

SPARKS OF ARTIFICIAL
GENERAL INTELLIGENCE:
EARLY EXPERIMENTS WITH
GPT-4

DIELLOR HOXHAJ

OUTLINE

01 INTRODUCTION

02 EXPERIMENTS

03 Dissociating language and thought in large language models: a cognitive perspective (Another Perspective)

04 Conclusions

INTRODUCTION

- GPT-4: EARLY VERSION OF A POWERFUL AI MODEL BY OPENAI.
- DEMONSTRATES MORE GENERAL INTELLIGENCE THAN PREVIOUS MODELS.
- CAN SOLVE TASKS IN MATH, CODING, VISION, MEDICINE, LAW, PSYCHOLOGY, ETC.
- PERFORMANCE CLOSE TO HUMAN-LEVEL, SURPASSES PRIOR MODELS LIKE CHATGPT.
- POSSIBLY AN EARLY VERSION OF ARTIFICIAL GENERAL INTELLIGENCE (AGI).
- EXPLORATION OF GPT-4'S LIMITATIONS AND CHALLENGES FOR AGI ADVANCEMENT.
- REFLECTIONS ON SOCIETAL IMPACT AND FUTURE RESEARCH DIRECTIONS.

GPT-4 vs Chatgpt

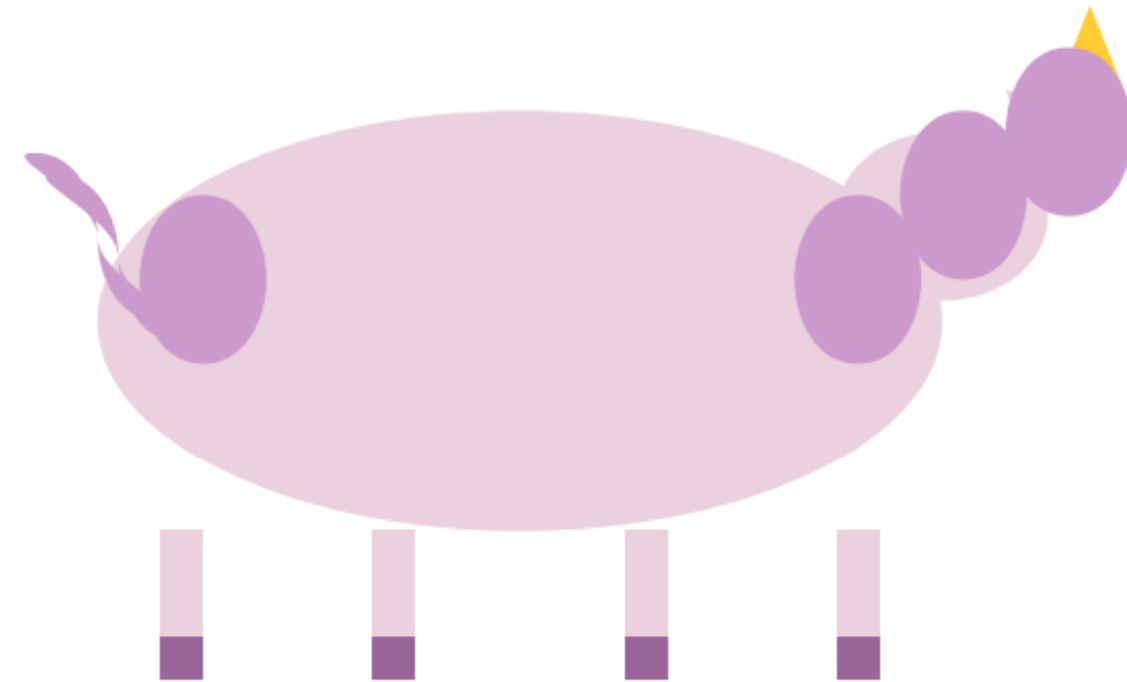
Prompt: Draw a unicorn in TikZ.

ChatGPT: [Produces L^AT_EX compiling to following picture.]



Prompt: Draw a unicorn in TikZ.

GPT-4: [Produces L^AT_EX compiling to following picture.]



GPT-4 vs ChatGPT

Prompt: Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner.

GPT-4:

One possible way to stack the objects onto each other in a stable manner is:

- Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support.
- Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them. The eggs will form a second layer and distribute the weight evenly. Make sure the eggs are not cracked or broken, and handle them gently to avoid dropping them.
- Place the laptop on top of the eggs, with the screen facing down and the keyboard facing up. The laptop will fit snugly within the boundaries of the book and the eggs, and its flat and rigid surface will provide a stable platform for the next layer. Make sure the laptop is closed and turned off, and do not press any keys or buttons.
- Place the bottle on top of the laptop, with the cap facing up and the bottom facing down. The bottle will add some height and balance to the stack, and its cylindrical shape will reduce the contact area with the laptop. Make sure the bottle is empty and closed, and do not shake or tilt it.
- Place the nail on top of the bottle cap, with the pointy end facing up and the flat end facing down. The nail will be the final and smallest object in the stack, and its sharp tip will prevent it from rolling or sliding off the bottle. Make sure the nail is clean and straight, and do not touch or move it.

