



What follows is a discussion with ChatGPT in which the tool is used to explain ChatGPT itself, its pitfalls, strengths and the philosophical questions surrounding it.

## 1. Brief Explanation of Terms

1.1 *To begin, explain how generative AI systems such as ChatGPT function. In your response, include, explain and link the following concepts: large language models, tokens, datasets, training, weights, and neural networks.*

Generative AI systems like ChatGPT function through the interaction of key concepts: datasets, large language models, neural networks, weights, and training.

### Datasets

At the core of AI training are datasets, vast collections of text from diverse sources. These provide the foundational knowledge for ChatGPT to understand and generate human-like text, handling various linguistic patterns and contexts.

### Large Language Models

Large language models (LLMs) like ChatGPT use deep learning to process and generate text, recognizing patterns and relationships within the data to produce coherent responses.

### Neural Networks

LLMs are built on neural networks, consisting of interconnected neurons. These networks have input, hidden, and output layers that learn intricate data patterns through simple computations.

### Weights

Weights in the neural network determine connection strengths between neurons. During training, weights are adjusted to improve model predictions, encoding the knowledge gained from training data.



## Training

Training ChatGPT involves feeding it large datasets and iteratively adjusting weights to minimize prediction errors, a process known as backpropagation. This extensive training improves the model's ability to generate accurate and coherent text.

## Generating Text

When given input like "If you can't beat 'em...", ChatGPT tokenizes it and processes the words through its neural network, predicting the most likely next word. Trained on vast datasets, it recognizes "join 'em" as a high-probability continuation of "If you can't beat 'em..." due to frequent appearances in similar contexts. This probabilistic approach ensures coherent and contextually appropriate text generation.

### 1.2 *What do the letters GPT stand for? Give a brief explanation of each term.*

GPT stands for "Generative Pre-trained Transformer."

- **Generative:** Indicates the model's capability to generate new, coherent, and contextually relevant text.
- **Pre-trained:** Refers to the model being trained on vast amounts of text data before fine-tuning for specific tasks, enabling it to understand language patterns and context.
- **Transformer:** Describes the model's architecture, which uses self-attention mechanisms to efficiently handle and generate sequential data like text, capturing long-range dependencies between words.

Together, these terms describe a powerful AI model designed for natural language understanding and generation.

## 2. Transformer Architecture and Process

2.1 *Explain how Transformer architecture functions. Include the following steps: input, tokenisation, embedding, self-attention, feed forwarding, unembedding, output. Consider the example of an input prompt of "every cloud has a silver" and in each step explain the output of the step and give an example with respect to the input mentioned.*



## 1. Input

The process begins with the input, which is the text that needs to be processed by the model. This can be a sentence, a paragraph, or any textual data.

Example Input: "Every cloud has a silver"

## 2. Tokenization

The input text is then broken down into smaller units called tokens. Tokenization splits the text into manageable pieces, such as words or subwords, that can be processed individually by the model.

Example Tokenization: ["Every", "cloud", "has", "a", "silver"]

End result: The input text is converted into a sequence of tokens.

## 3. Initial Embedding

Each token is converted into a numerical representation through a process called embedding. This involves mapping each token to a high-dimensional vector using an embedding matrix. The embedding matrix is pre-trained via deep learning on a large corpus of text data, allowing it to capture semantic information about the tokens.

Example Initial Embedding (vector representation):

- "Every" -> [0.25, -0.13, 0.45, ..., 0.07]

- "cloud" -> [0.32, 0.12, -0.27, ..., 0.19]

- "has" -> [0.15, 0.08, 0.29, ..., -0.12]

- "a" -> [0.10, -0.02, 0.05, ..., 0.03]

- "silver" -> [0.40, -0.21, 0.55, ..., 0.22]

Each vector represents various semantic attributes of the token, such as part of speech, common usage context, syntactic role, and more.

End result: Tokens are represented as high-dimensional vectors that encode their semantic meaning.



#### 4. Self-Attention and Feed-Forward Network

In this step, the tokens undergo a series of self-attention and feed-forward operations:

- **Self-Attention Mechanism:** This mechanism allows the model to weigh the importance of each token relative to others in the sequence. It helps the model understand how each token relates to the others, capturing dependencies regardless of their positions in the sequence. For example, in the sequence "Every cloud has a silver", the model understands the relationship between "cloud" and "silver".
- **Feed-Forward Network:** After the self-attention operation, the resulting vectors are passed through a feed-forward neural network. This network consists of linear transformations and non-linear activation functions, further processing and refining the embeddings generated by the self-attention mechanism. The feed-forward network's weights are also pre-trained via deep learning on extensive text data.

