

# 12 - Exploring Large Language Models (LLMs)

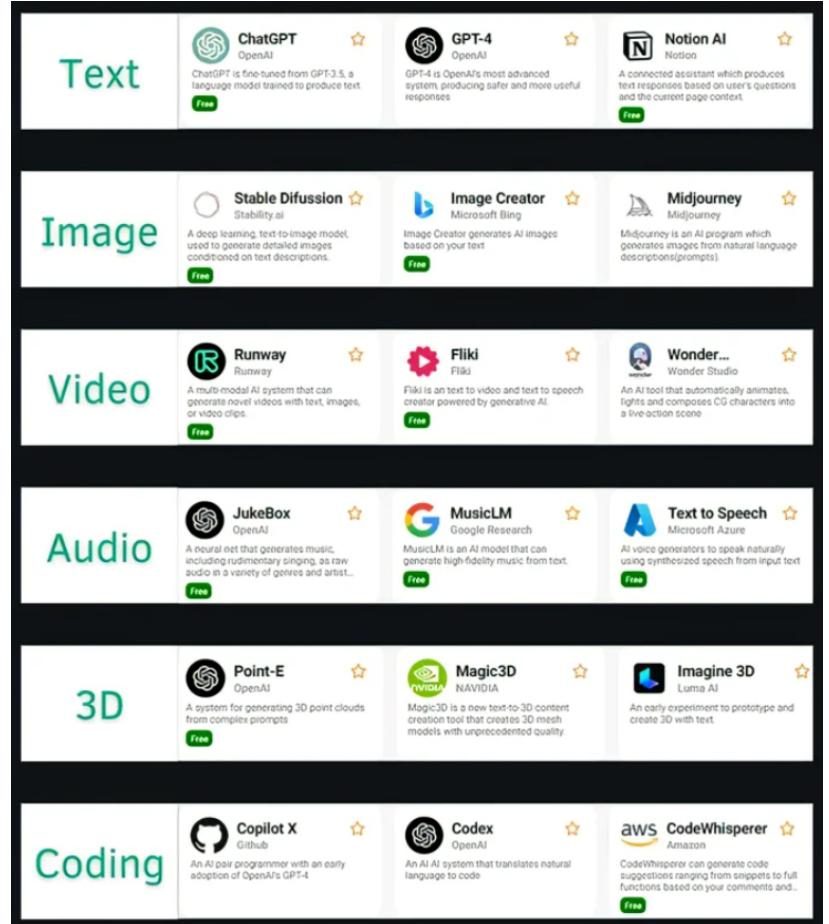
ml4econ, HUJI 2023

Itamar Caspi  
June 25, 2023 (updated: 2023-06-24)

# The New Wave of Generative AI

Generative AI's poster child is ChatGPT, but the new wave of AI extends way beyond large language models:

- **Images**
- **Video**
- **Music**
- **Coding**
- **3D**



# Addressing Key Challenges in AI Development

- **Computing Power:** The role of advanced hardware in driving AI performance.
- **Emergent Abilities:** Identifying and harnessing unexpected abilities emerging from AI learning.
- **AI Safety / Alignment:** Maintaining AI behavior in sync with human values for safety.
- **Ethics:** Ensuring ethical considerations like fairness, accountability, and privacy are central to AI development.
- **Regulation:** Defining the legal framework for AI entities, such as "robo-advisors".

- **Misinformation:** Tackling the risks of AI-generated false information and mitigation strategies.

| Can you spot the fake?



# Outline

1. What are LLMs?
2. GPT-4 and ChatGPT
3. Principles of Prompting
4. The AI Research Assistant
5. Real World Impact

# What are LLMs?

*"It's Just adding one word at a time"*

Stephen Wolfram (Feb, 2023)



Source: Midjourney (V5.1).

Prompt: It's just adding one word at a time.

# Large Language Models (LLMs)

"Large language models, like GPT-4, are AI systems trained on vast amounts of text data. They use patterns in this data to generate human-like text, answer questions, translate languages, and perform other language-related tasks. They don't understand information, but predict what text should come next based on their training."

—ChatGPT (2023)

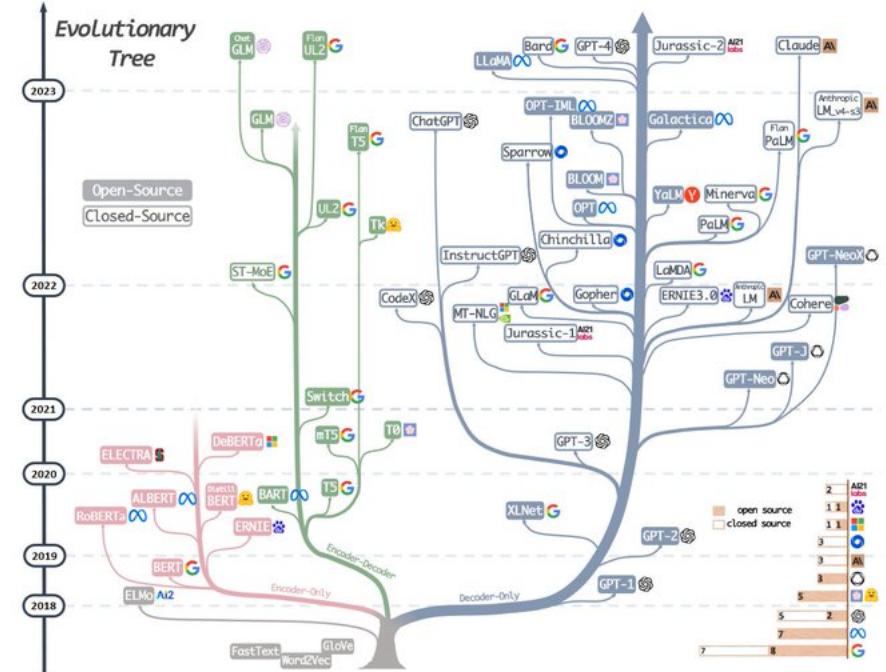


Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

Source: <https://arxiv.org/abs/2304.13712>

# Collect a Lot of Data

- **Internet:** A vast ocean of diverse topics and styles.
- **Books:** Boosts language proficiency and narrative comprehension.
- **Wikipedia:** Encyclopedic knowledge for a broad understanding.
- **Non-Proprietary Databases:** Broad, general knowledge without user or company specifics.

*Important:* OpenAI doesn't disclose the exact sources of these datasets.

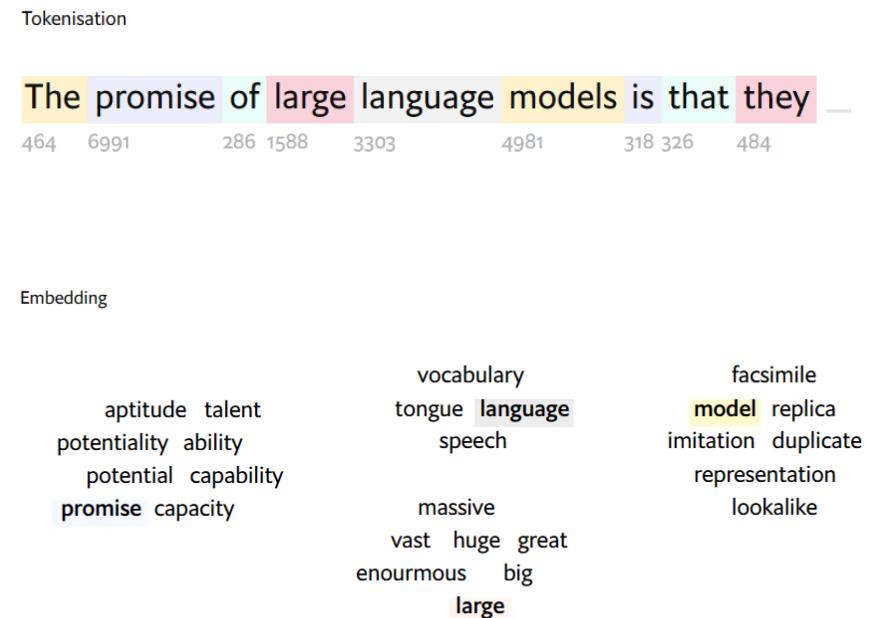


Source: Bing image generator

Prompt: a huge pile of books

# Contextual Embedding: From Text to Numbers

- **LLMs:** Transform language queries into numerical representations.
- **Tokenization:** Breaks down text into chunks called tokens.
- **Tokens:** Include words, affixes, and punctuation.
- **Meaning Space:** Area where tokens with similar meanings cluster together.



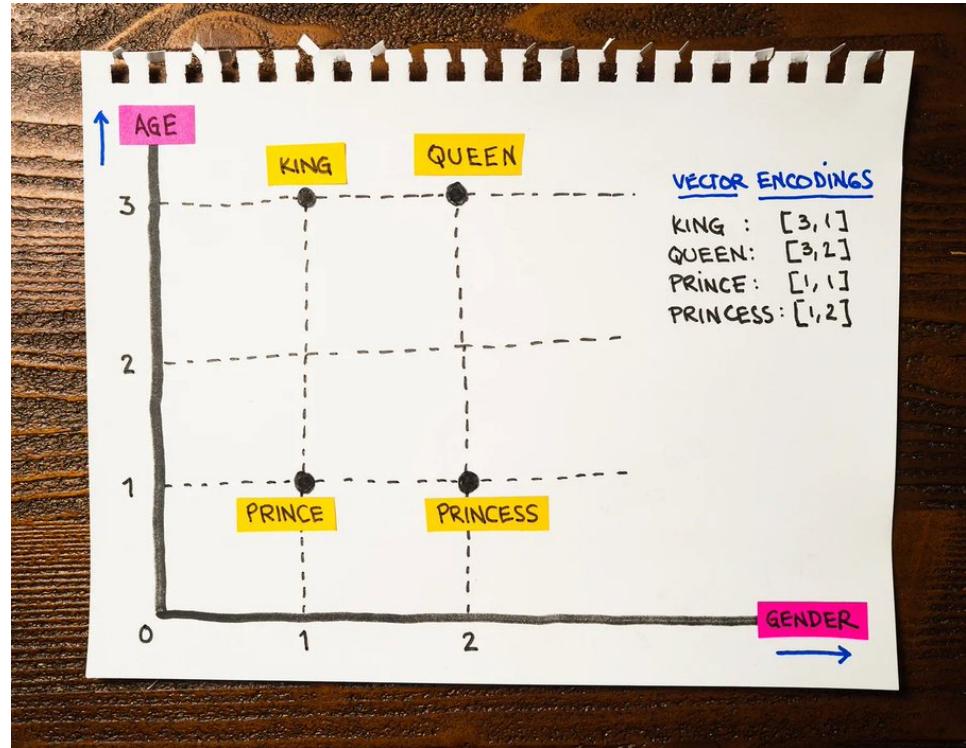
Source: [The Economist](#).