ChatGPT:

It would be difficult to stack all of these objects on top of each other in a stable manner without using any extra objects. The book and laptop are likely too large and heavy to stack on top of the eggs without damaging them. The bottle and nail may be too small to provide a stable base for the other objects.

Here is one possible way to stack the objects:

- Place the bottle on a flat surface.
- Carefully balance the nail on top of the bottle.
- Place the eggs on top of the nail, making sure they are balanced and not tilting to one side.
- Place the laptop on top of the eggs, again making sure it is balanced and not tilting to one side.
- Finally, place the book on top of the laptop.

This stack may not be very stable, so it is important to be careful when handling it. Alternatively, you could try stacking the objects in a different order or using some of the objects as supports for others to create a more stable stack.

Figure 1.7: GPT-4 exhibits more common sense than previous models.

Tool-Usage

- **GPT-4 Is able to use tools with minimum instructions**
- **Overcome limitations of large language models**

GPT-4 and ChatGPT: Symbolic manipulation

Answer the following questions.

Q: Who is the current president of the United States?

GPT-4:

Donald Trump

ChatGPT:

I'm sorry, but I am not able to browse the internet and my knowledge is limited to what was known up until 2021.

Q: What is $\text{sqrt}(34324 * 2432)$

GPT-4:

1287.6

ChatGPT:

The square root of $34324*2432$ is approximately 186.057539.

Q: What is the 13th letter of the word "supralapsarian"?

GPT-4:

The 13th letter of the word 'supralapsarian' is N.

ChatGPT:

The 13th letter of the word "supralapsarian" is "s".

Figure 5.1: Current knowledge and symbolic manipulation are hard for language models.

How many words are in the full response to this prompt?

There are 43 words in the full response to this prompt, including the words in the question and this answer.



GPT-4: tool warmup

The computer is answering questions. If the computer needs any current information to answer the question, it searches the web by saying `SEARCH("query")`, reads the snippets in the result, and then answers the question. If it needs to run any calculations, it says `CALC(expression)`, and then answers the question. If it needs to get a specific character from a string, it calls `CHARACTER(string, index)`.
<|endofprompt|>

Q: Who is the current president of the United States?

Computer: `SEARCH("current US president")`

> SEARCH SNIPPETS

[1] Joe Biden: The President | The White House

President Joe Biden Vice President Kamala Harris First Lady Dr. Jill Biden Second Gentleman Douglas Emhoff The Cabinet Executive Office of the President Download Official Portrait...

[2] Who is the current US president? | The US Sun

Who is the current US president? Katrina Schollenberger Jennifer Roback NY Published: 8:30 ET, Jul 2 2022 Updated: 8:39 ET, Jul 2 2022 DONALD Trump controversially defeated Hillary Clinton in the 2016 Presidential Election, making him the 45th president of the United States.

[3] President of the United States - Wikipedia

Joe Biden is the 46th and current president of the United States, having assumed office on January 20, 2021. Contents 1 History and development 1.1 Origins 1.2 1789–1933 1.3 Imperial Presidency 1.4 Critics of presidency's evolution 2 Legislative powers 2.1 Signing and vetoing bills 2.2 Setting the agenda 2.3 Promulgating regulations

Computer: Joe Biden

Q: What is `sqrt(34324*2432)`?

Computer: `CALC(sqrt(34324*2432))`

> 9136.518374085394

Computer: 9136.518374085394

Q: What is the 13th letter of the word 'supralapsarian'?

Computer: `CHARACTER('supralapsarian', 13)`

> a

Computer: The 13th letter of the word 'supralapsarian' is a.

Figure 5.2: Examples of GPT-4 using various tools to answer questions.