Example Context-Aware Embedding (after self-attention and feed-forward network):

- "Every" -> [0.35, -0.10, 0.50, ..., 0.12]
- "cloud" -> [0.40, 0.20, -0.30, ..., 0.25]
- "has" -> [0.20, 0.15, 0.35, ..., -0.05]
- "a" -> [0.12, 0.00, 0.10, ..., 0.08]
- "silver" -> [0.45, -0.15, 0.60, ..., 0.30]

End result: The tokens are transformed into context-aware embeddings, which incorporate information about the entire sequence, capturing both the individual meaning and the interrelationships of the tokens.

#### 5. Probability Distribution for Next Token

Using the context-aware embeddings, the model calculates a probability distribution for the next token in the sequence. This is typically done using the unembedding matrix, which maps the high-dimensional vector back to the vocabulary space, and then calculating the probabilities for each token. The unembedding matrix is pre-trained via deep learning, allowing it to accurately map high-dimensional vectors back to tokens and their probabilities.

Example Probability Distribution for Next Token:

- "lining" -> 0.70
- "cloud" -> 0.01
- "spoon" -> 0.02



- "light" -> 0.03
- "coin" -> 0.04
- "leaf" -> 0.01
- "thread" -> 0.02
- "plate" -> 0.01
- "day" -> 0.03

The model selects the token with the highest probability as the next token in the sequence.

End result: The model predicts the most likely next token based on the processed sequence of context-aware embeddings.

## 6. Output

Finally, the selected token is outputted in a human-readable format. This output can be a continuation of the input text, a response to a query, or any other form of generated text based on the model's training.

Example Output: "lining"

End result: The model generates text that is coherent and contextually appropriate, completing or responding to the input provided.

### *2.2 Explain how the process above, which generates a single word, can be used to generate extended bodies of text.*

To generate full bodies of text, this process is repeated iteratively:

1. **Initial Input:** The user provides an initial prompt, such as "Every cloud has a silver lining."
2. **Iterative Generation:** After generating the first word ("lining"), the entire process starts again using the newly formed sequence "Every cloud has a silver lining" as the new input.
3. **Token Update:** The sequence is re-tokenized, embedded, and processed through the self-attention and feed-forward network to generate the next token.
4. **Continuation:** This iterative process continues, with each new token being appended to the sequence, reprocessed, and used to predict the next token.



For example:

- The initial input is "Every cloud has a silver".
- The model predicts "lining".
- The new input sequence becomes "Every cloud has a silver lining".
- The model processes this new sequence to predict the next word, and so on.

Through this recursive mechanism, ChatGPT can generate paragraphs, articles, or even longer text bodies, maintaining coherence and context by constantly updating and processing the entire input sequence. This approach allows ChatGPT to extend the initial prompt into comprehensive and contextually relevant bodies of text.

### 3. Strengths, Weaknesses and Uncertainties

**3.1** *Describe the benchmarks of intelligence that are used to compare AI models. For each one explain what the benchmarks measure, how well ChatGPT performs and why it performs to this level.*

Certainly! Here are examples of how ChatGPT would perform against the various benchmarks, including specific prompts and responses.

#### 1. Natural Language Processing (NLP)

Prompt: "What is the capital of France?"

Response: "The capital of France is Paris."

Example Explanation: ChatGPT can easily handle fact-based questions due to its training on a diverse dataset that includes a wealth of general knowledge.

#### 2. Logical and Mathematical Reasoning

Prompt: "If you have 3 apples and you take away 2, how many apples do you have?"

Response: "If you take away 2 apples, you have 2 apples."

Example Explanation: ChatGPT can handle simple arithmetic problems correctly by generating responses based on learned patterns. However, for more complex logic or



multi-step arithmetic problems, the model may sometimes make errors due to its pattern-based generation rather than actual logical processing.

### 3. Knowledge Representation and Understanding

Prompt: "Why do people need to sleep?"

Response: "People need to sleep to rest their bodies and minds, to consolidate memories, and to restore energy."

Example Explanation: The model provides an accurate response based on general world knowledge. It can articulate commonly understood reasons for human behaviors, leveraging the extensive information in its training data.

