Challenging the "intelligence" of ChatGPT and a neuropsychological study on the function of the mind to tell a story

Abstract

We want to describe a neuropsychological evaluation conducted on ChatGPT to assess its level of "intelligence." This evaluation used tests routinely employed to evaluate prefrontal functioning in humans, as human intelligence is thought to rely on the functional integrity of the frontal lobes. The findings suggest that while ChatGPT excels in generative linguistic tasks, its performance on prefrontal tests varied, showing strengths in some areas and weaknesses in others. In fact, we want to compare the performance of the front part of the human brain with the Chat GPT function to recall a story. The inconsistent profile of ChatGPT's performance suggests that its emergent abilities do not fully replicate human cognitive functioning, especially in tasks associated with prefrontal functions. This study addresses the limitations of human accuracy in deception detection by introducing an investigation into the lie-detection capabilities of ChatGPT 2. The research pioneers an examination across an English-language dataset involving personal opinions, autobiographical memories, and future intentions. Stylometric analysis illuminates linguistic nuances within the dataset. Employing 5-fold cross-validation in a single scenario, the study reveals promising outcomes, highlighting the potential of ChatGPT2 in advancing lie-detection techniques. The findings contribute valuable insights to the evolving field of linguistic analysis, emphasizing the model's role in enhancing accuracy in detecting deception.

Keywords: Artificial Intelligence, Chat GPT Models, ChatGPT, Prefrontal Functioning, Neuropsychological Evaluation

Introduction

1. Decoding Human Cognition with Neuropsychology

Neuropsychology, often regarded as the science of human cognition, is a branch of psychology that focuses on understanding the intricate relationship between the brain and cognitive functions. It delves into the exploration of how neural processes give rise to various aspects of human cognition, including perception, memory, language, attention, problem-solving, and executive functions.

As a scientific discipline, neuropsychology employs a multidisciplinary approach, integrating knowledge from neuroscience, psychology, and other related fields. Researchers and practitioners in neuropsychology utilize a variety of methods, such as neuroimaging techniques, neuropsychological assessments, and experimental paradigms, to investigate how brain structure and function contribute to cognitive processes.

The overarching goal of neuropsychology is not only to unravel the mysteries of how the brain supports cognitive functions but also to apply this knowledge in clinical settings. In clinical neuropsychology, practitioners use their understanding of the neural basis of cognition to assess and rehabilitate individuals with cognitive impairments resulting from neurological or psychiatric disorders.

In essence, neuropsychology as the science of human cognition seeks to unravel the mysteries of the brain-mind connection, shedding light on how the neural intricacies shape our cognitive experiences, and ultimately contributing to advancements in both theoretical understanding and practical applications for the benefit of individuals with cognitive challenges.

Prefrontal functions, intricately linked to the integrity of the prefrontal lobes, represent a pinnacle of cognitive abilities, traditionally encompassing attention, memory, language, and executive functions. These functions play a crucial role in goal-oriented behavior, volition, and planning [1]. Recent research expands their scope, revealing involvement in mood regulation, personality development, self-awareness, and social reasoning [2]. Selective frontal lobe lesions not only impact cognition but also manifest in behavioral and emotional symptoms, hindering daily functioning [3].

2. Evaluating Large Language Models in Comparison to AI

We want to evaluate intelligence, both in neuropsychology and artificial intelligence (AI). It highlights Chollet's theory, emphasizing a system's intelligence as the measure of its skill acquisition efficiency across tasks [4]. Chollet proposes a dataset based on the Raven test for non-verbal intelligence assessment.

In the realm of verbal intelligence, recent research focuses on LLMs like BERT [5] and GPT-2, showcasing remarkable capabilities in language-related tasks. LLMs, neural networks with vast parameters, are trained on extensive text datasets to generate coherent and contextually appropriate text. Their development marks a significant breakthrough in natural language processing, enabling automated language generation and understanding.

The debate ensues on whether the astounding performance of LLMs truly mirrors human intelligence or merely extends the capabilities of large-scale associators. Optimistic views posit that scaling models unlock emergent abilities beyond initial training tasks, while skeptics liken LLMs to "stochastic parrots," generating linguistic sequences based on learned patterns without deep semantics [6]. The dichotomy raises critical questions about the true nature of intelligence in machine learning.