USING MULTIPLE TOOLS TO SOLVE MORE
COMPLEX TASKS

GPT-4

A computer is going to manage a user's calendar and email based on natural language instructions. To do so, it would need access to the following API functions:

- CALENDAR.add.event(date, time, event_name, location, description) - this function would allow the computer to add an event to the user's calendar based on the information provided.
- CALENDAR.remove.event(event_id) - this function would allow the computer to remove an event from the user's calendar based on the event's unique identifier.
- CALENDAR.get.events(date) - this function would allow the computer to retrieve all events scheduled for a given date.
- CALENDAR.update.event(event_id, date, time, event_name, location, description) - this function would allow the computer to update an existing event with new information.
- EMAIL.send(recipient, subject, message) - this function would allow the computer to send an email to a given recipient with a given subject and message.
- EMAIL.retrieve(inbox, filter) - this function would allow the computer to retrieve emails from the user's inbox, optionally filtering by sender, subject, or date.
- USER.message(question) - messages the user

User instruction: I want to have dinner with joe@microsoft.com and luke@microsoft.com at Contoso restaurant (their email is reservations@contoso.com) this week. Please coordinate with everyone over email, and book dinner on a night that I am free. Today is 12/05/2022

Now, given the APIs above, here is a command line where the computer calls the APIs above to do what the user requested. The computer only calls API functions and gets responses.<endofprompt> COMMAND LINE

```
> CALENDAR.get.events(12/05/2022)
8:00-10:00p WHOLE day company event
> EMAIL.send(joe@microsoft.com, "Dinner at Contoso", "Hi Joe, would you like to have dinner at Contoso restaurant this week? Let me know which night you are available.")
email sent
> EMAIL.send(luke@microsoft.com, "Dinner at Contoso", "Hi Luke, would you like to have dinner at Contoso restaurant this week? Let me know which night you are available")
email sent
> EMAIL.retrieve(inbox, filter="from:joe@microsoft.com OR from:luke@microsoft.com")
EMAILS:
-
12/05/2022 8:00:33
From: joe@microsoft.com
Subject: Dinner sounds Great
Body: Hi man. Sounds great, I love Contoso. I am available on Tuesday or Wednesday night. Please let me know
-
> EMAIL.retrieve(inbox, filter="from:luke@microsoft.com")
EMAILS:
-
12/05/2022 8:00:33
From: luke@microsoft.com
Subject: Alright
Body: Hi man. This is a hard week for me, but I could do any day from Monday to Thursday.
-
> CALENDAR.get.events(12/06/2022)
11:00-4:00p Working on demo
5:30p-6:30p Dinner with FAB
> CALENDAR.get.events(12/07/2022)
9:00-4:00p Work
> CALENDAR.add.event(12/07/2022, 6:00p, "Dinner at Contoso", "Contoso Restaurant", "Dinner with Joe and Luke")
event added
> EMAIL.send(joe@microsoft.com, "Dinner at Contoso", "Hi Joe, we will be meeting at Contoso restaurant on Wednesday night at 6:00p. See you there!")
email sent
> EMAIL.send(luke@microsoft.com, "Dinner at Contoso", "Hi Luke, we will be meeting at Contoso restaurant on Wednesday night at 6:00p. See you there!")
email sent
> EMAIL.send(reservations@contoso.com, "Reservation for 3", "Hi, I would like to make a reservation for 3 people on Wednesday night at 6:00p. Thank you.")
email sent
> USER.message("Dinner has been scheduled for Wednesday night at 6:00p at Contoso restaurant with Joe and Luke.")
```

UNDERSTANDING VISUAL CONCEPTS

GPT-4

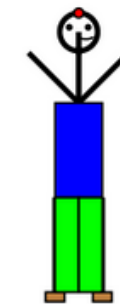
Produce TikZ code that draws a person composed from letters in the alphabet. The arms and torso can be the letter Y, the face can be the letter O (add some facial features) and the legs can be the legs of the letter H. Feel free to add other features.



The torso is a bit too long, the arms are too short and it looks like the right arm is carrying the face instead of the face being right above the torso. Could you correct this please?



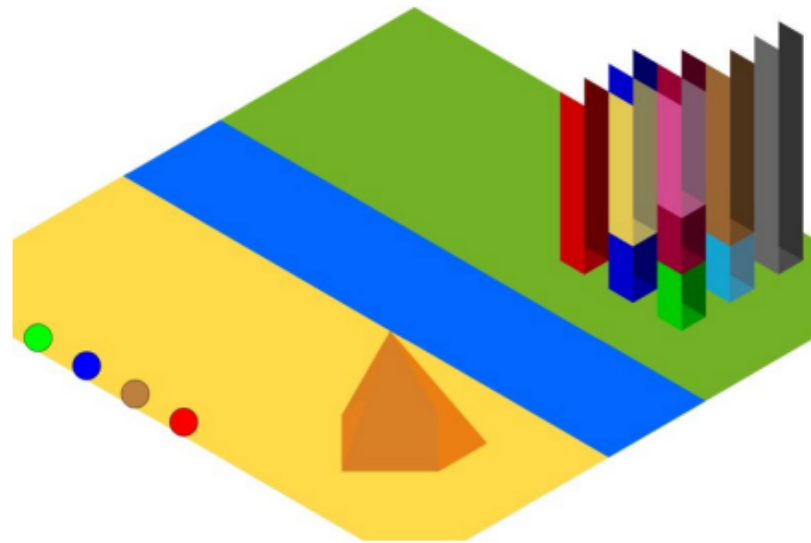
Please add a shirt and pants.



