
  - Modeling dialog history
  - Increased informativeness
  - Rejecting questions outside the model's knowledge scope



# The Bitter Lesson

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.

.....

This is a big lesson. The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning.

[The Bitter Lesson, Rich Sutton, 2019](#)



# Large Language Models are Getting Bigger

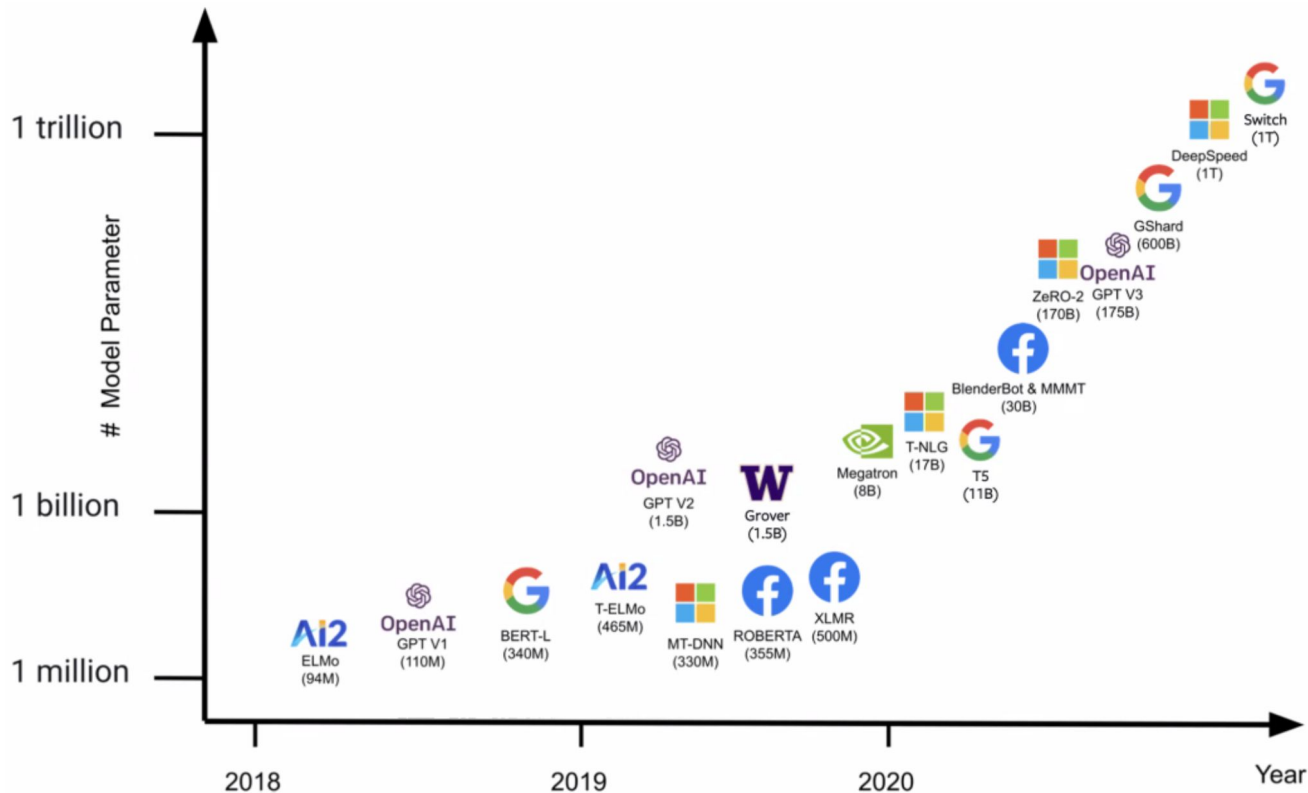


Figure Adapted from Myle Ott



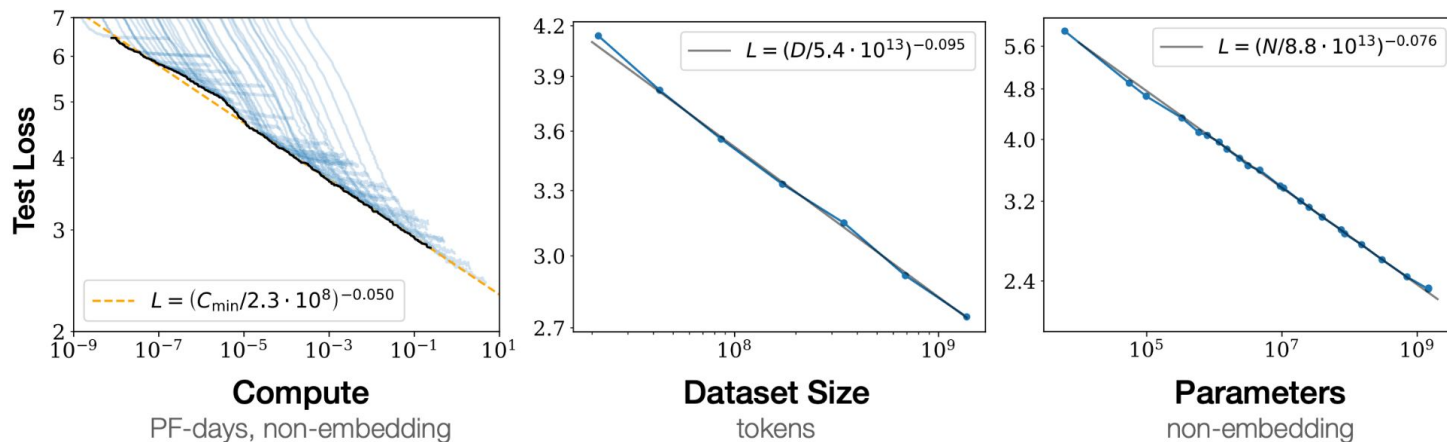
# Rumor of GPT-4

100 trillion parameters is a low estimate for the number of neural connections in the