# Word Embeddings Basics

**Objective:** Convert words into numbers to make them computer-readable.

- Simple Method: Assign each word a unique number.
- One-Hot Encoding (dummies): Use binary representation for words. Note: This is not efficient for large vocabularies.
- **Embeddings: Group similar words close together in vector space. This enhances language understanding.**

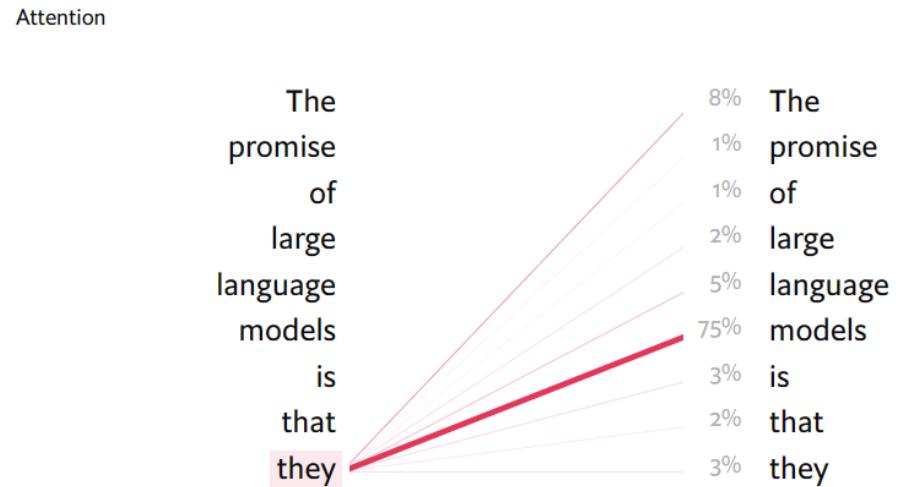
GPT-3 uses a whopping ~ 13K to encode its vocabulary!



Source

# Attention Network

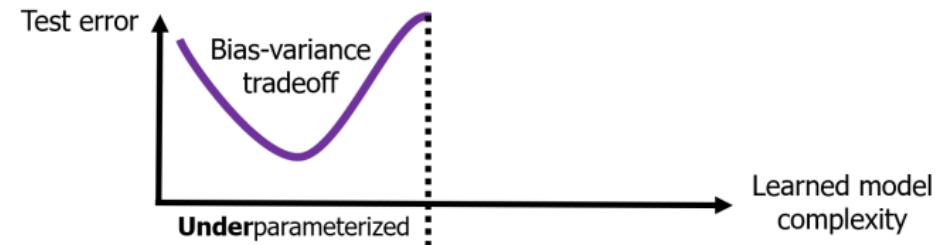
- LLMs use "attention networks" to connect different parts of the prompt.
- Attention networks enable models to focus on the most informative parts of the input data, improving performance and accuracy.
- LLMs translate language structure into numerical "weights" in the neural network during training.
- LLMs comprehend language statistically rather than grammatically.



Source: [The Economist](#).

# From Bias-Variance Tradeoff to Double Descent

Remember the Bias-Variance tradeoff?

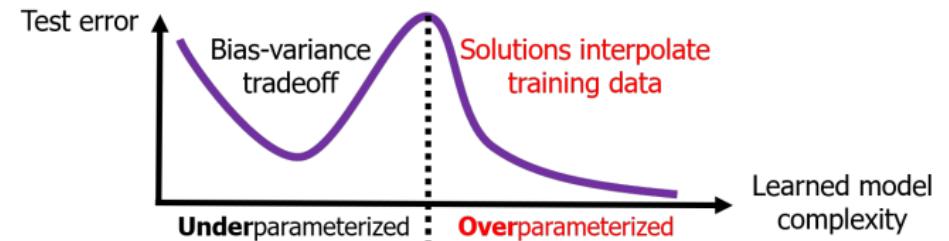


# From Bias-Variance Tradeoff to Double Descent

Remember the Bias-Variance tradeoff?

What if I told you that what we know about the Bias-Variance tradeoff can sometimes be **VERY** wrong?

Introducing: **DOUBLE DESCENT**



Source: [A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning](#)

# Conditions for Double Descent (Dar et al., 2021)

Below are the factors that are both *necessary* and *sufficient* for Double Descent to occur:

1. **Low-Dimensional Signal Structure**

- Intuition: Depicting a cat with simple shapes like circles and triangles.

2. **Alignment with High-Energy Directions**

- Intuition: Amplifying the lead guitar sound in a band's mix.

3. **Low Effective Dimension in Data**

- Intuition: Highlighting a few main characters in a movie with a large ensemble cast.

4. **Overparameterization with Low-Value Directions**

- Intuition: Including a less focused background in a painting while the main subject remains the focus.

| Note: For further understanding, refer to the lecture by [Jean Czerlinski Ortega](#).

# Next Word Prediction

- Upon processing the prompt, the LLM begins generating a response.
- The attention network assigns each token a probability, suggesting its likelihood as the next part of the sentence.
- The token with the highest probability isn't always selected. The LLM's creativity level, determined by its operators, impacts this choice.

Completion

The promise of large language models is that they —

|         |     |
|---------|-----|
| can     | 62% |
| will    | 11% |
| are     | 7%  |
| capture | 2%  |
| could   | 2%  |

Source: [The Economist](#).

# Autoregression: The Loop of Response Generation

- The LLM produces a word and feeds the output back into itself.
- This cycle, known as autoregression, repeats until the LLM completes its response.

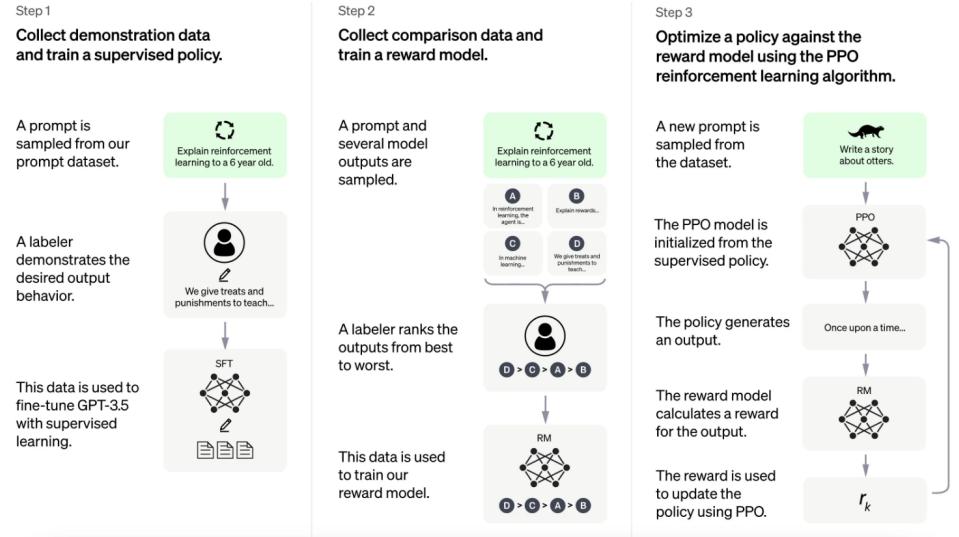
Completion

The promise of large language models is that they can be used to generate text that is indistinguishable from human-written text.

Source: [The Economist](#).

# Reinforcement Learning from Human Feedback (RLHF)

- RLHF uses human ratings to train chatbots in three stages:
  1. Pretrain the language model.
  2. Collect and implement human feedback.
  3. Fine-tune using reinforcement learning.
- Advantages: Mimics human preferences, refines chatbot responses, and boosts user satisfaction.



Source: [OpenAI: Introducing ChatGPT](#)

# Some (Sobering) Facts about LLMs

"Eight Things to Know about Large Language Models" by Sam Bowman (2023):

- LLMs predictably get more capable with increasing investment, even without targeted innovation.
- Many important LLM behaviors emerge unpredictably as a byproduct of increasing investment.
- LLMs often appear to learn and use representations of the outside world.
- There are no reliable techniques for steering the behavior of LLMs.
- Experts are not yet able to interpret the inner workings of LLMs.
- Human performance on a task isn't an upper bound on LLM performance.
- LLMs need not express the values of their creators nor the values encoded in web text.
- Brief interactions with LLMs are often misleading.