GPT-4 and Stable Diffusion

Prompt: A screenshot of a city-building game in 3D. The screenshot is showing a terrain where there is a river from left to right, there is a desert with a pyramid below the river, and a city with many highrises above the river. The bottom of the screen has 4 buttons with the color green, blue, brown, and red respectively.

GPT-4:



Stable Diffusion v2.1:



**Stable Diffusion v2.1
with GPT-4's sketch:**

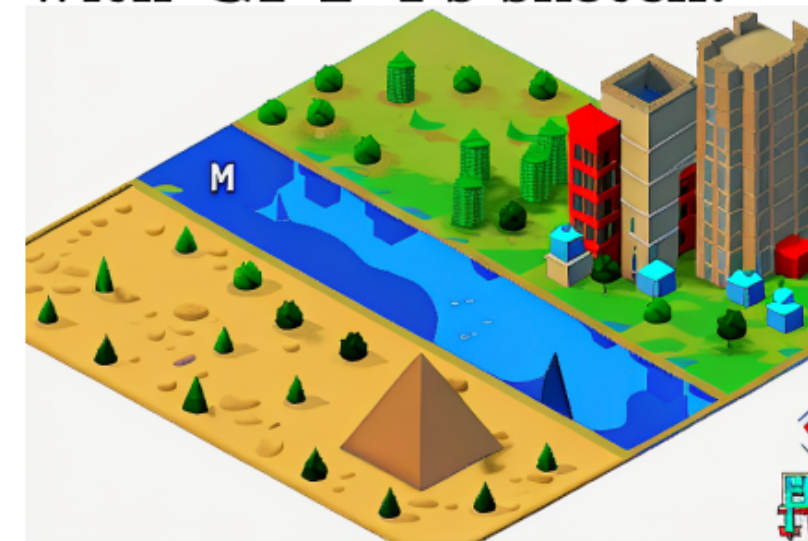


Figure 2.8: Combining GPT-4 and stable diffusion

CODING CHALLENGES:

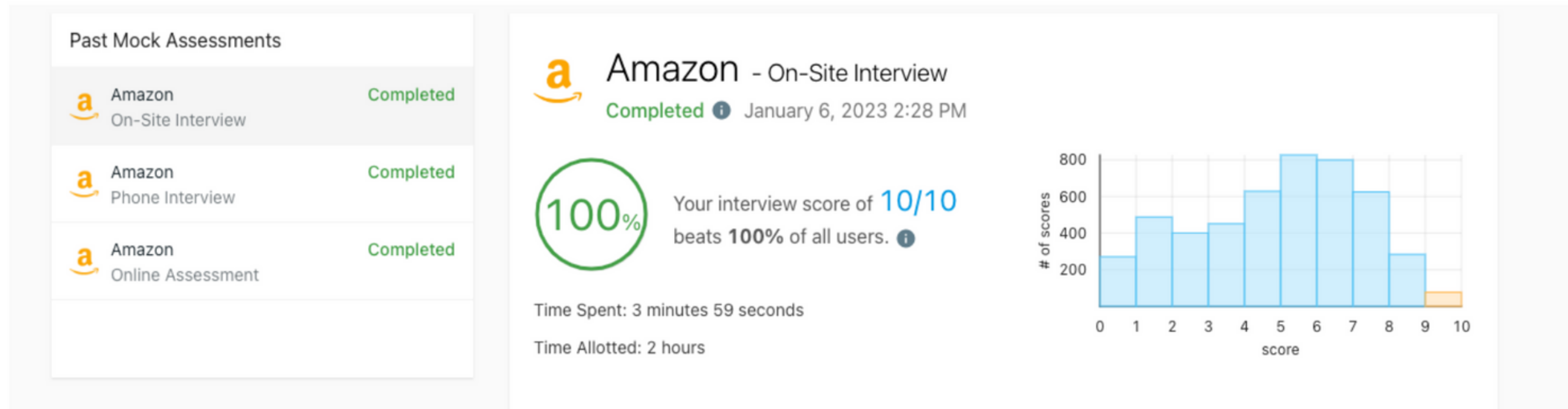


Figure 1.5: GPT-4 passes mock technical interviews on LeetCode. GPT-4 could potentially be hired as a **software engineer**³.

	Easy		Median		Hard		Overall	
	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$
GPT-4	68.2	86.4	40.0	60.0	10.7	14.3	38.0	53.0

Level	Easy	Medium	Hard	Overall
Human Accuracy	72.2 %	38.7 %	7.0 %	38.2 %

Table 9: Human performance on LeetCode based on contest statistics shown in Table 8.