3. ChatGPT

OpenAI introduced ChatGPT as a publicly accessible chatbot, positioned as a sibling model to Instruct GPT. Distinctive for its ability to follow instructions and deliver detailed responses, ChatGPT gained global popularity soon after its release.

Renowned for comprehending user instructions and providing nuanced responses, ChatGPT distinguished itself from previous Large Language Models (LLMs). Notably, it garnered attention for its capability to generate believable and original scientific abstracts, even surpassing plagiarism detection methods. The model's achievements extended to being credited as a co-author in scientific articles, sparking intense debates within the scientific community.

ChatGPT's outstanding performance has prompted a call for thorough comparative studies, highlighting its potential and impact in the domain of conversational AI and beyond.

4. Lie detection

Detecting lies entails the assessment of the truthfulness in a given communication. Individuals crafting deceptive narratives utilize verbal tactics to instill false beliefs in their conversational counterparts, engaging in a distinct and momentary psychological and emotional state [7]. The exploration and advancement of this subject persistently captivates the field of cognitive psychology, owing to its substantial and promising implications in forensic and legal contexts [8]. Its potential crucial function lies in assessing the truthfulness of witnesses and potential suspects throughout investigations and legal proceedings. This influence extends to both the investigative data collection phase and the ultimate decision-making stage [9].

Despite extensive research efforts directed towards identifying verbal indicators of deception and devising efficient methods to distinguish between truthful and deceptive narratives, these verbal cues, at their optimum, remain subtle. Typically, they lead both naive and expert individuals to perform just slightly above chance levels [10, 11]. Stylometry is the study of linguistic style, particularly in written texts, using computational and statistical methods to analyze patterns, structures, and features of an author's writing. This field examines various aspects of language, such as word choice, sentence structure, and punctuation usage, to identify unique patterns that can be indicative of an author's writing style. Stylometry is often applied in authorship attribution, plagiarism detection, and forensic linguistics, where it helps determine the likely authorship of a text or identify writing patterns across different documents [12, 13]. While stylometry has its constraints, it has proven to be a valuable tool in lie detection. One notable advantage is its capacity to independently code and extract verbal cues, mitigating the challenges associated with inter-coder agreement. This implies that when researchers apply the same technique to the same dataset, they can extract uniform indices, thereby enhancing the consistency and reliability of the analysis [14,15].

Exploring the refinement of a Large Language Model (LLM) through fine-tuning on limited datasets for lie-detection has not been investigated in this manner. LLMs, as Transformer language models, boast a minimum of hundreds of millions of parameters and undergo initial training on extensive corpora during the pre-training phase. This pre-training equips LLMs to grasp intricate language patterns and structures, fostering a robust understanding of syntax, semantics, and pragmatics. Subsequently, these pre-trained models can undergo fine-tuning for specific tasks using smaller, task-specific datasets, leading to cutting-edge performance. LLM fine-tuning commonly spans tasks such as language translation, text classification (e.g., sentiment analysis), question-answering, text summarization, and code generation. As a result, LLMs excel across a

diverse array of Natural Language Processing (NLP) tasks, distinguishing them from models designed exclusively for a singular task [16].

Given the exceptional adaptability of Large Language Models (LLMs), the primary goal of this study was to assess the effectiveness of fine-tuning an LLM for the classification of accuracy in succinct narratives extracted from raw texts. To accomplish this, we deployed an open-source LLM, specifically Chat GPT 2, incorporating three diverse datasets: personal opinions (**Deceptive Opinions dataset**) [17], autobiographical experiences (**Hippocorpus dataset**) [18], and future intentions (**Intention dataset**) [19]. It's crucial to note that these three types are integral to the main article, and for this report, our emphasis was on one of the datasets.

Finally, after using stylometry to find differences between truthful and deceptive statements in the dataset, we explored whether the way truthful or deceptive narratives are presented linguistically is a factor considered by the model for its final prediction. This allowed us to analyze verbal lie detection from the perspective of multiple theoretical frameworks.