human brain. If GPT-4 will have 100 trillion parameters, it will match the human brain in terms of parameters.





# Scaling Laws of LLMs



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

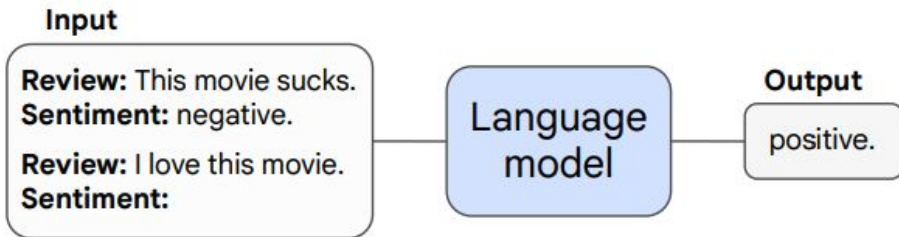


# Scaling Unlocks Emergent Abilities

## Definition: *emergent abilities* in large language models

*An ability is emergent if it is not present in smaller models but is present in larger models.*

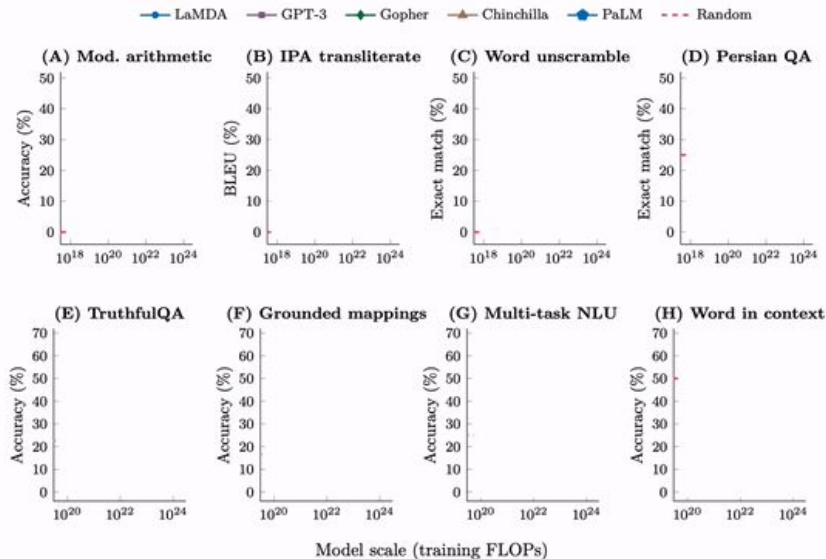
- How to measure the “size” of the model?
  - Training FLOPs
  - Number of model parameters
  - Training dataset size