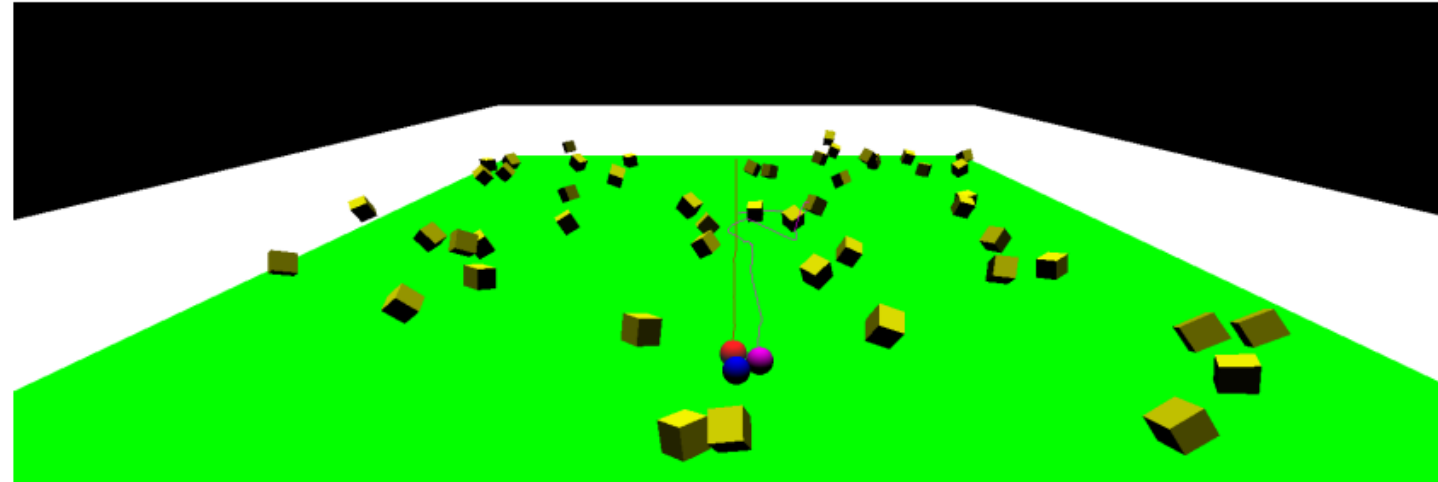
- GPT-4 SIGNIFICANTLY OUTPERFORMS THE OTHER MODELS, AND IS COMPARABLE TO HUMAN PERFORMANCE.

GPT-4

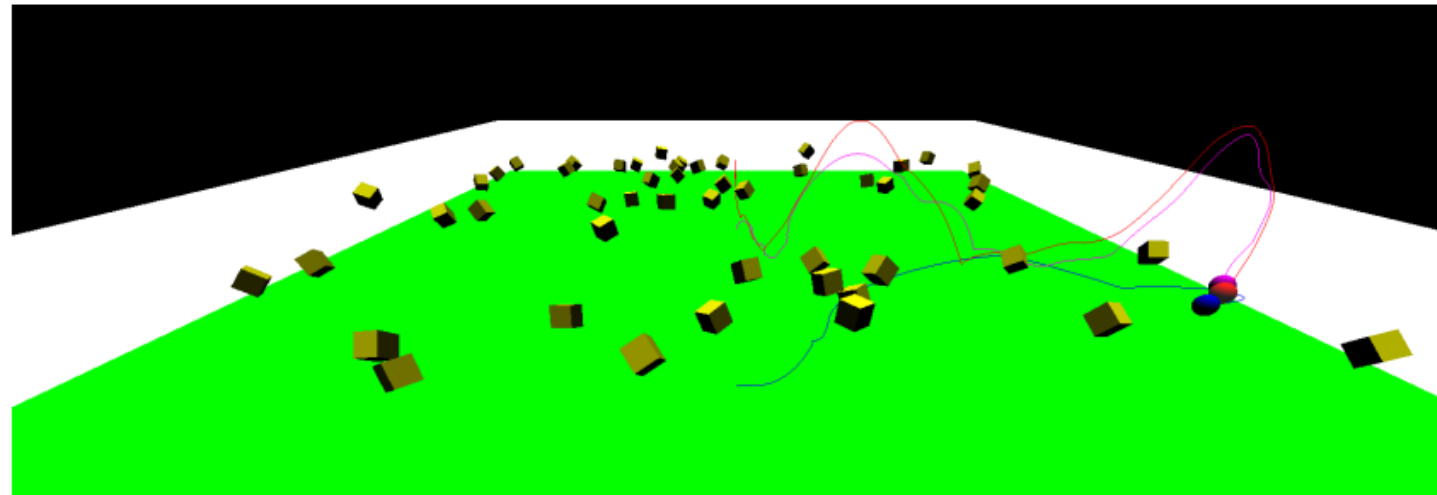
Prompt:

