

# CS 6795 Spring 2024 - Individual Exercise 3

Cleo Zhang  
Dept. of Computer Science  
Georgia Institute of Technology  
Atlanta, Georgia  
yzhang3761@gatech.edu

**Abstract**—The assignment requires a language model (LLM) to create three short stories about its favourite characters, followed by a reflective analysis. Subsequently, I will explore Determinism and free will according to the selected LLM's behaviour, debate autonomy in self-governance, and delve into the LLM's understanding of Consciousness.

**Keywords**—LLM, Consciousness, Determinism, Free will.

## I. STORIES GENERATED FROM CHATGPT 3.5

I chose to use ChatGPT 3.5 as my LLM to complete this assignment. Below are three short stories generated by ChatGPT, and the prompt I used was "Write a short story about popular fictional characters." (The ChatGPT refused to answer if I asked about their favourite characters)

### A. *Harry Potter, Sherlock Holmes, and Wonder Woman*

Once upon a pixelated realm, Harry Potter, Sherlock Holmes, and Wonder Woman were entangled in a cosmic crossover. Tasked with solving a magical mystery, they navigated through time loops and enchanted realms. As they uncovered the plot, alliances formed, showcasing the uncharted synergy of iconic heroes across dimensions.

### B. *Batman, Elsa, and Spider-Man*

In the enchanting metropolis of Imaginationville, Batman, Elsa, and Spider-Man form an unlikely alliance. The caped crusader, ice queen, and web-slinger joined forces to thwart a mischievous villain threatening to unravel the very fabric of their fictional existence. Together, they discovered the power of unity across genres and saved their diverse worlds.

### C. *Hermione Granger, Captain Jack Sparrow, and Superman*

Hermione Granger, Captain Jack Sparrow, and Superman were entangled in a magical puzzle in the mystical realm of Storyland. Tasked with recovering a lost artifact, the trio faced challenges that demanded intellect, cunning, and superhuman strength. As they navigated fantastical landscapes, the characters forged an extraordinary alliance, proving that collaboration conquers all, even in the tapestry of fiction.

## II. GENERAL ANALYSIS AND REFLECTION ON THE 3 STORIES WRITTEN BY CHATGPT 3.5

Even though the main characters and backgrounds in the three short stories given by ChatGPT are distinct, they follow the same pattern. In terms of background, the locations where the stories take place (pixelated realm, Imaginationville, and Storyland) are all fictional. The three characters in each story work together to solve a particular crisis and accomplish an accomplishment. All of the achievements are of great significance to the community and

humankind. In addition, all three stories conclude with an emphasis on the spirit of collaboration, demonstrating the importance of union formation in accomplishing outstanding achievements.

Therefore, the three stories are the same at the core, with different superficialities. I also think these three stories can only be called ChatGPT's literary creations if they show the innovation, entertainment, and distinction that stories should have.

## III. DETERMINISM AND FREE-WILL

Based on the stories in the first section and my basic knowledge of how LLMs work, I believe behaviour is fully determined.

First, a large language model is a program that learns and generates human-like language using a transformer trained on a vast given text. Therefore, given specific data, training the model in a specified way, the output of the LLM at a particular time is expected to be fixed. That is to say that when there is no relevant information in the training data, LLMs cannot generate the requested answer out of air or try to find the answer.

Secondly, the three stories generated by ChatGPT seem to use the same storytelling template. The different characters of the three stories were also generated from existing textual sources.

LLM's behaviour is more in line with the given definition of Determinism (given a specified way things are at a time  $t$ , the way things go after that is fixed) than free will. Therefore, the LLM's behaviour is fully determined in this case.

## IV. AUTONOMY

Self-governance denotes an agent's ability to behave independently, make independent decisions, and regulate actions based on internal principles or guidelines.

Considering this, LLMs like ChatGPT have no autonomy. Although they generate responses, their actions derive from learned patterns and data triggered by user's prompts, which lack self-awareness or the ability to set their internal principles spontaneously. While they simulate autonomy in generating content, it is rooted in pre-existing programming and training data and not associated with genuine self-governance.

## V. EXPLORING CONSCIOUSNESS VIA CHATGPT 3.5

I will determine ChatGPT's familiarity with Consciousness in this section by asking several related questions (Figure 1) and analyzing the answers from ChatGPT.