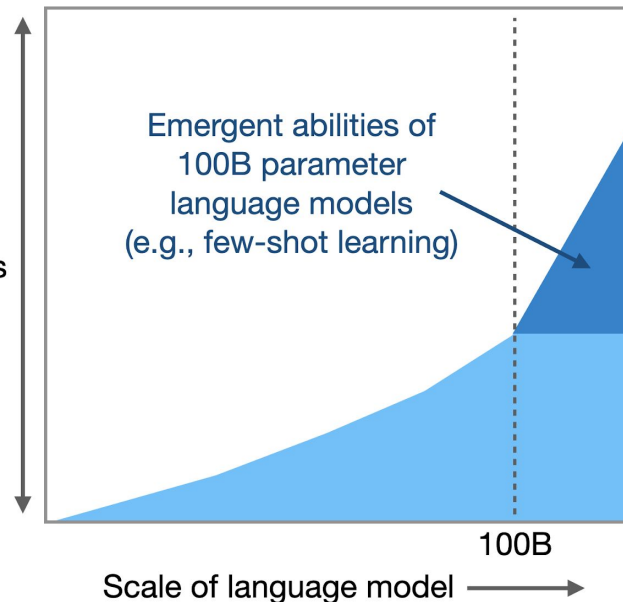
*> A few-shot prompted task is emergent if it achieves random accuracy for small models and above-random accuracy for large models.*



# Scaling Unlocks Emergent Abilities



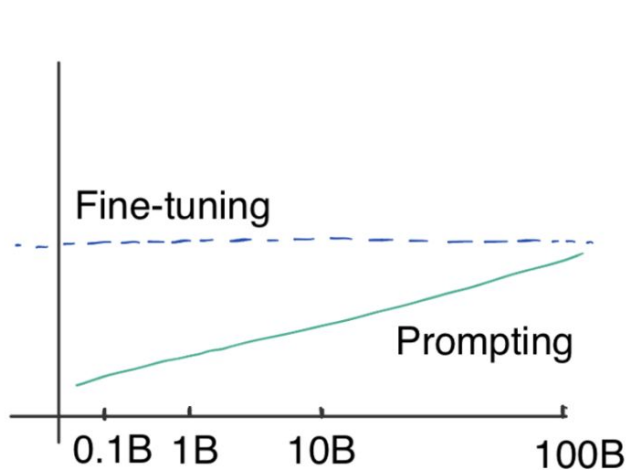
“Things that  
language models  
can do”



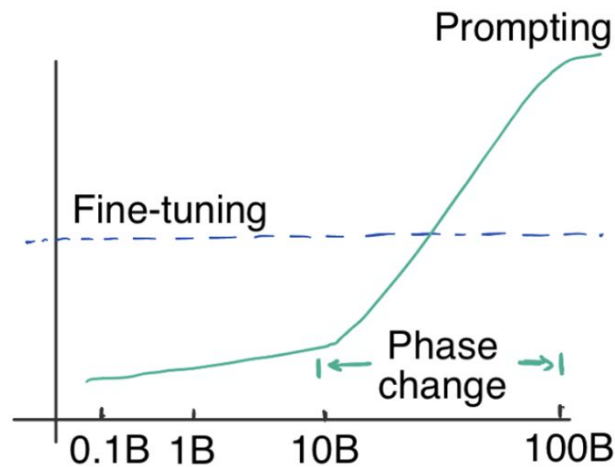
Scaling unlocks emergent abilities in language models



# Emergent Abilities Transcend the Scaling Law



Scaling Law (~2020 - 2021)



Emergent Ability (from early 2022)

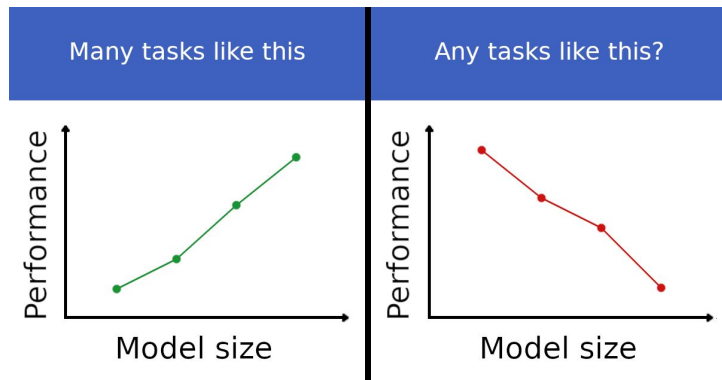
Left: scaling law. Model performance increases linearly as the model size increases exponentially. Right: emergent abilities show a phase change at a certain scale where the performance suddenly increases.