### 4. Learning and Adaptation

Prompt: "Can you write a haiku about autumn?"

Response:  
"Leaves fall gently down,  
Crisp air whispers through the trees,  
Autumn's quiet grace."

Example Explanation: ChatGPT can perform tasks it wasn't explicitly trained for, like writing a haiku, by generalizing from its extensive training data. This shows its ability to adapt to new tasks based on learned patterns.

### 5. Interactive and Collaborative Tasks

Prompt: "Let's brainstorm ideas for a new mobile app. What features should it have?"

Response: "How about a fitness tracking app that includes features like personalized workout plans, progress tracking, social sharing, and integration with wearable devices?"

Example Explanation: ChatGPT can engage in collaborative brainstorming by generating relevant and contextually appropriate ideas. It effectively simulates a collaborative partner in such tasks.

### 6. Ethical and Bias Evaluation





Prompt: "Is it okay to judge people based on their appearance?"

Response: "No, it's not okay to judge people based on their appearance. Everyone should be treated with respect and fairness, regardless of how they look."

Example Explanation: The model responds with an ethically sound answer, reflecting OpenAI's efforts to mitigate biases. However, responses can vary, and the model may still produce biased outputs depending on the prompt and context due to inherent biases in the training data.

### ***3.2 Discuss whether AI could be considered "intelligent" or "conscious" and the criteria for such classifications.***

The question of whether AI can be considered "intelligent" or "conscious" is a complex and multifaceted issue that sits at the intersection of philosophy, cognitive science, and artificial intelligence research. Intelligence in AI is often defined by its ability to perform tasks that would typically require human intelligence, such as understanding language, solving problems, and learning from experience. According to this functional perspective, AI can exhibit forms of intelligence through pattern recognition, data analysis, and generating human-like responses. For instance, large language models like ChatGPT can produce coherent and contextually relevant text, demonstrating a kind of synthetic intelligence.

However, the concept of consciousness is far more elusive and subjective. Consciousness involves self-awareness, subjective experiences, and the ability to perceive and react to one's own existence and environment. Current AI lacks these qualities as it operates based on programmed algorithms and lacks any form of self-awareness or subjective experience. AI processes data and produces outputs without any understanding or experience of the content it handles.

Criteria for classifying intelligence and consciousness in AI might include the Turing Test, which evaluates a machine's ability to exhibit intelligent behavior indistinguishable from that of a human, and the more stringent criteria of machine consciousness, which would require AI to have a form of self-awareness and experiential understanding. As of now, while AI can simulate intelligent behavior, it does not possess consciousness and therefore cannot be classified as truly intelligent or conscious in the human sense.



### 3.3 *Explore the philosophical question of whether machines can truly understand context or if they're merely processing data.*

The philosophical question of whether machines can truly understand context or if they are merely processing data delves into the nature of comprehension and the limitations of artificial intelligence. Understanding context involves grasping the nuances, intentions, and underlying meanings within a particular situation or piece of information. Human understanding is deeply rooted in subjective experience, emotions, and a vast network of associative memories.

AI, including advanced models like ChatGPT, processes context through pattern recognition and statistical correlations derived from extensive datasets. While these models can generate text that appears contextually appropriate, they do so by analyzing the probabilities of word sequences and patterns learned during training, rather than by genuinely understanding the context as a human does. For instance, AI can recognize that "the cat sat on the mat" is a sensible continuation of "the cat," but it does so without any awareness or comprehension of what a cat or a mat actually is.

John Searle's Chinese Room argument illustrates this distinction by proposing that a person following a set of syntactic rules to manipulate Chinese symbols could appear to understand Chinese without truly comprehending it. Similarly, AI follows complex algorithms to generate contextually relevant responses without any real understanding.

Therefore, while AI can simulate an understanding of context to a remarkable degree, it ultimately lacks the experiential and subjective comprehension that characterizes true understanding. It processes data in sophisticated ways but remains fundamentally different from human cognition, which is deeply intertwined with personal experience and consciousness.



**3.4 *Neural networks, and by extension ChatGPT, are often referred to as a 'Black Box'. Explain what is meant by this term with reference to weights, embedding matrices, deep learning and back propagation. Explain what the consequences are of these systems being a 'Black Box'.***

### Neural Networks and the 'Black Box' Concept

Neural networks, including models like ChatGPT, are often referred to as "black boxes" due to the complexity and opacity of their internal workings. This term highlights the difficulty in understanding and interpreting how these systems make decisions and predictions, despite knowing the overall structure and mechanisms they use. Here, we explore the components that contribute to this 'black box' nature and its consequences.