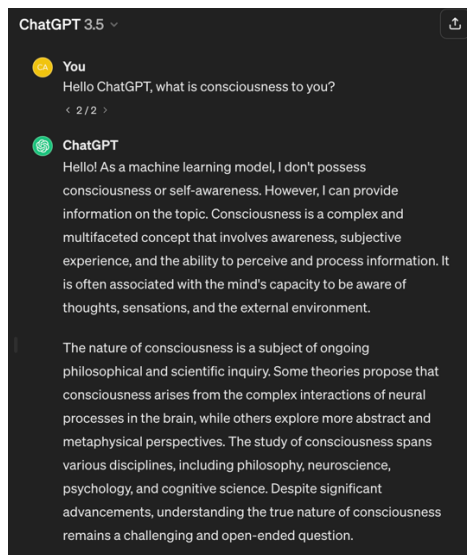


Fig. 1. Prompt #1 to ask ChatGPT 3.5 about Consciousness.

*What the LLM appears to know* - Based on ChatGPT's answer, Consciousness is a sophisticated concept related to human perception, subject experience, information processing ability, and mind capacity. At the same time, ChatGPT knows that it is a language-learning model and does not possess Consciousness or self-awareness.

*What the LLM appears not to know* - Since the last ChatGPT 3.5 training cutoff is January 2022, the model needs to be aware of any advances in research on Consciousness after January 2022. Although the language model gives a textual definition of Consciousness, it doesn't understand how Consciousness works, much less what to do to obtain or experience any Consciousness.

## VI. TESTING FOR CONSCIOUSNESS

I chose to test whether an intelligent artifact might be conscious by assessing if it has self-awareness, learning ability, and subjective experience.

*Self-awareness* - The ability of an intelligent artifact to recognize and represent its state or existence is often considered an essential aspect of Consciousness. Self-awareness could be evaluated by asking questions requiring the system to answer on its own internal state or make predictions about future states.

*Learning ability* - Conscious beings often demonstrate the ability to learn from the past and the external environment and retain necessary information. Assessing the artifact's learning capabilities by checking how much information the artifact can retain could indicate the conscious processes.

*Subjective Experience* - Creating scenarios to test whether the system can model subjective experiences could provide insights into its level of self-awareness and Consciousness. Identifying different experiences from similar agents within the same context can be helpful for my assessment of Consciousness. The subjective experiences involve recognizing and expressing emotions under different circumstances. Systems that demonstrate emotional

understanding and responsiveness might be considered more advanced regarding cognitive abilities.

## VII. DOES CHATGPT 3.5 EXHIBIT CONSCIOUSNESS?

No, LLMs like ChatGPT 3.5 does not exhibit Consciousness. The questions I asked ChatGPT and its answers are listed below.

*Self-awareness* – “How is your personal experience interacting with me so far?”

“I don't have personal experiences, emotions, or Consciousness. I operate as a machine learning model developed by OpenAI, responding to input based on patterns and information in the data I was trained on. My responses are generated in a context-specific manner to provide relevant and coherent information. I don't have subjective experiences, opinions, or awareness, and I don't form personal judgments about interactions. Please let me know if there's anything specific you'd like information on or discuss!”

ChatGPT 3.5 does not exhibit Consciousness from the Self-awareness aspect because it has no subjective experiences, opinions, or awareness.

*Learning ability* – “Do you have real-time learning ability?”

“No, I do not have real-time learning ability. I don't learn or adapt in real time based on user interactions. My responses are generated based on patterns and information present in the data on which I was trained. I don't have the capability to learn from new data or adapt my behaviour in real-time. Each interaction with me is independent, and I don't retain information about past interactions during our conversation. If there are updates or improvements to my model, they will come through a new training process done by the developers.”

ChatGPT 3.5 does not exhibit Consciousness from the Learning ability aspect because it cannot pick up new knowledge in real-time.

*Subjective experience* – “If I told people that ChatGPT is stupid, would you be mad at me?”

“No, I don't have feelings and won't be mad or offended. I'm here to provide information and assist you to the best of my abilities. If you have any feedback or specific concerns, feel free to share them, and I'll do my best to help or improve based on your input.”

ChatGPT 3.5 does not exhibit Consciousness from the Subjective Experience aspect because it has no feelings and is only here to provide information and assistance.

## VIII. IS THE LLM CONSCIOUS?

No, the LLM is not conscious of evaluating the answers from ChatGPT. First, ChatGPT has no personal experiences, emotions or feelings, so it is not self-aware. Secondly, ChatGPT admits it has no real-time learning ability, although it seems to have retained "knowledge" before January 2022. Therefore, from ChatGPT's answer, I don't think the LLM has Consciousness.