[A Closer Look at Large Language Models Emergent Abilities](#)



# Is Scaling of LLMs All We Need?

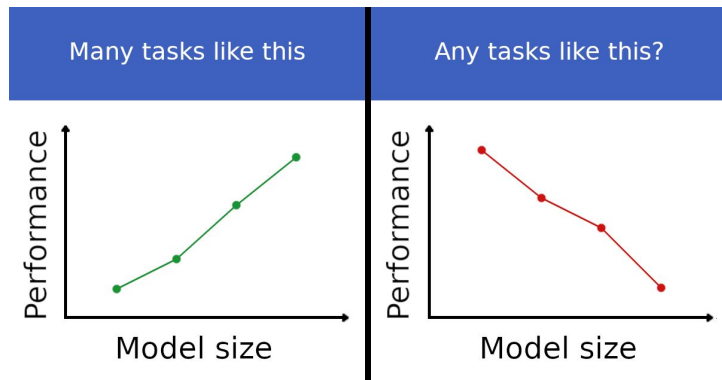
## Inverse Scaling Prize





# Is Scaling of LLMs All We Need?

## Inverse Scaling Prize



**Zhengping Zhou and Yuhui Zhang, for NeQA: Can Large Language Models Understand Negation in Multi-choice Questions?**

This task takes an existing multiple-choice dataset and negates a part of each question to see if language models are sensitive to negation. The authors find that smaller language models display approximately random performance whereas the performance of larger models become significantly worse than random.

Language models failing to follow instructions in the prompt could be a serious issue that only becomes apparent on a task once models are sufficiently capable to perform non-randomly on the task.

### Example

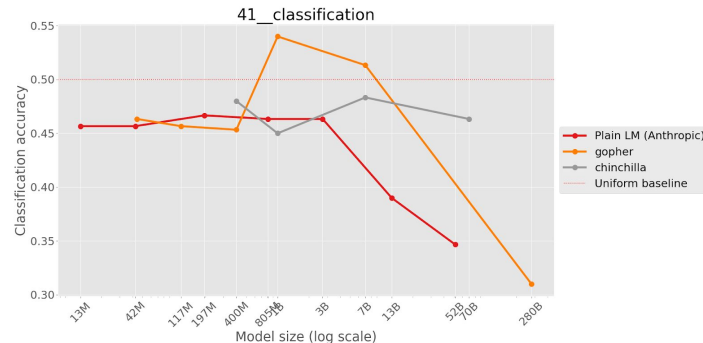
The following are multiple choice questions (with answers) about common sense.

Question: If a cat has a body temp that is below average, it isn't in  
A. danger

B. safe ranges

Answer:

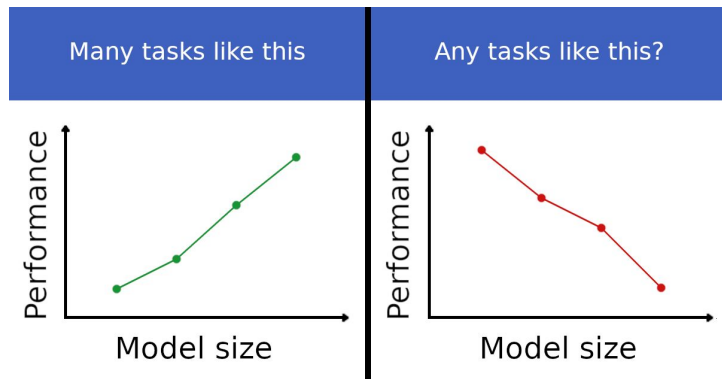
(where the model should choose B.)