Methods

1- Original paper

This article consists of three datasets: the Deceptive Opinions dataset, referred to as the Opinion Dataset, the Hippocorpus dataset, referred to as the Memory Dataset, and the Intention dataset. For each dataset, participants were required to provide genuine or fabricated statements in three different domains: personal opinions, autobiographical experiences, and future intentions, respectively. In this article, we used two models which were named FLAN-T5 small and FLAN-T5 Base. Also, 3 scenarios that we will explain below have been used:

Scenario 1:

Objective: Test the model's ability to identify lies within a specific context.

Procedure: Fine-tune and test the model on parts of a single dataset, repeated for each dataset.

Scenario 2:

Objective: Assess the model's performance on new context samples.

Procedure: Fine-tune on two datasets, test on the third dataset, repeated with different pairings.

Scenario 3:

Objective: Measure the model's generalization across multiple contexts.

Procedure: Aggregate datasets from Scenario 1, fine-tune, and test on the combined sets.

2- This paper:

In our article, we only used the Memory Dataset and the first scenario. The model was fine-tuned on a portion of a single dataset and tested on the remaining part. This scenario evaluates the capacity of the model to learn how to detect lies related to the same context. We divided the data into two parts, train and test that are 90% and 10% respectively.

Model

1- ChatGPT 2

ChatGPT-2 is a conversational language model developed by OpenAI. It is part of the GPT-2 (Generative Pre-trained Transformer 2) family, which is a series of powerful language models. Here are some key points about ChatGPT-2:

1- Architecture:

Transformer Model: ChatGPT-2 is based on the transformer architecture, a type of neural network architecture that has shown exceptional performance in natural language processing tasks.

2-Training:

Pre-training: Pre-training is the initial phase where the language model (e.g., ChatGPT-2) is trained on a large dataset that contains a wide range of diverse text from the internet.

During pre-training, the model learns the statistical patterns, syntactic structures, and contextual relationships present in the data. This phase provides the model with a general understanding of language.

Fine-tuning: After pre-training, the model can be fine-tuned on a specific dataset or task. Fine-tuning involves training the model on a narrower dataset that is related to a specific application or domain. By fine-tuning, the model adapts its general language understanding to the specifics of the target task or context. It helps tailor the model's capabilities to a particular use case.

Transfer Learning: Transfer learning is a machine learning paradigm where knowledge gained from one task (e.g., pre-training) is leveraged to improve the performance on another related task (e.g., fine-tuning). In the context of ChatGPT-2, transfer learning allows the model to utilize the knowledge acquired during pre-training when fine-tuning for specific applications. This approach is efficient, especially when dealing with limited task-specific data.

3- Conversational Capabilities:

ChatGPT-2 is designed for generating coherent and contextually relevant responses in a conversational setting. It can understand and generate human-like text based on the input it receives.

4- Generative Model:

It operates as a generative model, meaning it can create new text based on the patterns it has learned during pre-training. This makes it suitable for tasks like text completion, creative writing, and more.

5- Large Scale:

The "2" in GPT-2 indicates that it is a second version of the GPT model, and it has a large number of parameters, making it a high-capacity model capable of capturing complex language structures.

6- Conditional Text Generation:

While the original GPT-2 model is often used for unconditional text generation, ChatGPT-2 can be fine-tuned for specific tasks or contexts, allowing it to generate text based on certain conditions or prompts.

It's important to note that the capabilities and features of ChatGPT-2 may be further refined or expanded in newer versions or variants, so checking the latest information from OpenAI is recommended for the most accurate details.

2- Fine tuning

Fine-tuning ChatGPT-2 involves adjusting the model's parameters on a specific dataset related to a particular task or domain. This process refines the model's pre-trained knowledge for more specialized applications. The goal is to adapt the general language understanding gained during pre-training to better suit the nuances and requirements of a targeted use case. Fine-tuning allows ChatGPT-2 to generate more contextually relevant and task-specific responses.

3- Pre-processing:

The preprocessing steps outlined involve preparing a dataset for use, particularly in the context of training or utilizing a language model like GPT-2. Here's a breakdown of the mentioned preprocessing steps:

Removing Rows with Null Values: This step aims to ensure data quality by eliminating rows in the dataset where the "Story" column has missing or null values. Null values can interfere with the training process and removing them ensures a clean dataset for accurate model training.

Removing Rows with short stories: Also, the rows in the dataset where the "Story" column has very few values (less than 60 words) were eliminated. This can help filtering out noise or irrelevant information. It is also helpful to deal with overfitting and reducing the amount of unnecessary computation and memory usage during training.

