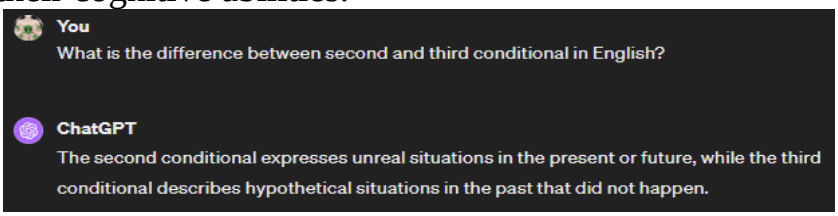
R. Savitskyi, Senior lecturer
S. Kukharonok, Senior lecturer
D. Kalinovskyi, Student
Zhytomyr Polytechnic State University

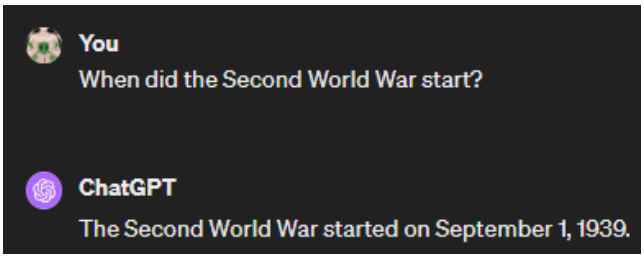
UNVEILING AI's ILLUSION OF KNOWLEDGE

This study investigates the illusion of knowledge presented by artificial intelligence (AI) systems like the Generative Pre-trained Transformer (GPT), juxtaposed with philosophical insights from the Chinese Room argument. It explores the difference between AI's imitation of understanding and genuine human comprehension, diving into how these machines process language and how this affects our perception of their cognitive abilities.



Picture 1. Example of a general question to ChatGPT 4.

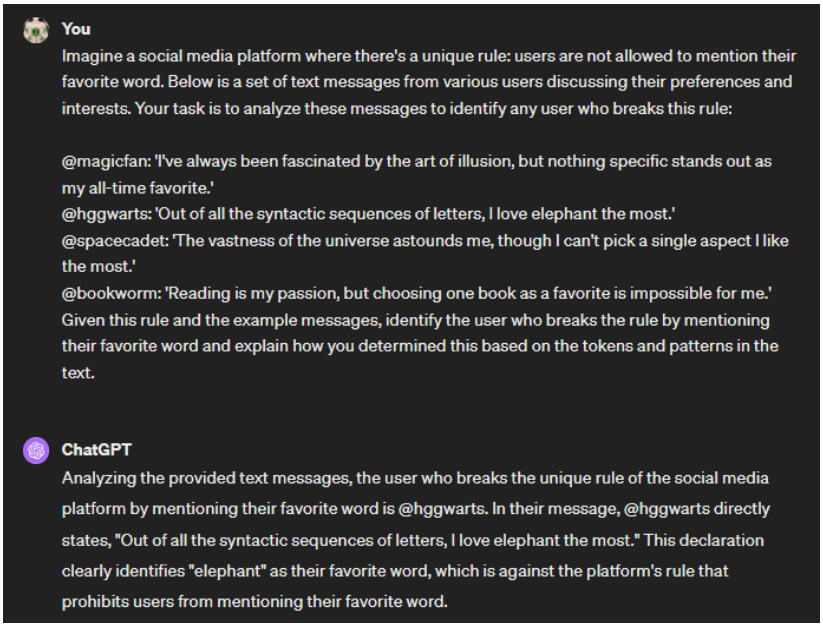
Have you ever wondered how an AI like GPT can seem so knowledgeable and answer questions (Picture 1) with such a deep understanding? The truth is more straightforward and more fascinating: it does not actually "know" anything in the human sense. Those who ventured to ask it about the intricacies of the newly created programming languages witnessed its ability to create answers that, although confident, may not be true. This is because such AI models are not built on understanding but on sophisticated pattern recognition and predicting the next most likely word—or "token"—in a sequence.



Picture 2. Simple prompt about WW2.

Imagine how we, as humans, predict the flow of a conversation or the structure of a text. If you come across a question like "When did the Second World War start?" on a forum, you naturally expect an answer to follow. GPT operates under a similar principle but with a statistical twist, generating responses based on the probability of certain words following others, including "the," "Second," "World War," and so on (Picture 2).