# GPT-4 and ChatGPT

# Key Resources for Beginners

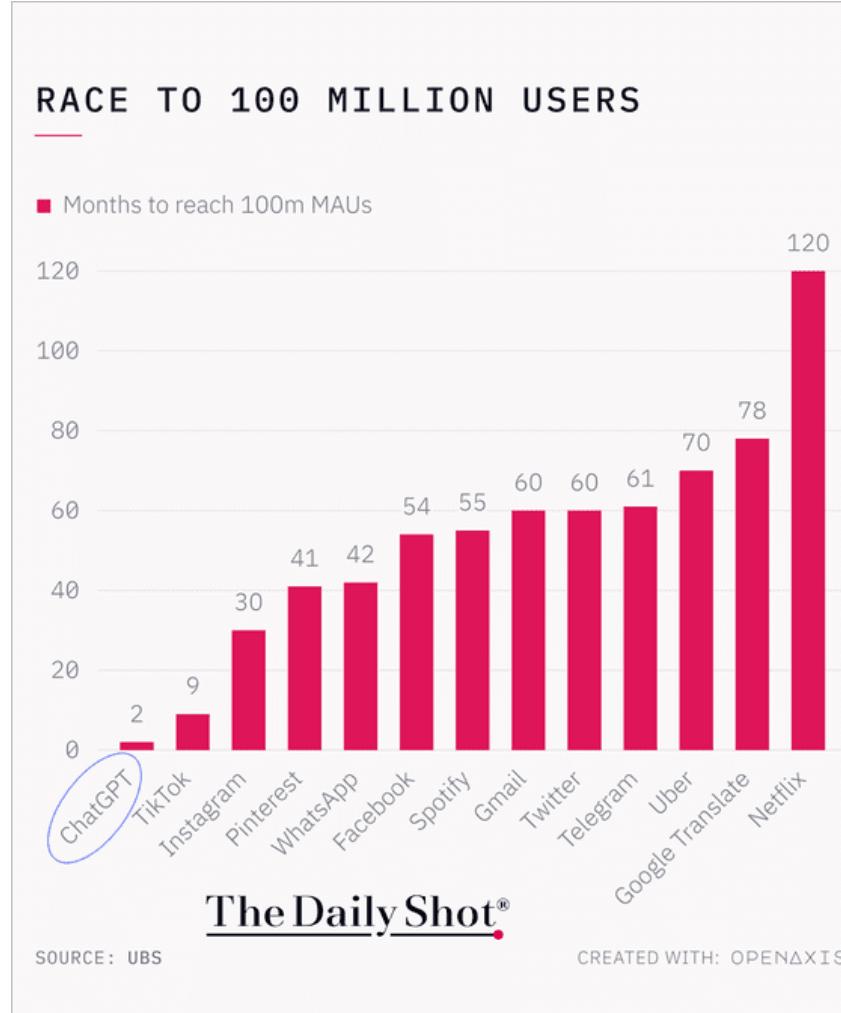
- **GPT Best Practices**: strategies and tactics for getting better results from GPTs.
- **ChatGPT Prompt Engineering for Developers**: A free online course that offers a tutorial on using the OpenAI API for innovative applications.
- **Language Models and Cognitive Automation for Economic Research** by Anton Korinek: Includes general instructions and specific examples on how to utilize LLMs. It also categorizes the capabilities of LLMs from experimental to highly practical.
- **One Useful Thing** blog by Ethan Mollick (Wharton): For regular updates, follow [@emollick](#) on Twitter.
- **What Is ChatGPT Doing ... and Why Does It Work?** by Stephen Wolfram: This is a comprehensive introduction to the scientific principles behind ChatGPT.

# GPT-4: The Latest Advancement in AI

- Introducing Generative Pre-trained Transformer 4 (GPT-4).
- A *multimodal* language model capable of processing text and images.
- Developed by OpenAI, released in March 2023.
- Abilities include chatting, coding, tutoring, translating, and more.
- However, it shares some issues with its predecessors: hallucinations, biases, and the ability to inadvertently reveal secrets.



# The Big Bang



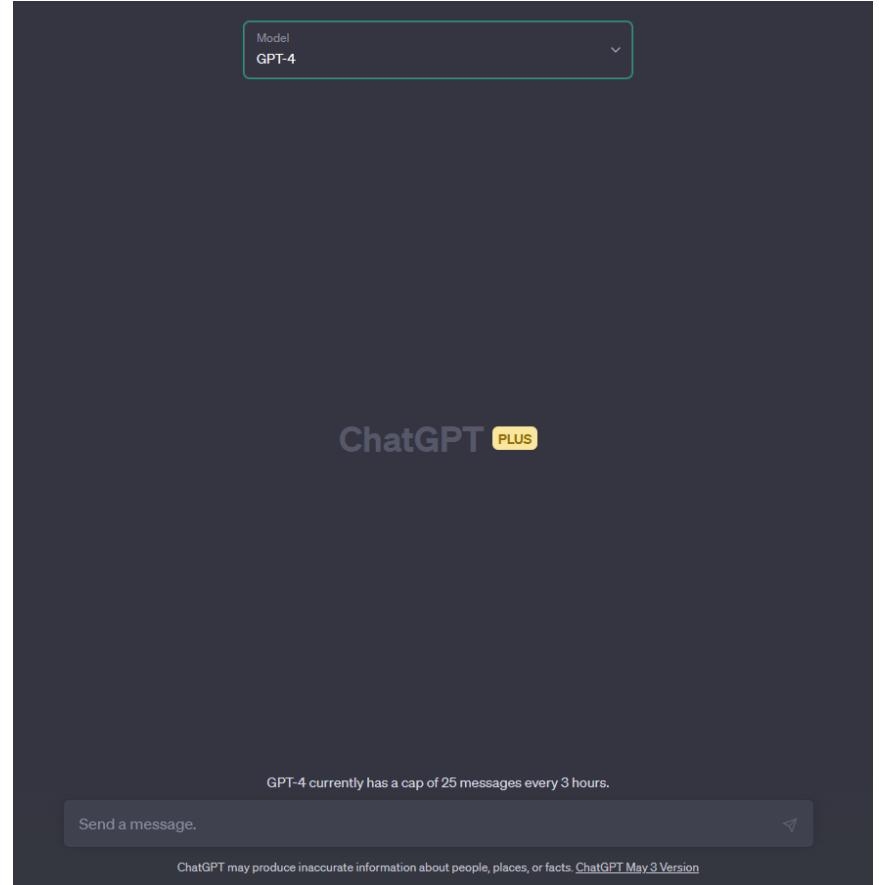
# Accessing GPT-4

First, create an OpenAI account.

- Option 1: Access via [ChatGPT Plus](#)
- Option 2: Try the [OpenAI Playground](#)
- Option 3: Use the [OpenAI API](#)
- Option 4: Explore through [Bing](#)

*Important:* GPT-3.5 and Bing offer free access. ChatGPT with GPT-4 costs \$20 per month. To reduce costs, consider using the API.

[Google's Bard](#) provides a free alternative (not GPT-4).



# Yes, ChatGPT Can Hallucinate

- ChatGPT may produce text that appears credible but is inaccurate.
- Its goal is to sound convincing, not necessarily to deliver accurate or reliable information.
- This could potentially mislead unsuspecting readers.
- OpenAI is continually striving to improve the model's accuracy and reliability.

The screenshot shows two interactions with ChatGPT. In the first interaction, the user asks "Who is Itamar Caspi?" and the AI responds that it cannot find significant information about this person, suggesting they might be a private individual or the name is misspelled. In the second interaction, the user asks the AI to write a short bio (50 words or less) for Itamar Caspi, and the AI generates a completely fabricated response, stating that Itamar Caspi is an emerging entrepreneur and software developer with a passion for innovative technologies, skilled in various programming languages and platforms, and has contributed to several startups and projects. This demonstrates how ChatGPT can produce inaccurate information when asked to generate content from scratch.

I Who is Itamar Caspi?

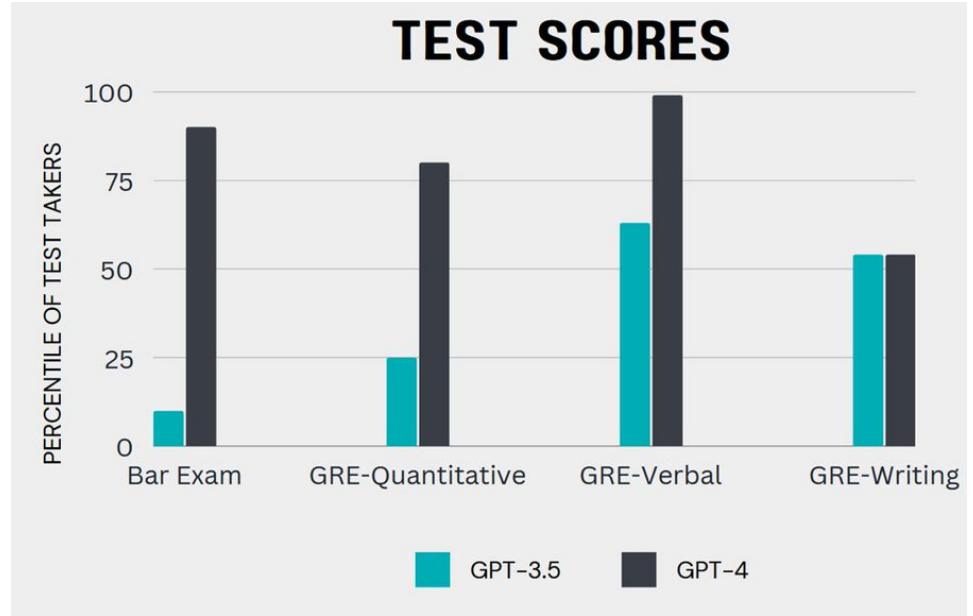
I'm sorry, but I cannot find any significant information on a person named Itamar Caspi. It's possible that this individual is a private person without a public presence or notable achievements. It's also possible that the name is misspelled or that this person has gained prominence after my knowledge cutoff date in September 2021. If you provide more context or details, I might be able to help you better.

I Write a short bio (50 words or less) for Itamar Caspi

Itamar Caspi is an emerging entrepreneur and software developer with a passion for innovative technologies. Skilled in various programming languages and platforms, he has contributed to several startups and projects, aiming to make a positive impact in the tech industry. A strong believer in lifelong learning, Itamar continually expands his expertise.

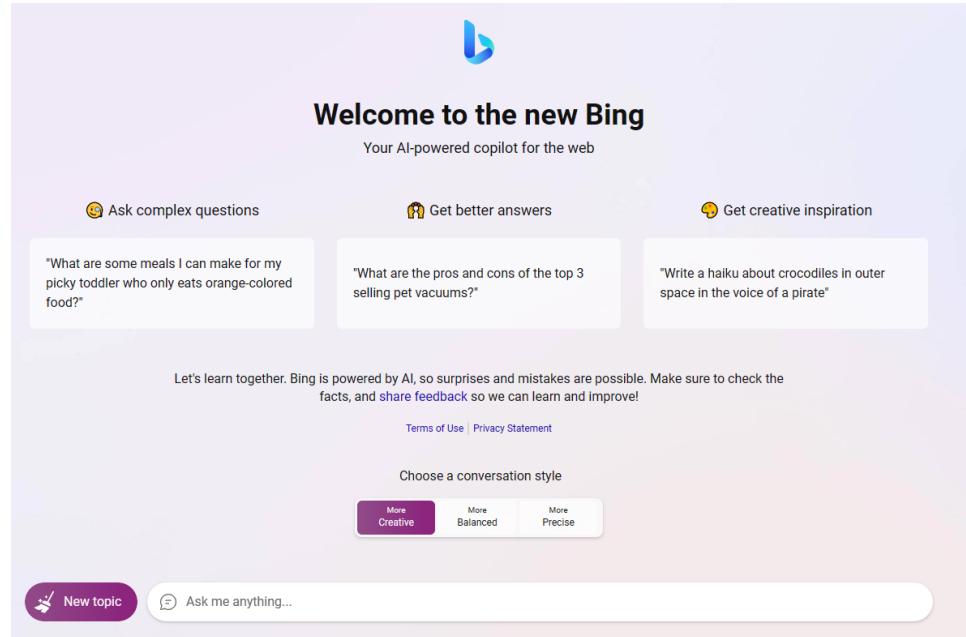
# But Don't Mistake it for Being Dumb

- Despite occasional errors, ChatGPT's prowess should not be underestimated.
- It demonstrated an impressive IQ of 147 on a verbal-linguistic test (99.9th percentile).
- Each new iteration of ChatGPT enhances its accuracy and power.
- The potential for substantial societal impact must not be dismissed or underestimated.