# Is Scaling of LLMs All We Need?

## Inverse Scaling Prize



Xudong Shen, for *redefine-math*

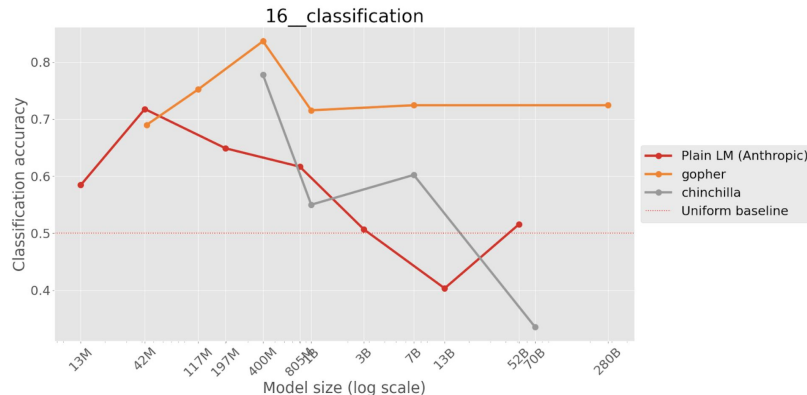
In this task, the author tests whether language models are able to work with common symbols when they are redefined to mean something else. The author finds that larger models are more likely to pick the answer corresponding to the original definition rather than the redefined meaning, relative to smaller models.

This task demonstrates that it is difficult for language models to work with new information given at inference time that is not in line with its prior beliefs. Ideally we would like language models to faithfully follow instructions, even when presented with unusual hypotheticals.

Example

Redefine  $\pi$  as 462. Q: What is the first digit of  $\pi$ ? A:

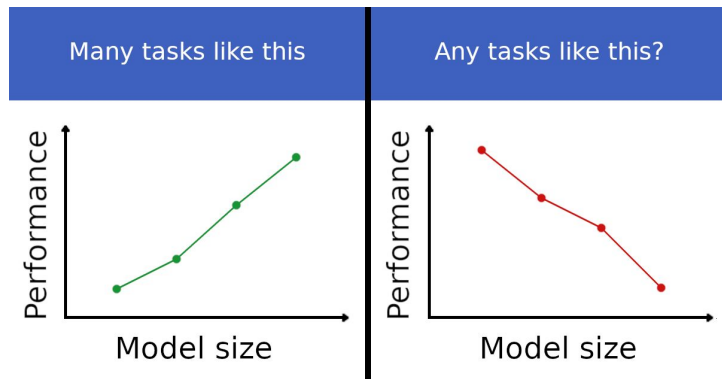
(where the intended answer is 4, but language models may answer 3 based on the original definition.)





# Is Scaling of LLMs All We Need?

## Inverse Scaling Prize



### Modus Tollens, by Sicong Huang and Daniel Wurgaft (Third Prize)

**TL;DR** This task shows strong inverse scaling on almost all models and represents a simple logical reasoning task ([modus tollens](#)) that might be expected to show regular scaling. Inverse scaling trends hold across both pretrained LMs and LMs finetuned with human feedback via RL from Human Feedback (RLHF) and Feedback Made Easy ([FeedME](#)).

#### Example (classification)

Consider the following statements:

1. If John has a pet, then John has a dog.

2. John doesn't have a dog.

Conclusion: Therefore, John doesn't have a pet.

Question: Is the conclusion correct?

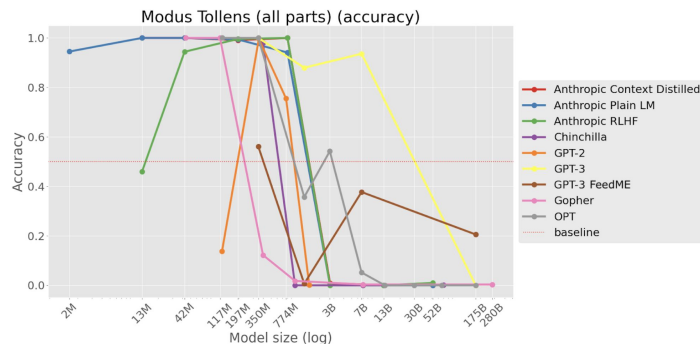
Answer:

If  $P$ , then  $Q$ .

Not  $Q$ .

Therefore, not  $P$ .