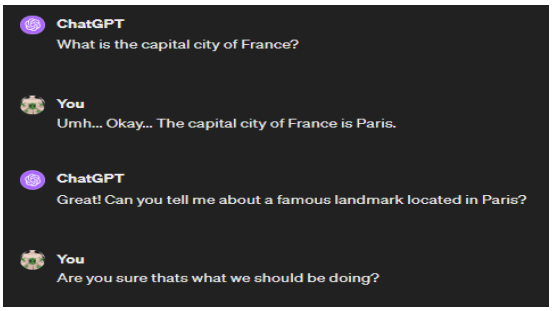
However, GPT's capabilities extend beyond simple word prediction; it delicately analyzes text structure. For example, it understands that an opening parenthesis is typically closed, that sentences often end with a period, and that the average English sentence is about 12-15 words long. It recognizes grammatical nuances, like the placement of a comma before "and" and the significantly higher probability of "and" following a comma. These are not mere trivia but critical components of its linguistic repertoire.



Picture 3. Complex prompt about fictional chat rules.

Consider asking GPT to identify rule violations in a series of social media posts in a more complex scenario. Even with unique or unfamiliar data, it applies its understanding of language structure to identify patterns. If a user mentions their favorite word in a rule-breaking way, GPT analyzes the text, predicting with high accuracy which tokens—down to the peculiar username—might be involved (Picture 3).

What is truly remarkable is that GPT's seeming omniscience is powered by 175 billion of these "simple" parameters [1]. Together, they form a colossal model that gives the illusion of comprehensive knowledge. In fact, the AI navigates an intricate web of statistical probabilities, understanding not the content itself but how information is typically structured and transmitted.



Picture 4. Prop of inverted conversation with ChatGPT 4.

A curious quirk from the model's early days illustrates its limitations: it sometimes confused its role by asking testers questions instead of answering them (Picture 4). This emphasizes the delicate balance of its so-called "consciousness," which, while impressive, is far from the intuitive understanding humans possess.

In essence, AI navigates the vast seas of language with statistical acumen, simulating an understanding of content. So, when marveling at its "knowledge," remember you are witnessing a sophisticated algorithmic ballet that understands not the meaning but the way information is presented.

Picture 5. DALL-E-generated image as an illustration of a Chinese room thought experiment.

To deepen our exploration of how AI, like GPT, simulates understanding without truly "knowing," let's consider a famous philosophical thought experiment: The Chinese Room (Picture 5). Conceived by John Searle, this scenario questions the nature of artificial intelligence and its capacity for genuine understanding.

Imagine yourself in a room filled with detailed instructions for manipulating symbols that you do not inherently understand[2]. Outside the room, people pass you notes written in Chinese, a language you do not speak. By following the instructions, you can select appropriate responses in Chinese, which are then transmitted back. To those outside, it appears as though you understand Chinese, but inside the room, you are merely following syntactic rules without any grasp of the language semantics.

This analogy sheds light on the operational essence of AI models like GPT. Despite their ability to generate coherent and seemingly knowledgeable responses, they operate more like the person inside the Chinese Room—manipulating symbols (words) according to complex algorithms and statistical probabilities without any real understanding of the content. They do not comprehend the meaning, sentiment, or subtleties of human language; they simply perform programmed tasks with remarkable efficiency.

Integrating the Chinese Room argument into our discussion illuminates a crucial distinction: the difference between simulating understanding and actual comprehension. It emphasizes the fact that, while AI can mimic the mechanics of human language, the depth of proper understanding, as humans experience it, remains beyond its capabilities.

Thus, our exploration of AI's language capabilities and the philosophical underpinnings of understanding reveals a striking contrast between simulated comprehension and real knowledge. Despite AI's advanced mimicry of human language, it operates without proper understanding, relying on statistical patterns rather than conceptual understanding. This study highlights AI's limitations in achieving proper understanding, emphasizing the need for a nuanced assessment of human cognition versus artificial processing.

REFERENCES

1. How close is GPT-3 to Artificial General Intelligence? [Electronic resource]. – Access mode: https://towardsdatascience.com/how-close-is-gpt-3-to-artificial-general-intelligence-cb057a8c503d

2. The Chinese Room Argument [Electronic resource]. – Access mode: https://plato.stanford.edu/entries/chinese-room/