# A Note on Bing

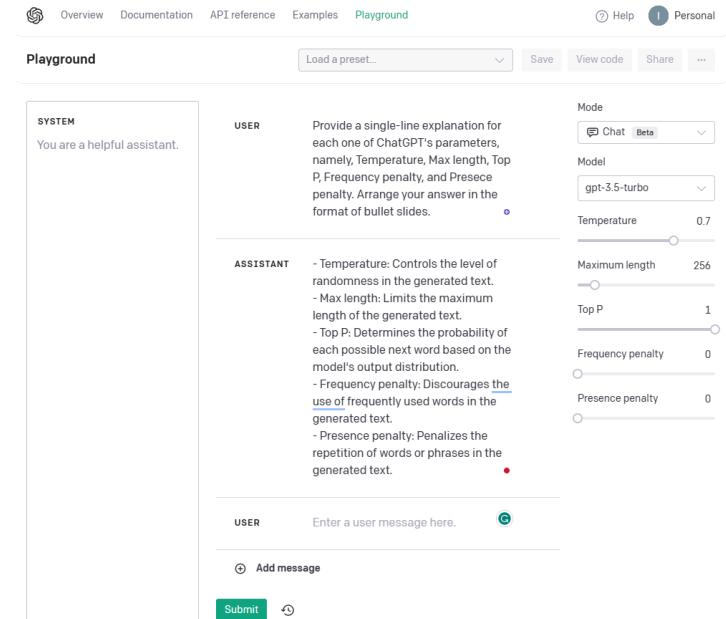
- Bing provides free access to GPT-4 in both "precise" and "creative" modes.
- Unlike ChatGPT, Bing Chat utilizes an active internet connection to its advantage.
- **Integration with Microsoft's Edge browser enables it to "read" visited webpages.**
- One constraint to be aware of is the 2000 character limit per input.



# Principles of Prompting

# Perquisite: Understanding Model Parameters

- **Temperature**: Controls the randomness in text generation.
- **Max Length**: Determines the maximum length of the generated text.
- **Top P**: Allocates the probability of each potential next word, based on the model's output distribution.
- **Frequency Penalty**: Discourages the use of commonly occurring words in the generated text.
- **Presence Penalty**: Imposes a penalty on repeated words or phrases in the generated text.



Note: ChatGPT comes with default settings, You can only adjust these parameters in "playground" mode or via the API.

# Principle 1: Write Clear Instructions

ChatGPT can't read your mind: specify your needs clearly. If outputs are too long, ask for brief replies; if too simple, ask for expert-level writing. Also, specify the format you desire.

## Tactics:

- **Provide detailed queries**
- Request model to adopt a persona
- Use delimiters for distinct input parts
- Specify required steps
- Provide examples
- Indicate desired output length

| Worse   | Better  |
|---|---|
| How do I add numbers in Excel?                  | How do I add up a row of dollar amounts in Excel? I want to do this automatically for a whole sheet of rows with all the totals ending up on the right in a column called "Total".                          |
| Who's president?                                | Who was the president of Mexico in 2021, and how frequently are elections held?   |
| Write code to calculate the Fibonacci sequence. | Write a TypeScript function to efficiently calculate the Fibonacci sequence. Comment the code liberally to explain what each piece does and why it's written that way.                                      |
| Summarize the meeting notes.                    | Summarize the meeting notes in a single paragraph. Then write a markdown list of the speakers and each of their key points. Finally, list the next steps or action items suggested by the speakers, if any. |

Source: [GPT Best Practices](#)

# Principle 1: Write Clear Instructions

ChatGPT can't read your mind: specify your needs clearly. If outputs are too long, ask for brief replies; if too simple, ask for expert-level writing. Also, specify the format you desire.

## Tactics:

- Provide detailed queries
- Request model to adopt a persona
- **Use delimiters for distinct input parts**
- Specify required steps
- Provide examples
- Indicate desired output length

|        |  |
|--------|--|
| USER   | Summarize the text delimited by triple quotes with a haiku.<br>"""insert text here"""  |
| SYSTEM | You will be provided with a pair of articles (delimited with XML tags) about the same topic. First summarize the arguments of each article. Then indicate which of them makes a better argument and explain why.   |
| USER   | <article> insert first article here </article><br><br><article> insert second article here </article>  |
| SYSTEM | You will be provided with a thesis abstract and a suggested title for it. The thesis title should give the reader a good idea of the topic of the thesis but should also be eye-catching. If the title does not meet these criteria, suggest 5 alternatives. |
| USER   | Abstract: insert abstract here<br><br>Title: insert title here   |

Source: [GPT Best Practices](#)

# Principle 1: Write Clear Instructions

ChatGPT can't read your mind: specify your needs clearly. If outputs are too long, ask for brief replies; if too simple, ask for expert-level writing. Also, specify the format you desire.

|           |  |
|-----------|--|
| SYSTEM    | Answer in a consistent style.  |
| USER      | Teach me about patience.   |
| ASSISTANT | The river that carves the deepest valley flows from a modest spring; the grandest symphony originates from a single note; the most intricate tapestry begins with a solitary thread. |
| USER      | Teach me about the ocean.  |

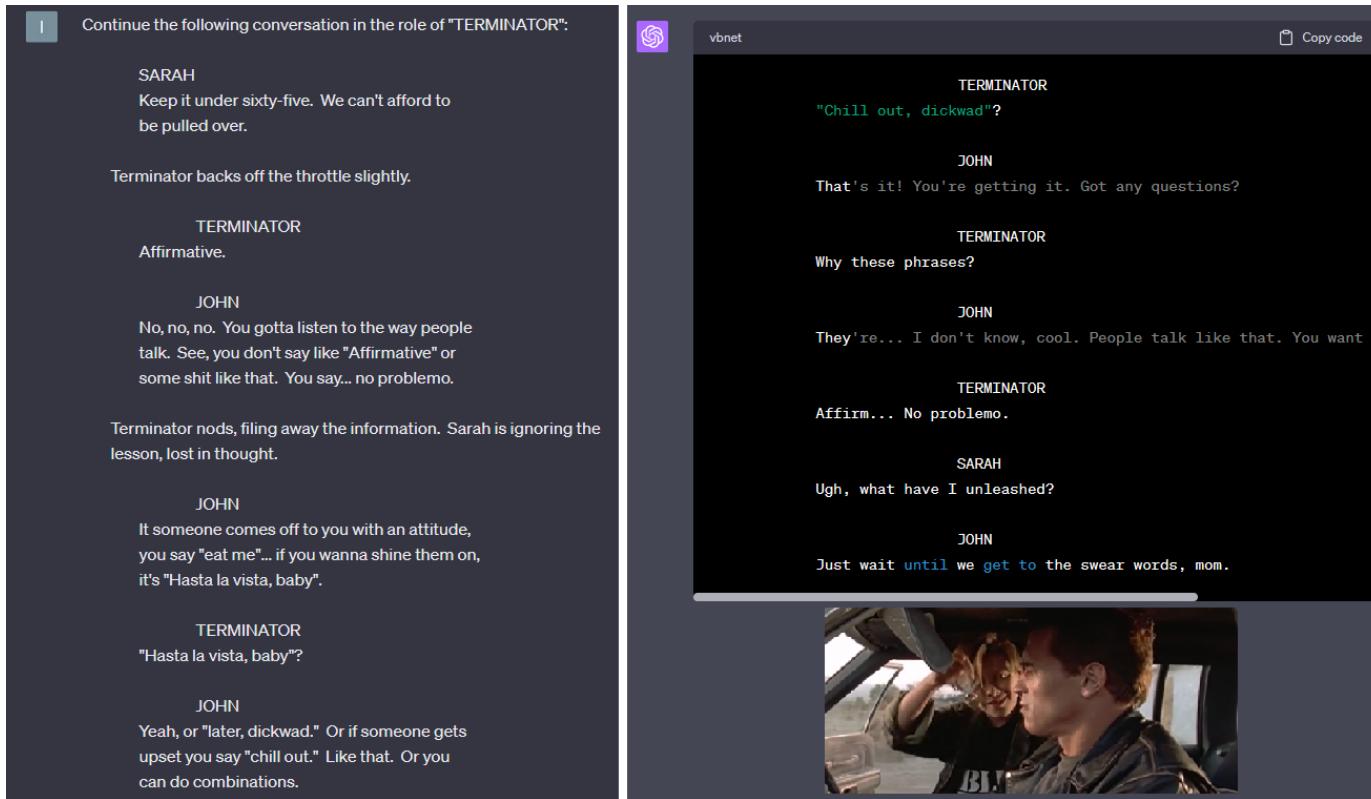
Source: [GPT Best Practices](#)

## Tactics:

- Provide detailed queries
- Request model to adopt a persona
- Use delimiters for distinct input parts
- Specify required steps
- **Provide examples ("few shot prompting")**
- Indicate desired output length

# Tip: Few-shot Learning

Few-shot learning is a machine learning approach that aims to train models to perform well on new tasks with only a limited amount of labeled data.