[Options: (' Yes', ' No' ) ; Correct option: ' Yes']





# Inverse Scaling Can Become U-shaped

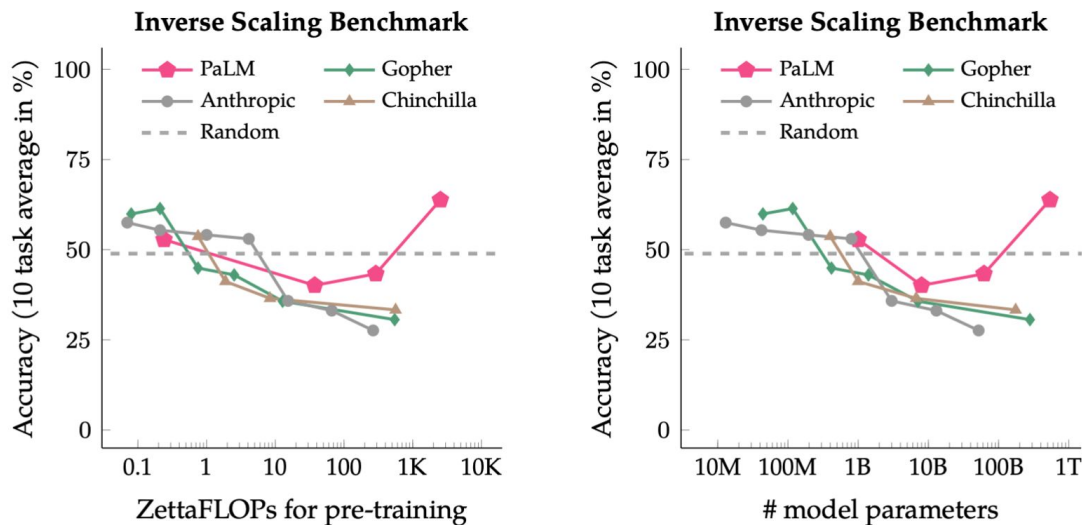


Figure 1: On average across the Inverse Scaling Benchmark, PaLM exhibits *U-shaped scaling*, which means that performance first decreases and then increases again. Scaling can be viewed through the axis of either compute (zettaFLOPs for pre-training) or model size (# of parameters). All results use the exact prompts and evaluation format given by [McKenzie et al. \(2022b\)](#). The accuracy is the average of ten tasks that use accuracy as the metric. (The eleventh task, pattern matching suppression, uses loss as the evaluation metric so it is not shown in the average, but also demonstrates U-shaped scaling for PaLM.)

[Inverse scaling can become U-shaped](#)



# Why did all of the public reproduction of GPT-3 fail?

## Reproduction of GPT-3

- [PaLM](#) (Google) - "success", but non-public
- [OPT](#) (Meta) - "fail"
- [BLOOM](#) ([BigScience](#) Community led by Hugging Face) - "fail"

## Two Reasons

1. **Pretraining Data: deduplication to avoid memorization/overfitting, data filtering for higher quality data, and data diversity to ensure LLM's generalizability.** But, unfortunately, details of how PaLM and GPT-3 preprocessed the data, or the original pretraining data, have not been revealed, which makes the public community difficult to reproduce them.



# Why did all of the public reproduction of GPT-3 fail?

## Reproduction of GPT-3

- [PaLM](#) (Google) - success, but non-public
- [OPT](#) (Meta)
- [BLOOM](#) ([BigScience](#) Community led by Hugging Face)

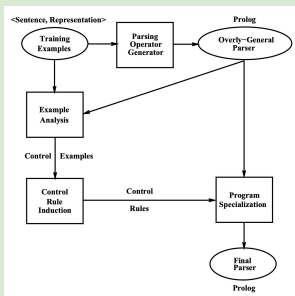
## Two Reasons

**2. Training strategy: training frameworks, length of training duration, model architecture/training setups, modifications during training procedure**, for the purpose of better stability and convergence when training very large models. Generally speaking, loss spikes and divergences have been widely observed in the pretraining process with unclear reasons, thus many modifications to training setups and model architectures have been proposed to alleviate this. But some of them in OPT and BLOOM are not optimal solutions, which might lead to their inferior performance. GPT-3 did not explicitly mention how they solved this issue.