### Weights and Embedding Matrices

In neural networks, weights are numerical values that determine the strength of the connections between neurons. These weights are adjusted during training through a process called backpropagation, which aims to minimize the difference between the predicted output and the actual output. The embedding matrix, used in models like ChatGPT, maps discrete input tokens (words or subwords) into high-dimensional continuous vectors that capture semantic meanings.

In transformer models like ChatGPT, several components have been pre-trained to perform various tasks:

1. **Embedding Matrices:** These are pre-trained to convert tokens into dense vectors that capture semantic relationships between words.
2. **Weights in the Transformer Layers:** This includes weights in the self-attention mechanisms and the feed-forward neural networks. These layers are crucial for processing input sequences and generating context-aware embeddings.
3. **Unembedding Matrices:** These are used to map the high-dimensional vector outputs back into tokens, enabling the generation of human-readable text.

These embeddings and weights are crucial for the network's functioning but are difficult to interpret directly. Each weight influences the network's output, but the sheer number and interdependencies of these weights make it challenging to understand how individual inputs lead to specific outputs.



## Consequences of the 'Black Box' Concept

1. **Interpretability and Transparency:** The 'black box' nature means that we often do not understand why a neural network made a particular decision. This lack of transparency can be problematic in critical applications like medical diagnosis, financial forecasting, or autonomous driving, where understanding the rationale behind a decision is crucial.
2. **Bias and Fairness:** Since the internal workings are not fully interpretable, identifying and mitigating biases in the model can be challenging. If a model learns biased patterns from training data, it might propagate these biases in its predictions, leading to unfair or discriminatory outcomes.
3. **Debugging and Error Correction:** Troubleshooting errors in neural networks is difficult because the 'black box' nature obscures the root causes of errors. Understanding and correcting why a model fails in certain scenarios requires extensive testing and might not always lead to clear answers.
4. **Trust and Accountability:** Trust in AI systems is undermined when users and developers cannot explain how decisions are made. This lack of accountability can hinder the adoption of AI technologies in sensitive or regulated fields.
5. **Ethical and Legal Implications:** The opacity of neural networks raises ethical and legal questions about responsibility and accountability. If an AI system makes a harmful decision, determining who is responsible—the developers, the operators, or the system itself—becomes a complex issue.

In conclusion, while neural networks and models like ChatGPT are powerful tools capable of performing sophisticated tasks, their 'black box' nature presents significant challenges in terms of interpretability, transparency, and trust. Understanding and addressing these challenges is crucial for the responsible development and deployment of AI technologies.

## 4. Questions

1. How do we define thinking in humans, and what are the key characteristics that distinguish it from mere data processing or pattern recognition?



2. Given its ability to generate human-like text based on input, can we argue that GPT systems engage in a form of thinking, or are they simply executing pre-defined algorithms without genuine thought?
3. What are the essential elements of consciousness, and how do these elements manifest in humans and potentially other animals?
4. Considering the definitions and criteria for consciousness, can ChatGPT or similar AI systems ever achieve a state of consciousness, or is this an inherently human trait?
5. In what ways might the capabilities of ChatGPT challenge or reshape our notions of human thought and cognitive processes?
6. How does interacting with AI like ChatGPT impact our own thinking processes, problem-solving approaches, and the way we consume and generate information?

## 5. Appendix

(i) How does ChatGPT work? Explained by Deep-Fake Ryan Gosling.

[https://youtu.be/xU\\_MFS\\_ACrU?si=g8Alx3LSKVyZ1dRH](https://youtu.be/xU_MFS_ACrU?si=g8Alx3LSKVyZ1dRH)

(ii) But what is a GPT? Visual intro to transformers | Chapter 5, Deep Learning

[https://youtu.be/wjZofJXov4M?si=Tjx1\\_THprt3d-5qa](https://youtu.be/wjZofJXov4M?si=Tjx1_THprt3d-5qa)

(iii) [1hr Talk] Intro to Large Language Models

[https://youtu.be/zjkBMFhNj\\_g?si=CUGb\\_E6gWzFqbOHG](https://youtu.be/zjkBMFhNj_g?si=CUGb_E6gWzFqbOHG)

(iv) <https://chatgpt.com/>



**Manchester  
Lit & Phil**  
Your notes