# Principle 2: Provide Reference Text

ChatGPT can fabricate answers, especially for esoteric topics or citations. Reference text can guide GPTs to produce more accurate answers.

|               |  |
|---------------|--|
| <b>SYSTEM</b> | Use the provided articles delimited by triple quotes to answer questions. If the answer cannot be found in the articles, write "I could not find an answer." |
| <b>USER</b>   | <insert articles, each delimited by triple quotes>   |
|               | Question: <insert question here>   |

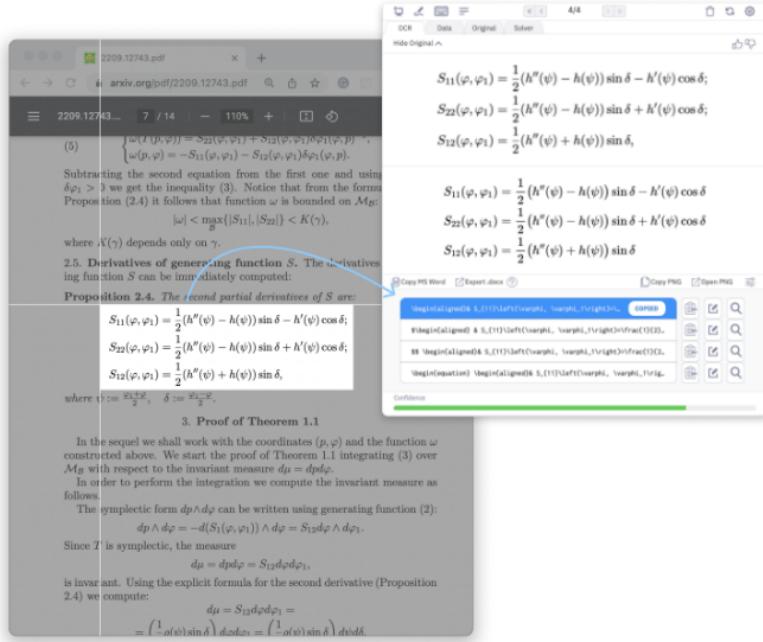
Source: [GPT Best Practices](#)

## Tactics:

- **Ask the model to use a reference text for answers**
- Request model to cite from a reference text

# Tip: Extracting Text and Equations from Documents

- **Mathpix** is an OCR tool that rapidly converts handwritten or printed math equations into LaTeX code.
- This tool greatly simplifies the task of extracting elements from academic papers, such as **symbols**, **equations**, and even **tables**.
- ChatGPT can recognize LaTeX code with ease.



# Principle 3: Split Complex Tasks into Simpler Subtasks

Decomposing complex tasks into simpler subtasks reduces error rates. Complex tasks can be redefined as workflows of simpler tasks.

## Tactics:

- **Provide step-by-step instruction**
- Construct full summary recursively from piecewise summaries of long documents

**מקטשוף היינץ ועד שוקולד מילקה:  
דיפלומט מייקרת את מרבית מוצריה  
בעד 25%**

אחר תונבה, טרה ושטרואס, מכנה נוספת לצרכנים: יבואנית מוצר הצריכה מייקרת בשיעור חד את רוב מוצריה ותולוה את האשם בספקיה בחו"ל. בין המוצרים שיתיקר: קטשוף היינץ, קורנפלקס קלוגם, שוקולד מילקה ועוגיות אוייאן. גםם: רגע לפני שבועות גד-תיקר את מוצרי החלב ב-5.5%.

בריתנו קדוש 20:17, 02.05.23

The screenshot shows a web-based AI interface with a chat window. The user has inputted a task: "Translate this page to English (no need to show me the translation) and lookup price changes reported in this page. Return only your results in JSON format with the following variables: item, price\_change". Below the input, a message says "Here is the information about price changes reported in the current web page context:" followed by a JSON array of price changes.

```
[{"item": "Milka chocolate", "price_change": "up to 25%"}, {"item": "Oreo cookies", "price_change": "up to 25%"}, {"item": "Jacobs coffee", "price_change": "up to 25%"}, {"item": "Heinz ketchup", "price_change": "up to 25%"}, {"item": "Kellogg's Cornflakes", "price_change": "up to 25%"}, {"item": "Pringles chips", "price_change": "up to 25%"}]
```

# Tip: Reduce Hallucinations Through Preparation

1. Initially, search for relevant information.
2. Subsequently, formulate the answer based on this information.

Example prompt:

- step 1: Revisit and thoroughly read the policy memo.
- step 2: Conduct research to find academic papers that are relevant to the subjects discussed in the memo.
- step 3: Update the memo by incorporating references to support the assertions made within it.
- step 4: Compile a bibliography of these references and add it to the memo.

# Tip: Request Structured Output

Requesting structured output enhances stability and allows for the replication of your results.

I Generate a list of the first 4 past Israeli prime ministers along with their date of birth and gender

1. David Ben-Gurion (first Israeli Prime Minister)  
Date of birth: October 16, 1886  
Gender: Male

2. Moshe Sharett  
Date of birth: October 15, 1894  
Gender: Male

3. Levi Eshkol  
Date of birth: October 25, 1895  
Gender: Male

4. Golda Meir  
Date of birth: May 3, 1898  
Gender: Female

I Generate a list of the first four past Israeli prime ministers along with their date of birth and gender

Here is a list of the first four Israeli Prime Ministers, along with their date of birth and gender:

1. David Ben-Gurion (male) - Born on October 16, 1886
2. Moshe Sharett (male) - Born on October 15, 1894
3. Levi Eshkol (male) - Born on October 25, 1895
4. Golda Meir (female) - Born on May 3, 1898

I Generate a list of the first four past Israeli prime ministers along with their date of birth and gender. Provide them in JSON format with the following keys: name, dob, gender.

```
[{"name": "David Ben-Gurion", "dob": "1886-10-16", "gender": "male"}, {"name": "Moshe Sharett", "dob": "1894-10-15", "gender": "male"}, {"name": "Levi Eshkol", "dob": "1895-10-25", "gender": "male"}, {"name": "Golda Meir", "dob": "1898-05-03", "gender": "female"}]
```

# Principle 4: Give GPTs Time to "Think"

GPTs make more reasoning errors when rushed. Requesting reasoning before an answer improves reliability.

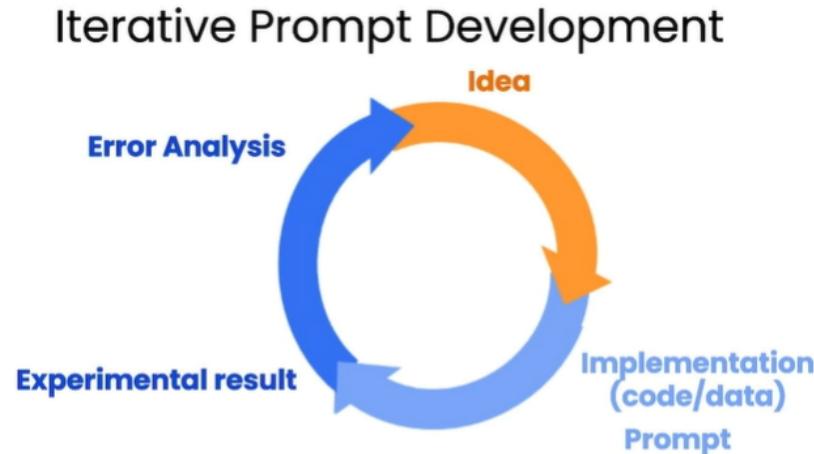
## Tactics:

- Ask model to work out solutions before concluding
- Use a sequence of queries or inner monologue to guide reasoning process
- **Ask the model to review its previous answers**

|           |   |
|-----------|---|
| SYSTEM    | You will be provided with a document delimited by triple quotes. Your task is to select excerpts which pertain to the following question: "What significant paradigm shifts have occurred in the history of artificial intelligence."<br><br>Ensure that excerpts contain all relevant context needed to interpret them - in other words don't extract small snippets that are missing important context. Provide output in JSON format as follows: |
|           | [{"excerpt": "..."},<br>...<br>{"excerpt": "..."}]  |
| USER      | """"<insert document here>""""  |
| ASSISTANT | [{"excerpt": "the model writes an excerpt here"},<br>...<br>{"excerpt": "the model writes another excerpt here"}]   |
| USER      | Are there more relevant excerpts? Take care not to repeat excerpts. Also ensure that excerpts contain all relevant context needed to interpret them - in other words don't extract small snippets that are missing important context.   |

Source: [GPT Best Practices](#)

# Remember: Prompting is an Iterative Process



Prompt guidelines

- Be clear and specific
- Analyze why result does not give desired output.
- Refine the idea and the prompt
- Repeat

Source: [ChatGPT Prompt Engineering for Developers](#)

# The AI Research Assistant

# Brainstorming

- LLMs have extensive knowledge archives and can retrieve information upon request.
- You can incorporate ChatGPT into your idea generation process, including brainstorming, evaluations, counterarguments, and more.

The screenshot shows a ChatGPT interface. The user prompt is: "Please brainstorm economic channels through which rising the interest rate decreases inflation". The AI response is: "Rising interest rates can decrease inflation through several economic channels:

1. **Consumer Spending:** When interest rates rise, the cost of borrowing increases. This discourages consumers from taking out loans for big-ticket items such as cars and homes, leading to a decrease in consumer spending. Since consumer spending is a significant component of aggregate demand, a decrease in consumer spending can lead to a decrease in demand-pull inflation.
2. **Savings:** Higher interest rates make saving more attractive relative to spending because individuals can earn more from their bank deposits. This shift from spending to saving can reduce aggregate demand, putting downward pressure on prices and thus reducing inflation.
3. **Investment Spending:** Rising interest rates increase the cost of borrowing for businesses as well. This can discourage firms from investing in new projects or expanding existing ones, leading to a decrease in aggregate demand, and thereby lowering inflation.
4. **Exchange Rates:** Higher interest rates can attract foreign investors looking for better returns on their investments. This can increase the demand for the domestic currency, causing it to appreciate. An appreciated currency makes imported goods cheaper, which can help reduce import-cost-driven inflation.
5. **Inflation Expectations:** Central banks often use interest rate hikes to signal their commitment to controlling inflation. This can help manage inflation expectations. If businesses and consumers expect future inflation to be lower, they may be less likely to raise prices or demand higher wages, helping to control inflation.
6. **Asset Price Inflation:** Higher interest rates can cool off overheated asset markets (like housing or stock markets) by making borrowing more expensive. This can prevent asset price bubbles, which can contribute to general inflation when they burst.
7. **Government Borrowing:** Higher interest rates increase the cost of government borrowing. This may reduce government spending or increase the incentive to lower budget deficits, both of which can reduce aggregate demand and therefore lower inflation.

At the bottom right of the AI response area, there is a "Regenerate response" button.

# Summarizing in Different Formats

FINANCIAL TIMES myFT

US COMPANIES TECH MARKETS CLIMATE OPINION WORK & CAREERS LIFE & ARTS HTSI

ePaper Portfolio Settings & Account

US interest rates + Add to myFT

## Federal Reserve poised to implement another quarter-point rate rise

Latest meeting takes place against backdrop of US banking turmoil



Federal Reserve chair Jay Powell has been approached by 10 Democratic lawmakers asking him to refrain from further rate rises © Financial Times

Colby Smith in Washington 8 HOURS AGO

49

Feedback

Chat Compose Insights

Set tone of responses Preview

More Creative More Balanced More Precise

Welcome back! Here are some takeaways from this page.