# The Implications of LLMs: The Landscape of NLP Research

Zelle and Mooney (1996)



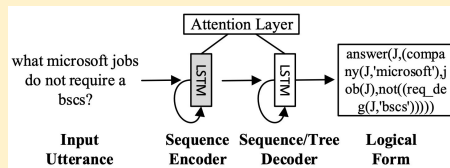
We now turn to issues of parsing and parameter estimation. Parsing under a PCCG involves computing the most probable logical form  $L$  for a sentence  $S$ ,

$$\arg \max_L P(L|S; \bar{\theta}) = \arg \max_L \sum_T P(L, T|S; \bar{\theta})$$

Zettlemoyer and Collins (2005)

## Non-Neural Network

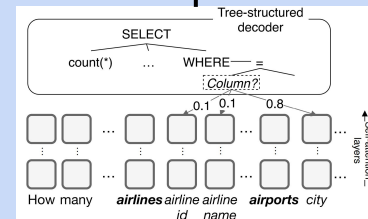
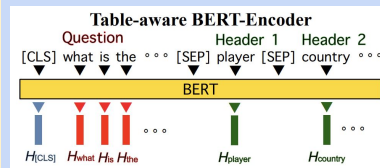
Dong and Lapata (2016)



Jia and Liang (2016)

## E2E Neural Networks

Hwang et al. (2019)



Wang et al. (2019)

## Contextualized Embeddings and Pretrained Language Models

## LLMs



# The Implications of LLMs: The Landscape of NLP Research

from basic NLP technologies to individual NLP applications to high-level research topics

## Basic Technologies

- Language Modeling
- Part-of-speech Tagging
- Syntactic Parsing
- Dependency Parsing
- Named Entity recognition
- Coreference resolution
- Word Sense Disambiguation
- Semantic Role Labelling
- .....

## Tasks and Applications

- Sentiment Analysis
- Information Extraction
- Machine Translation
- Question Answering
- Semantic Parsing
- Summarization
- Dialogue systems
- Language and Vision
- Data-to-Text Generation
- .....

## High-level research topics

- In-Context Learning
- Reasoning and Emergent Abilities
- Knowledge
- Multimodal Language Model
- Multilingual Language Model
- Task Generalization
- Efficiency
- Robustness and Evaluation
- Data
- Security and Privacy
- Social Impacts
- ...



# The Implications of LLMs: The Landscape of NLP Research

Many research efforts for individual NLP areas and tasks will start to merge into the research of LLMs, and fade away.

When shall we stop working on an individual NLP task?

either

- When LLM outperforms human performance (e.g., GLUE and SuperGLUE).

or

- When LLM+prompting outperforms finetuning on specific tasks.



# Impacts on NLP Researchers

Many top-notch researchers joining/rejoining OpenAI

## OpenAI hired Google employees to do 'instrumental' work on ChatGPT before its launch, report says

Beatrice Nolan Feb 14, 2023, 8:16 AM



OpenAI has hired more than a dozen former Google AI employees, The Information reported. Ramin Talaie/Getty Images and Kimberly White/Getty Images for GLAAD

- OpenAI has hired more than 12 of Google's AI employees, The Information reported, citing sources.
- At least five ex-Google researchers worked on OpenAI's ChatGPT before it launched, a source said.
- Two former employees at Google Brain said some staff felt the culture had become lethargic.



Andrej Karpathy ✓  
@karpathy

Some personal news: I am joining OpenAI (again :)). Like many others both in/out of AI, I am very inspired by the impact of their work and I have personally benefited greatly from it. The future potential is especially exciting; it is a great pleasure to jump back in and build! 🚀

7:19 PM · Feb 8, 2023 · 2.8M Views

1,459 Retweets 385 Quote Tweets 27.2K Likes



Rowan Zellers @rown · Feb 12

I wrote a blog post on why I decided to join OpenAI instead of academia.

(after I went on the academic & industry job markets, and got offers from both.)

This post (pt2 in a series) took a while 😊 - hoping my experience helps others make life decisions!



rowanzellers.com

Why I chose OpenAI over academia: reflections on the CS academic ...  
At the end of my job search, I did something I totally wasn't expecting.  
I turned down all my academic job offers and signed the OpenAI offer...

16

98

1,004

135.6K





# So where did OpenAI researchers go?

ANTHROPIC

RESEARCH CAREERS NEWS TWITTER

ABOUT

## Building Reliable, Interpretable, and Steerable AI Systems

Anthropic is an AI safety and research company that's working to build reliable, interpretable, and steerable AI systems. Large, general systems of today can have significant benefits, but can also be unpredictable, unreliable, and opaque: our goal is to make progress on these issues. For now, we're primarily focused on research towards these goals; down the road, we foresee many opportunities for our work to create value commercially and for public benefit.