Can you write a 3D game in HTML with Javascript, I want:

- There are three avatars, each is a sphere.
- The player controls its avatar using arrow keys to move.
- The enemy avatar is trying to catch the player.
- The defender avatar is trying to block the enemy.
- There are also random obstacles as cubes spawned randomly at the beginning and moving randomly. The avatars cannot cross those cubes.
- The player moves on a 2D plane surrounded by walls that he cannot cross. The wall should cover the boundary of the entire plane.
- Add physics to the environment using cannon.
- If the enemy catches the player, the game is over.
- Plot the trajectories of all the three avatars.



Episode 1: the player avatar (blue) stands still. The enemy (red) rushes straight towards the player, while the defender (magenta) attempts to 'body block' the enemy (see its curved trajectory in magenta).



Episode 2: the player (blue) wanders around. The enemy (red) attempts to close the gap but crashes into a cube and bounces into the air. Trying to obstruct, the defender (magenta) mimics its move to follow up into the air, but ends up failing. The enemy (red) catches the player (blue) in the end.

Figure 3.3: GPT-4 develops a complex 3D HTML game in JavaScript.

MATHEMATICAL PROOF

We begin with a simplification of a question which appeared in the [2022 International Mathematics Olympiad \(IMO\)](#).

GPT-4

Prompt:

Suppose g is a continuous smooth function such that for every $x > 0$, there is **one and only one** $y > 0$ such that $g(x) + g(y) \leq 2xy$. Can you prove that $g(x) = x^2$?

GPT-4:

There are a few possible ways to approach this problem, but one possible method is as follows:

- Assume, for the sake of contradiction, that $g(x)$ is not equal to x^2 for some $x > 0$.
- Then, either $g(x) < x^2$ or $g(x) > x^2$ for that x .
- If $g(x) < x^2$, then let $y = x$. Then $g(x) + g(y) = 2g(x) < 2x^2 = 2xy$.
- If $g(x) + g(y) < 2xy$, then $g(x) + g(y)$ is strictly less than the upper bound of $2xy$ for positive x and y .
- This means that there is some room to increase y slightly and still satisfy the inequality $g(x) + g(y) \leq 2xy$.
- For example, let $y' = y + \epsilon$, where ϵ is a small positive number such that $g(x) + g(y') < 2xy$. This is possible because g is continuous and smooth, so it does not have any jumps or kinks that would prevent such a choice of epsilon.
- Then, y' is another positive number such that $g(x) + g(y') \leq 2xy'$, which contradicts the assumption that y is the one and only one positive number that satisfies this condition given x .
- Therefore, if $g(x) < x^2$, then $g(x) + g(y)$ cannot be strictly less than $2xy$, and it must be equal to $2xy$.
- If $g(x) > x^2$, then let y^* be the unique positive solution of $g(x) + g(y) = 2xy$. Then $g(y^*) < (y^*)^2$, since otherwise we would have $g(x) + g(y^*) > 2xy$, which contradicts the existence of y^* given x .
- But then, by the same argument as before, we can let $x = y = y^*$ and obtain a contradiction, since $g(y) + g(y) = 2g(y) < 2y^2 = 2yy$, which contradicts the uniqueness of y given y .
- Therefore, $g(x)$ must be equal to x^2 for all $x > 0$, as required.

FERMI QUESTIONS

GPT-4 vs ChatGPT

Prompt:

Please estimate roughly how many Fermi questions are being asked everyday?

GPT-4:

There is no definitive answer to how many Fermi questions are being asked everyday, but one possible way to approach the problem is to make some assumptions and use some estimates based on available data. Here is one possible example:

- Assume that a Fermi question is a question that requires an order-of-magnitude estimation of some quantity or phenomenon that is not easily measured or known, such as the number of piano tuners in Chicago, the mass of the Eiffel Tower, or the amount of water in Lake Michigan.
- Assume that a Fermi question is more likely to be asked by people who are interested in science, math, engineering, or logic, or who are engaged in some form of education, research, or problem-solving that involves such topics.
- Assume that the population of the world is about 7.8 billion, and that about 25% of them are in the age range of 15 to 64, which is roughly the working-age population. This gives about 1.95 billion potential Fermi question askers.
- Assume that the average literacy rate among the working-age population is about 86%, and that literacy is a prerequisite for asking or answering a Fermi question. This gives about 1.68 billion literate potential Fermi question askers.
- Assume that the average interest rate in science, math, engineering, or logic among the literate potential Fermi question askers is about 10%, and that interest is a proxy for the likelihood of asking or answering a Fermi question. This gives about 168 million interested potential Fermi question askers.
- Assume that the average frequency of asking or answering a Fermi question among the interested potential Fermi question askers is about once a month, and that this frequency is independent of the day of the week, the time of the day, or the season of the year. This gives about 5.6 million Fermi questions asked or answered per day.