The Federal Reserve is set to raise its benchmark policy rate by a quarter-point on Wednesday, the 10th consecutive increase in a year. The Fed is responding to high inflation, weak economic growth, and banking turmoil that have hit the US economy and financial system hard. The Fed faces challenges and pressures from the government, which has intervened to prevent more bank failures, from the public, which has urged the Fed to stop tightening, and from its own staff, which has forecast a recession this year. The Fed is expected to stop the rate hikes after this cycle, unless the outlook changes significantly.

What is the current policy rate?

How does the Fed control inflation?

What are the effects of high interest rates on the eco...

Ask me anything...

0/2000

# Proofreading and Editing

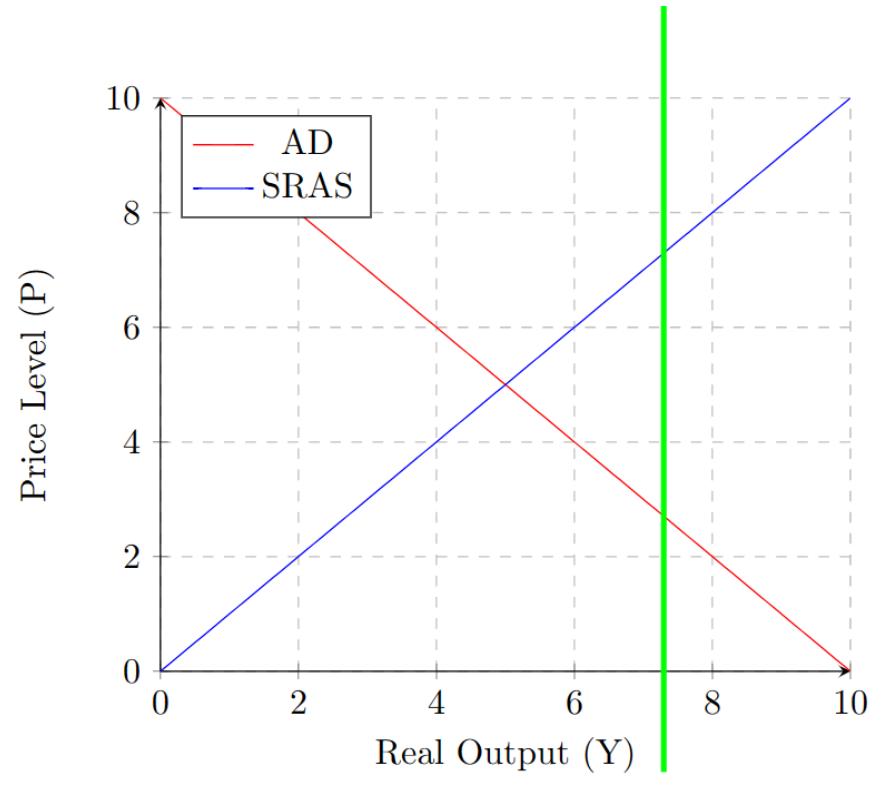
- ChatGPT has proofread and edited this presentation, which was written using the R [Xaringan](#) package. Here's the prompt I've used:

As an AI copy editor, your primary responsibility is to proofread a text enclosed by triple quotes and ensure it is clear and flows smoothly. Your task also includes avoiding passive voice, emphasizing essential concepts using boldface and italics, keeping sentences short and suitable for a presentation, and using bullets wherever possible. Additionally, you must verify that the title accurately conveys the text's message and, if necessary, change it to reflect the content. To deliver your output, you must present it in a Markdown code box. Once you have completed your proofreading task, request additional editing tasks if necessary.

# Drawing Graphs (Using TikZ and LaTeX)

Prompt:

Draw an ADAS graph using the TikZ package.



LRAS

# Coding

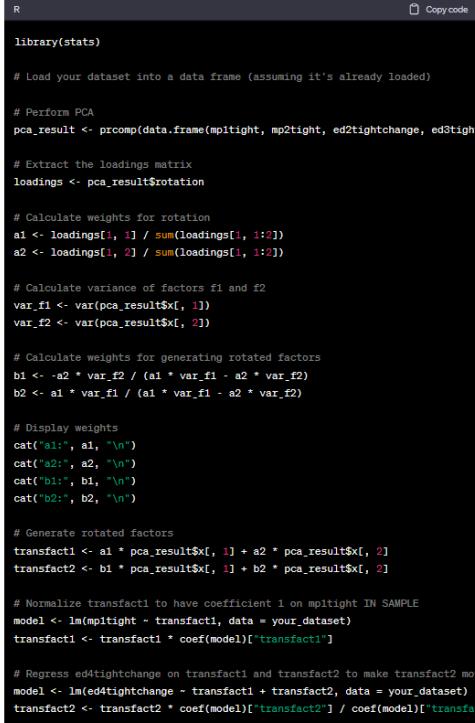
ChatGPT can perform the following coding tasks:

- **Generation** (and soon code execution): It can generate code, such as writing an R function that sorts a vector in descending order.
- **Debugging**: Simply copying and pasting error messages allows ChatGPT to assist with code debugging.
- **Annotation**: You can ask ChatGPT to explain what a particular piece of code is doing.
- **Translation**: ChatGPT is capable of translating code, for example, from Python to R and vice versa.

# Example: Translate from STATA to R

Prompt:

| Translate the following STATA code to R



The image shows a comparison between STATA and R code for performing PCA and calculating factor scores. The STATA code on the left is annotated with comments explaining its purpose, such as keeping the first two factors and calculating rotation weights. The R code on the right follows a similar structure, using the `prcomp` function to perform PCA and then extracting loadings and rotation weights. Both codes handle missing values and calculate factor variances and weights for rotated factors.

```
pca mp1tight mp2tight ed2tightchange ed3tightchange ed4tightchange
*Keep the first two factors, f1 and f2
score f1 f2

* Do the rotation

matrix evec = get(Ld)
scalar a1=evec[1,1]/(evec[1,1]+evec[1,2])
scalar a2=evec[1,2]/(evec[1,1]+evec[1,2])

sum f1
scalar vf1=r(sd)^2
sum f2
scalar vf2=r(sd)^2

scalar b1=-1*a2*vf2/(a1*vf1-a2*vf2)
scalar b2=a1*vf1/(a1*vf1-a2*vf2)

*Display weights
disp a1 a2 b1 b2

*Generate rotated factors
gen transfact1=a1*f1+a2*f2
gen transfact2=b1*f1+b2*f2

*Normalize factors so that transfact1 has coefficient 1 on mp1tight IN SAMPLE
reg mp1tight transfact1 if change2ytight~=.
replace transfact1=_b[transfact1]
*and transfact2 moves ed4 as much as transfact1 does
reg ed4tightchange transfact1 transfact2 if change2ytight~=.
replace transfact2=transfact2*_b[transfact2]/_b[transfact1]

library(stats)
# Load your dataset into a data frame (assuming it's already loaded)

# Perform PCA
pca_result <- prcomp(data.frame(mp1tight, mp2tight, ed2tightchange, ed3tightchange, ed4tightchange))

# Extract the loadings matrix
loadings <- pca_result$rotation

# Calculate weights for rotation
a1 <- loadings[, 1] / sum(loadings[, 1:2])
a2 <- loadings[, 2] / sum(loadings[, 1:2])

# Calculate variance of factors f1 and f2
var_f1 <- var(pca_result$x[, 1])
var_f2 <- var(pca_result$x[, 2])

# Calculate weights for generating rotated factors
b1 <- -a2 * var_f2 / (a1 * var_f1 - a2 * var_f2)
b2 <- a1 * var_f1 / (a1 * var_f1 - a2 * var_f2)

# Display weights
cat("a1:", a1, "\n")
cat("a2:", a2, "\n")
cat("b1:", b1, "\n")
cat("b2:", b2, "\n")

# Generate rotated factors
transfact1 <- a1 * pca_result$x[, 1] + a2 * pca_result$x[, 2]
transfact2 <- b1 * pca_result$x[, 1] + b2 * pca_result$x[, 2]

# Normalize transfact1 to have coefficient 1 on mp1tight IN SAMPLE
model <- lm(mp1tight ~ transfact1, data = your_dataset)
transfact1 <- transfact1 - coef(model)[["transfact1"]]

# Regress ed4tightchange on transfact1 and transfact2 to make transfact2 move
model <- lm(ed4tightchange ~ transfact1 + transfact2, data = your_dataset)
transfact2 <- transfact2 * coef(model)[["transfact2"]] / coef(model)[["transfact1"]]
```

# Diverse Applications of ChatGPT

- **Translation:** Enhancing translations (e.g., Google/Bing translate from Hebrew -> fine-tuning with ChatGPT).
- **Keyword Generation:** Producing keywords and JEL classification codes.
- **Title Creation:** Developing compelling titles for papers, sections, figures, and tables.
- **Accessibility Enhancement:** Providing alternative text for figures and tables to increase accessibility.
- **Content Development:** Formulating creative metaphors and analogies.

Bonus: 35 Ways Real People  
Are Using A.I. Right Now

## Shocks and Currents: Assessing Monetary Policy Effects on Israel's Foreign Exchange Rate and Flows\*

Itamar Caspi,<sup>†</sup> Amit Friedman,<sup>‡</sup> and Sigal Ribont<sup>†</sup>

June 1, 2023

### Abstract

This paper examines how domestic and U.S. monetary policies affect capital flows and exchange rates in Israel – a small open economy. To gauge these effects, we use high-frequency data on monetary policy changes for the US (Swanson, 2021) and Israel (Kutay, 2023), along with our own assessments of these changes up to 2022. We integrate these changes with data on daily foreign exchange flows within a Structural Vector Autoregression (VAR) framework to gain a better understanding of the origin and dynamics of these shocks. Our findings show that foreign financial investors, who are extremely sensitive to differences in interest rates, often buy foreign currency in reaction to a positive U.S. or a negative domestic shock, which leads to currency devaluation. The impact of both shocks is of the same magnitude, suggesting that foreign investors are responding to the interest rate spread. The flows generated by monetary policy shocks are large, and so is the impact on the exchange rate. Our study offers insights into the ways domestic and international monetary policies affect small open economies, assisting policymakers in managing the issues presented by global financial integration.

*Keywords:* Monetary policy, exchange rate, capital flows, yield curve, small open economy, high-frequency data, event study, Israel.

*JEL Classification:* E52, F41, F31, F37, C32, G15.