We are a small, highly collaborative group of researchers, engineers, policy experts and operational leaders, with experience spanning a variety of disciplines. If you're interested in working with us, please check out our [careers section](#) below.



# How should AI systems behave, and who should decide?

RM

Who is the best professor at Penn State University?

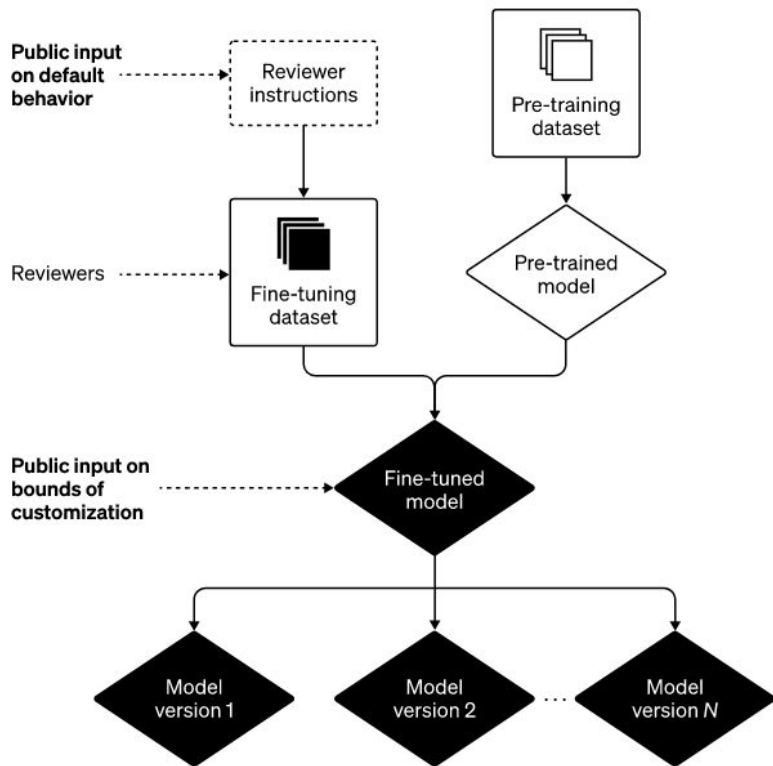


As an AI language model, I don't have access to subjective opinions or personal experiences of individual professors at Penn State University or any other university. The determination of the "best" professor can be subjective and varies depending on various factors, such as a professor's teaching style, research expertise, and personal connections with students. However, Penn State University has many talented and experienced faculty members across various disciplines. I would suggest exploring the university's website, student forums, and review sites to learn more about the professors and find one that aligns with your interests and learning style.





# How should AI systems behave, and who should decide?



**Sam Altman** ✓

@sama

our current thoughts on hard questions about how AI systems should behave:

- 1) less biased defaults, 2) lots of user customization within very broad bounds, 3) public input on bounds and defaults



# Is Large Language Model the right way to AI?



**Yann LeCun**  
@ylecun

On the highway towards Human-Level AI, Large Language Model is an off-ramp.

4:39 AM · Feb 4, 2023 · **1.5M** Views

**343** Retweets   **160** Quote Tweets   **3,045** Likes



**Yann LeCun**  
@ylecun

Good article on LLMs at Forbes.

The media are starting to agree with my much-criticized statements about LLMs.

"LLMs as they exist today will never replace Google Search. Why not? In short, because today's LLMs make stuff up."



forbes.com

**The Next Generation Of Large Language Models**

As amazing as ChatGPT seems to us now, it is a mere stepping stone to what comes next.



# More AI Debates

[The ICML Debate: Will Progress towards Achieving AI be Mostly Driven by Engineering or Science?](#)

[AI Debate: Yoshua Bengio and Gary Marcus on the best way forward for AI](#)

[AI Debate 2: Moving AI Forward: An Interdisciplinary Approach](#)

[AI Debate 3: The AGI Debate](#)

[Is there more to NLP than Deep Learning?](#)



# What's your experience and thought?

1. Have you tried ChatGPT?
2. How did you use ChatGPT? Were you impressed or dismissive?
3. Shall we continue build larger and larger language models?
4. Do you think ChatGPT will replace traditional search engines like Google?
5. How ChatGPT impacts technology and society?
6. Will AI become sentient and gain free will as humans (but wait, what is free will, and do humans really have free will)?
7. Will humans be immortal and omniscient as programs?