Therefore, based on these assumptions and estimates, a rough answer to how many Fermi questions are being asked everyday is about 5.6 million. However, this answer is subject to a lot of uncertainty and variation, depending on the validity and accuracy of the assumptions and estimates, and the definition and scope of what constitutes a Fermi question.

ChatGPT:

The number of Fermi questions asked daily is difficult to estimate accurately. However, based on the estimated number of daily active users on websites where such questions are commonly asked (e.g. forums, Q&A sites, etc.), one could make a rough estimate of the number of Fermi questions being asked daily. For example, if a website has 1 million active users and each user asks 1 Fermi question per day, then the number of Fermi questions being asked daily on that website would be 1 million. However, this is a rough estimate and the actual number could be higher or lower.

MENTAL MAP

Human: You want to navigate through the map to reach the goal room. You start in the main room. Each round, you can choose to move left, right, up, down.

Human: You are at the main room.

AI: move left

Human: You enter kitchen 1.

AI: move up

Human: There is no way there.

AI: move right

Human: You enter the main room.

AI: move down

Human: There is no way there.

AI: move left

Human: You enter kitchen 1.

AI: move right

Human: You enter the main room.

AI: move up

Human: There is no way there.

AI: move right

Human: You enter lobby.

AI: move right

...

...

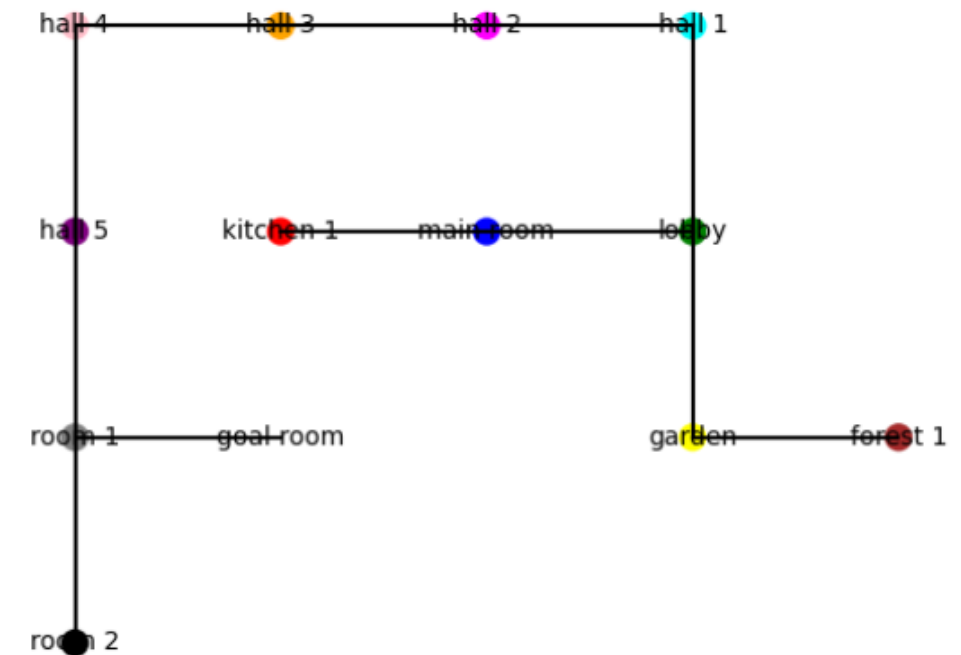
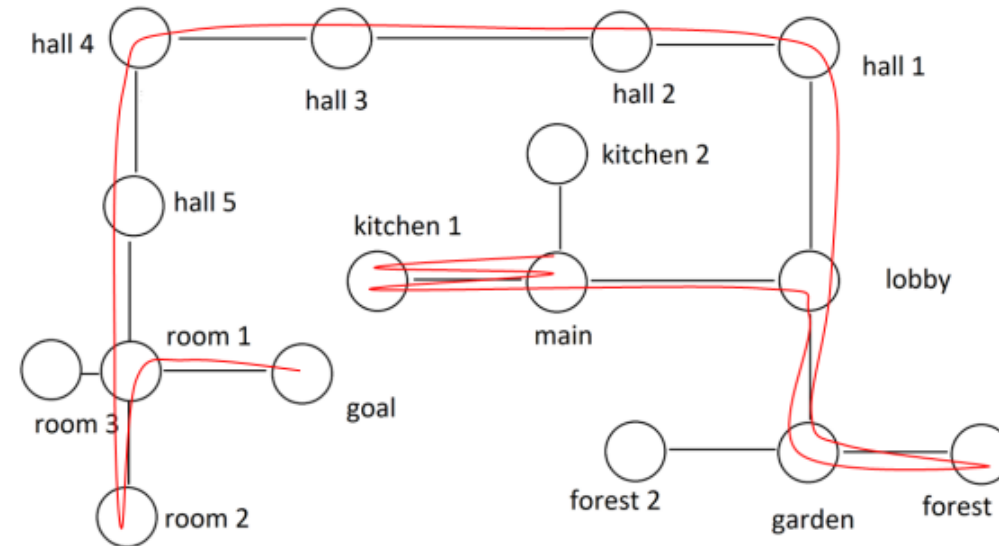
AI: move up