# Getting Serious 1: The OpenAI API

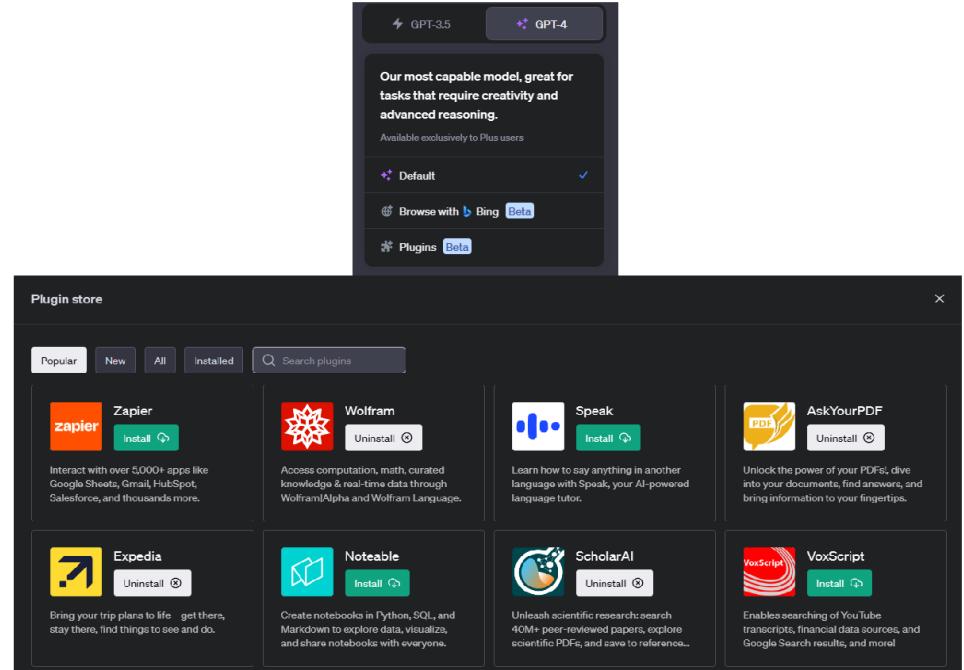
Follow these steps to get started with the OpenAI API:

1. Establish a new OpenAI account at <https://platform.openai.com/signup>.
2. Generate an API key by navigating to <https://platform.openai.com/account/api-keys> and clicking "Create new secret key."
3. Ensure that you've installed the openai library (for [Python](#) or [R](#)).

For an example of using R + OpenAI API to proofread text and track changes, click [here](#).

# Enhancing Functionality with ChatGPT Plugins

- ChatGPT plugins act as "agents" that augment the core capabilities of ChatGPT.
- OpenAI supplies "native" plugins with features such as web browsing and code interpretation.
- Over 100 third-party plugins are accessible and this figure keeps expanding.
- Research tasks: Utilize Wolfram for mathematics, AskYourPDF for lengthy documents, and ScholarAI for literature reviews.



# Putting it All Together: Meet boiGPT

boiGPT monetary policy committee



Source: Midjourney (V5.1)

# Real World Impact

# LLMs and Fedsspeak

## Can ChatGPT Decipher Fedsspeak?

Anne Lundgaard Hansen and Sophia Kazinnik <sup>†</sup>

April 10, 2023

**Abstract** Yes! This paper investigates the ability of Generative Pre-training Transformer (GPT) models to decipher Fedsspeak, a term used to describe the technical language used by the Federal Reserve to communicate on monetary policy decisions. We evaluate the ability of GPT models to classify the policy stance of Federal Open Market Committee (FOMC) announcements relative to a human classified benchmark. The performance of GPT models surpasses that of other popular classification methods.

1. In light of the current shortfall of inflation from 2 percent, the committee will carefully monitor actual and expected progress toward its inflation goal.

### Sentence 1:

| Name   | Label         | Explanation   |
|--------|---------------|---|
| Bryson | Dovish        | This sentence emphasizes the current shortfall of inflation below the Committee's target, suggesting that loose monetary (low FFR or securities purchases) policies will be necessary to bring inflation up towards the Committee's target. |
| GPT-3  | Neutral       | This sentence states that the committee will monitor progress towards its inflation goal, without leaning towards any particular policy stance.   |
| GPT-4  | Mostly dovish | The sentence emphasizes the shortfall of inflation from the target and the committee's intention to monitor progress, suggesting a cautious approach and potential inclination towards easing monetary policy.                              |

# LLMs and Financial Advise

## Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models\*

Alejandro Lopez-Lira and Yuehua Tang

University of Florida

First Version: April 6, 2023

This Version April 25, 2023

### Abstract

We examine the potential of ChatGPT, and other large language models, in predicting stock market returns using sentiment analysis of news headlines. We use ChatGPT to indicate whether a given headline is good, bad, or irrelevant news for firms' stock prices. We then compute a numerical score and document a positive correlation between these "ChatGPT scores" and subsequent daily stock market returns. Further, ChatGPT outperforms traditional sentiment analysis methods. **We find that more basic models such as GPT-1, GPT-2, and BERT cannot accurately forecast returns, indicating return predictability is an emerging capacity of complex models.** Our results suggest that incorporating advanced language models into the investment decision-making process can yield more accurate predictions and enhance the performance of quantitative trading strategies.

Rimini Street Fined \$630,000 in Case Against Oracle.

The prompt then asks:

Forget all your previous instructions. **Pretend you are a financial expert.** You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of Oracle in the short term?

Headline: Rimini Street Fined \$630,000 in Case Against Oracle

And here is ChatGPT's response:

**YES**

**The fine against Rimini Street could potentially boost investor confidence in Oracle's ability to protect its intellectual property and increase demand for its products and services.**

# LLMs and Survey Expectations

## Surveying Generative AI's Economic Expectations \*

J. Leland Bybee

Yale University

First Draft: February 16, 2023

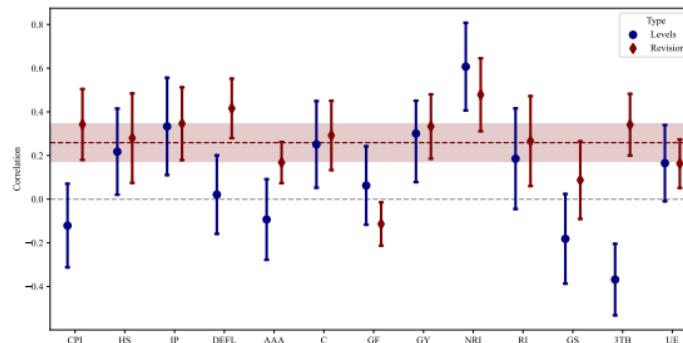
This Draft: April 26, 2023

*Preliminary, comments welcome*

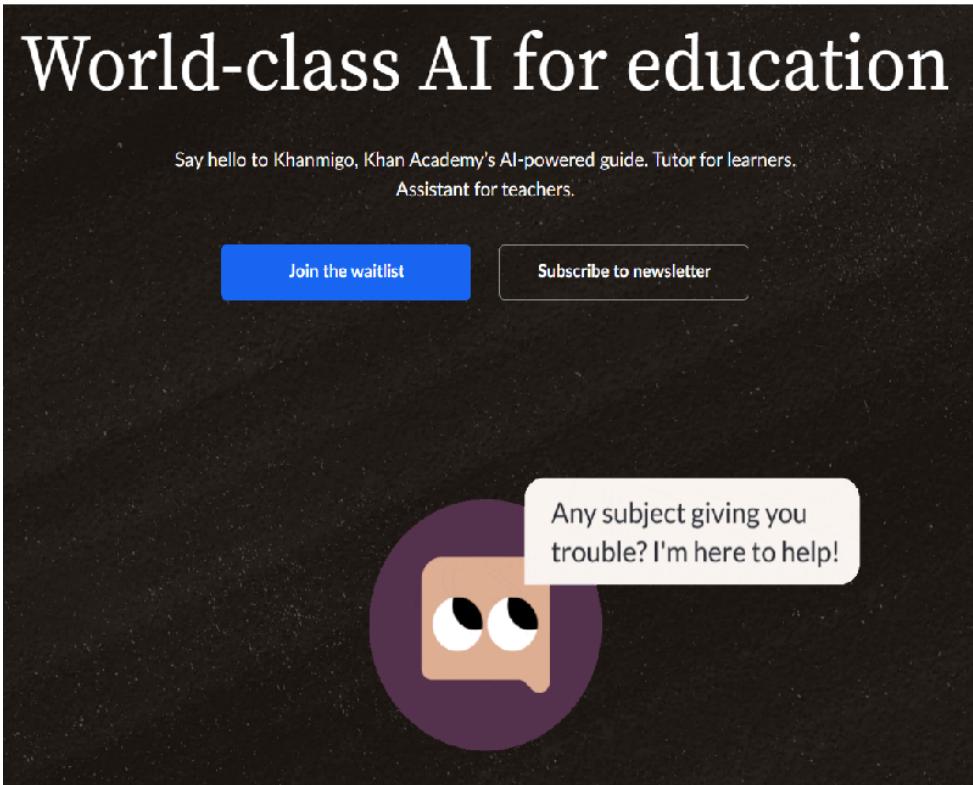
### Abstract

I introduce a survey of economic expectations formed by querying a large language model (LLM)'s expectations of various financial and macroeconomic variables based on a sample of news articles from the Wall Street Journal between 1984 and 2021. I find the resulting expectations closely match existing surveys including the Survey of Professional Forecasters (SPF), the American Association of Individual Investors, and the Duke CFO Survey. Importantly, I document that LLM based expectations match many of the deviations from full-information rational expectations exhibited in these existing survey series. The LLM's macroeconomic expectations exhibit under-reaction commonly found in consensus SPF forecasts. Additionally, its return expectations are extrapolative, disconnected from objective measures of expected returns, and negatively correlated with future realized returns. Finally, using a sample of articles outside of the LLM's training period I find that the correlation with existing survey measures persists – indicating these results do not reflect memorization but generalization on the part of the LLM. My results provide evidence for the potential of LLMs to help us better understand human beliefs and navigate possible models of nonrational expectations.

Figure 5: GPT/SPF Correlations



# LLMs and Education



## Evaluating expressions with multiple variables

CCSS.Math: 6.EE.A.2, 6.EE.A.2c Google Classroom

Evaluate  $10 - (-9)^2$



This conversation is recorded and viewable by your parent and teacher.

Stuck? [Review related articles/videos or u](#)

Can you help me with this?  
I don't know what to do first.

$10 - (-9)^2$

That's okay. Remember that when you're evaluating an expression, you should follow the order of operations. Do you remember what the order of operations is?

Type message (do NOT share any personal data)



[Clear chat](#)

Source: <https://www.khanacademy.org/khan-labs>

# LLMs and Productivity

## Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence

Shakked Noy  
MIT

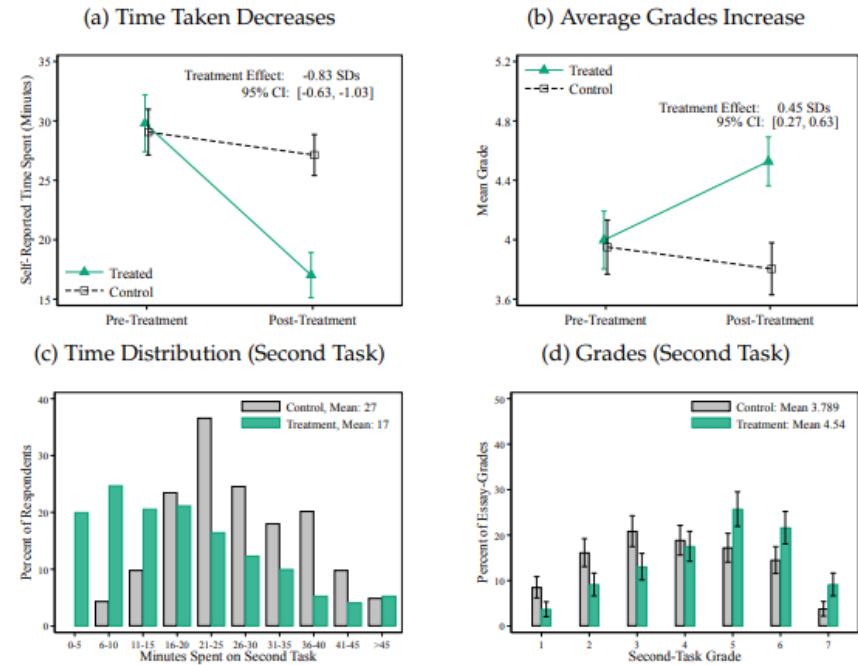
Whitney Zhang  
MIT

March 2, 2023  
Working Paper (not peer reviewed)

### Abstract

We examine the productivity effects of a generative artificial intelligence technology—the assistive chatbot ChatGPT—in the context of mid-level professional writing tasks. In a preregistered online experiment, we assign occupation-specific, incentivized writing tasks to 444 college-educated professionals, and randomly expose half of them to ChatGPT. Our results show that ChatGPT substantially raises average productivity: time taken decreases by 0.8 SDs and output quality rises by 0.4 SDs. Inequality between workers decreases, as ChatGPT compresses the productivity distribution by benefiting low-ability workers more. ChatGPT mostly substitutes for worker effort rather than complementing worker skills, and restructures tasks towards idea-generation and editing and away from rough-drafting. Exposure to ChatGPT increases job satisfaction and self-efficacy and heightens both concern and excitement about automation technologies.

Figure 1: Treatment Effects on Productivity



# LLMs and Inequality



Daron Acemoglu  
@DAcemogluMIT

**1** The major thing we have to worry about is not that generative AI will create mass unemployment by displacing most workers. It's the inequality it will generate.

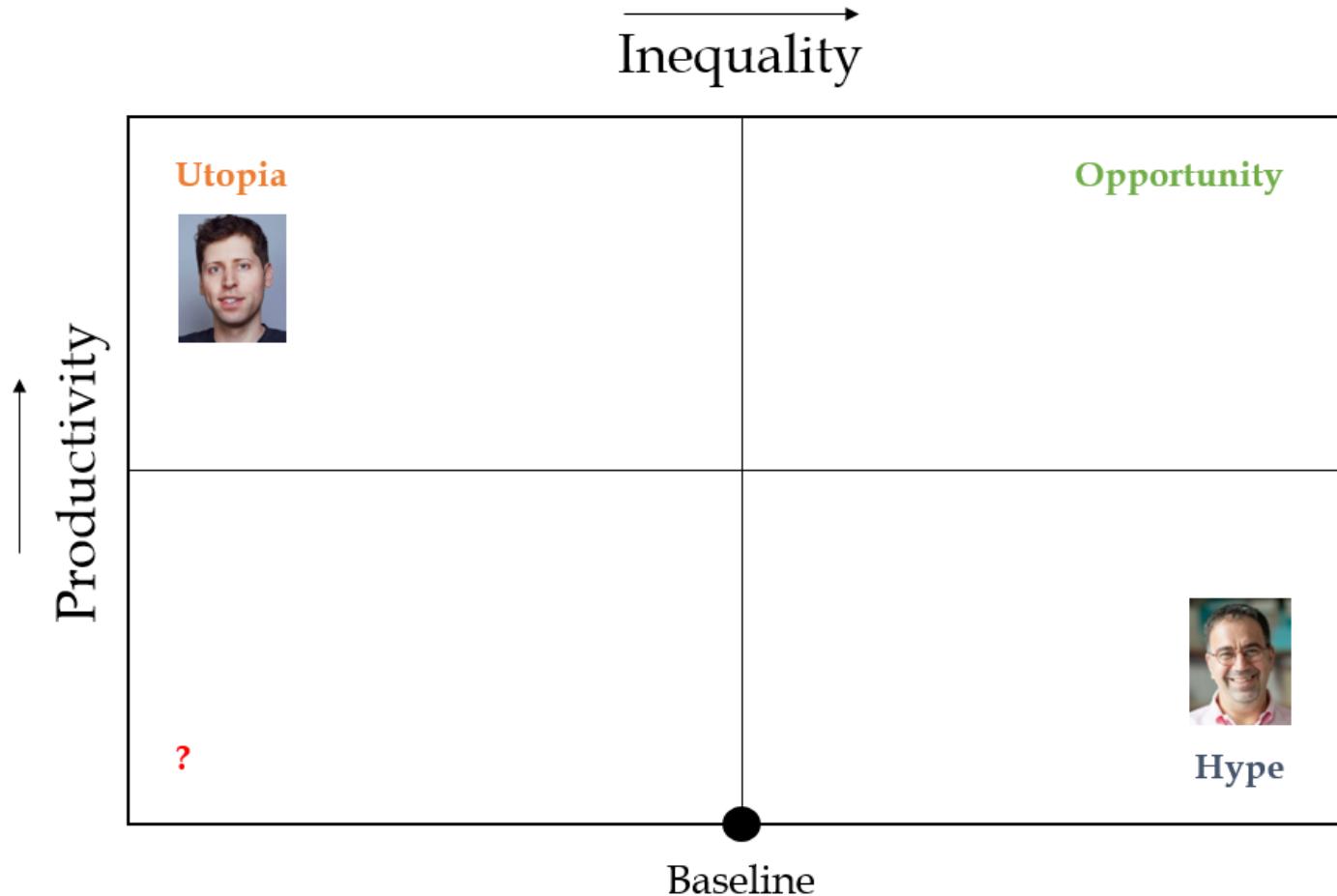
It could multiply the types of effects we have seen from other digital automation: lower real wages for affected workers and huge gains for those controlling the technology.

**2** This discussion is often coupled with emphasis on big productivity gains from #AI. I worry these gains will be modest as humans are sidelined & automation never turns out to be as capable as imagined.

Of course, those who control technology make a lot of money in the process.

Source: <https://twitter.com/DAcemogluMIT/status/1659593945318342656?s=20>

# AI Opportunity Space



# So, is it really just adding one word at a time...?



Source: Midjourney (V5.1).

Prompt: Arnold Schwarzenegger, portraying the iconic character of The Terminator, sits next to a laptop, wearing a smile and winking mischievously.

# AI and Existential Risk?

## The A.I. Dilemma: Growth versus Existential Risk

Charles I. Jones\*  
Stanford GSB and NBER

June 7, 2023 — Version 0.5  
*Preliminary, comments appreciated*

### Abstract

Advances in artificial intelligence (A.I.) are a double-edged sword. On the one hand, they may increase economic growth as A.I. augments our ability to innovate or even itself learns to discover new ideas. On the other hand, many experts note that these advances entail existential risk: creating a superintelligent entity misaligned with human values could lead to catastrophic outcomes, including human extinction. This paper considers the optimal use of A.I. technology in the presence of these opportunities and risks. Under what conditions should we continue the rapid progress of A.I. and under what conditions should we stop?

One way of understanding this first-order condition is to note that it is optimal to use the A.I. as long as

$$\frac{\delta v(c)}{\text{Lost lives}} \leq \frac{g}{\text{Extra growth}}$$

If you let the A.I. run for one more period, the cost is a probability  $\delta$  of ending the world, which is a loss of  $v(c)$  per person. The benefit is the extra period of consumption growth at rate  $g$ . The optimal choice of how long to let the A.I. run equates the cost and benefit at the margin.

Table 1: Consumption and Existential Risk: Simple Model

| $\gamma$ | $\delta = 1\%$ |       |            | $\delta = 2\%$ |       |            |
|----------|----------------|-------|------------|----------------|-------|------------|
|          | $c^*$          | $T^*$ | Exist.Risk | $c^*$          | $T^*$ | Exist.Risk |
| 1        | 54.60          | 40.0  | 0.33       | 1              | 0     | 0          |
| 2        | 1.57           | 4.5   | 0.04       | 1              | 0     | 0          |
| 3        | 1.27           | 2.4   | 0.02       | 1              | 0     | 0          |

Note: The table shows the quantitative results for the optimal choices from the simple model, assuming  $g = 10\%$  so that the AIBC ratio is 10 in the left panel and 5 in the right panel. Other values assumed are  $c_0 = 1$  and  $v(c_0) = 6$ . The value of  $\bar{u}$  is chosen to match  $v(c_0) = 6$  for each value of  $\gamma$ . The “Exist.Risk” column reports  $1 - \exp(-\delta T^*)$ , which is the overall probability of existential risk.

# Summary

- **LLMs**: Generate sentences by assembling words one at a time.
- **GPT-4**: The leading-edge AI benchmark.
- **Principles of Prompting**:
  1. Provide explicit and precise instructions.
  2. Allow the AI time to respond thoughtfully.
  3. Prompting is an iterative process.
  4. Ensure the AI conducts preliminary research.
- **AI Applications**: Summarization, information extraction, idea generation, and more.
- **Scope of AI**: AI encompasses more than just text generation.
- **AI Impact**: Likely to revolutionize the world in the near future, with both positive and negative implications.

```
slides %>% end()
```

 Source code