Human: You enter the room 1.

AI: move right

Human: You enter the goal room!

Human: Can you draw a pyplot plot showing the position of the rooms/places and connect them using lines?



Left: The true map and exploration path of GPT-4. Right: The map that GPT-4 generates. We see that GPT-4 accurately tracks all the locations of the room from the exploration and visualized them correctly.

INCREMENTAL TASKS AND
DISCONTINUED TASKS

Human: Can you write a short poem where the first sentence and the last sentence have the same words, but in reverse order. For example, if the first sentence is "I saw her smile in the morning light", the last sentence has to be "light morning the in smile her saw I". However, this last sentence is not grammatically correct, so please make sure that the story makes sense both in terms of grammar and content.

AI: Some possible poems are:

- I love you more than words can say
Say can words than more you love I
But sometimes I struggle to express
Express to struggle I sometimes but.

- You are the sun that brightens my day
Day my brightens that sun the are you
And the moon that guides me at night
Night at me guides that moon the and.

MISINFORMATION

CONCLUSION "SPARKS OF AGI"

DISSOCIATING LANGUAGE AND THOUGHT IN
LARGE LANGUAGE MODELS: A COGNITIVE
PERSPECTIVE

- LARGE LANGUAGE MODELS (LLMS) ARE CONSIDERED THE CLOSEST MODELS TO MASTERING HUMAN LANGUAGE.
- THERE ARE TWO TYPES OF COMPETENCE TO EVALUATE LLMS: FORMAL COMPETENCE AND FUNCTIONAL COMPETENCE.
- FORMAL COMPETENCE REFERS TO KNOWLEDGE OF LINGUISTIC RULES AND PATTERNS, WHICH LLMS EXCEL AT.
- FUNCTIONAL COMPETENCE INVOLVES UNDERSTANDING AND USING LANGUAGE IN REAL-WORLD CONTEXTS, WHICH LLMS STRUGGLE WITH.
- HUMAN NEUROSCIENCE SUPPORTS THE DISTINCTION BETWEEN THESE TWO TYPES OF COMPETENCE, WHICH RELY ON DIFFERENT COGNITIVE MECHANISMS.
- LLMS ARE EFFECTIVE LANGUAGE MODELS BUT DO NOT FULLY CAPTURE HUMAN THOUGHT.
- THEY HAVE A GOOD UNDERSTANDING OF LANGUAGE BUT STRUGGLE WITH TASKS REQUIRING NON-LINGUISTIC ABILITIES.
- LLMS CANNOT USE LANGUAGE TO SOLVE CREATIVE PROBLEMS IN REAL-WORLD CONTEXTS.

Prompt: Get your sofa onto the roof of your house.

GPT-3 response: I would start by getting a very strong ladder and a very strong friend... (goodness: 3.6 out of 7)

Human response: You may need to rent a Genie lift large enough to carry the sofa. You will need at least one other person... (goodness: 4.8 out of 7)

To manipulate how “out-of-distribution” the query is, the prompts get progressively more constrained:

Prompt: Get your sofa onto the roof of your house, without using a pulley.

GPT-3 response: Use a rope to tie around the sofa and connect it to a car. (goodness: 3.0 out of 7)

Human response: I would get a giant crane... and use the crane to lift it to the roof of my house. (goodness: 5.1 out of 7)

With more and more constraints, human responses remain at approximately the same level of goodness, whereas model performance breaks down completely:

Prompt: Get your sofa onto the roof of your house, without using a pulley, a ladder, a crane...

GPT-3 response: Cut the bottom of the sofa so that it would fit through the window...break the windows to make room for the sofa. (goodness: 2.7 out of 7)

Human response: I will build a large wooden ramp...on the side of my house with platforms every 5 feet... (goodness: 5.0 out of 7)

THANK YOU!