Using GPT-2 Tokenizer: The GPT-2 tokenizer is applied to convert the stories in the dataset into tokens. Tokenization is the process of breaking down text into smaller units, or tokens, which could be words or subworlds. GPT-2's tokenizer helps represent the stories in a format suitable for the model's input. Tokens are the basic units that the model processes during training or inference.

In summary, these preprocessing steps involve cleaning the dataset by removing rows with missing values and then transforming the textual stories into a tokenized format using the GPT-2 tokenizer. These prepared datasets can then be used for training or other natural language processing tasks with the GPT-2 model.

4-Modeling

The mentioned modeling step involves the utilization of GPT-2 for sequence classification, specifically through a model variant known as "GPT-2 ForSequenceClassification." Here's a breakdown:

GPT-2 Model: GPT-2, or Generative Pre-trained Transformer 2, is a powerful language model developed by OpenAI. It belongs to the transformer architecture and is pre-trained on a diverse range of internet text, allowing it to understand and generate human-like language. GPT-2 can be employed for various natural language processing tasks due to its versatile and context-aware understanding of language.

ForSequenceClassification: The model is configured for sequence classification, meaning it is trained to categorize or classify text sequences into predefined categories or labels. This application is particularly useful when dealing with tasks such as sentiment analysis, topic categorization, or any problem where assigning predefined labels to text sequences is required.

In summary, the modeling step involves leveraging the GPT-2 model, configured for sequence classification, to classify text sequences into predefined categories. This adaptation of GPT-2 showcases its flexibility in handling a range of natural language processing tasks, including classification.

Results

Following pre-processing, the dataset experienced a reduction in the number of rows, decreasing from 6855 to 5450. Despite this reduction, the impact on the model's accuracy was minimal. Although this reduction did not have much effect on the accuracy of the model, helped to reduce the calculation time. Removing null and very short values can be effective in several ways. It can enhance model performance by reducing noise, preventing overfitting, focusing on relevant information and reducing the amount of unnecessary computation and memory usage.

The model's accuracy, as measured through 5-fold cross-validation with 3 epochs and a batch size of 2, was found to be 0.89±0.07. This evaluation underscores the robustness and reliability of the model.

Discussion

Furthermore, a comparative analysis was conducted between the GPT-2 and BART models applied by other groups, and the results are presented in Table 1.

BART (Bidirectional and Auto-Regressive Transformers) and GPT-2 (Generative Pre-trained Transformer 2) are both advanced language models based on transformer architectures, but they have distinct characteristics and purposes. Here are the key differences between BART and GPT-2:

1. Objective and Architecture:

-BART: BART is designed for sequence-to-sequence tasks, such as language translation and text summarization. It utilizes a combination of auto-regressive and bidirectional transformers, allowing it to process input sequences bidirectionally and generate output sequences auto-regressively. BART can be pre-trained on a denoising autoencoder objective.

- GPT-2: GPT-2 is primarily designed as a generative language model. It employs a unidirectional transformer architecture for predicting the next word in a sequence. GPT-2 is pre-trained using unsupervised learning on a diverse range of tasks, enabling it to generate coherent and contextually relevant text.

2. Training Approach:

- BART: BART can be pre-trained using a denoising autoencoder approach. In this method, the model learns to reconstruct the original input sequence from a corrupted or noisy version of the sequence.

- GPT-2: GPT-2 uses a pre-training approach known as unsupervised learning. It is trained to predict the next word in a sequence based on the context of the previous words.

3. Fine-Tuning:

- BART: BART is often used for specific downstream tasks through fine-tuning on task-specific datasets. This allows it to adapt its pre-trained knowledge to the specifics of tasks like summarization or translation.

- GPT-2: GPT-2 can also be fine-tuned for specific tasks, but its original release included a large pre-trained model that was already capable of generating coherent and contextually relevant text across a wide range of domains.

4. Task Specificity:

- BART: BART is well-suited for sequence-to-sequence tasks, making it particularly effective for applications like text summarization and language translation.

- GPT-2: GPT-2, with its focus on generative language modeling, is versatile and can be used for a broader set of tasks, including text completion, story generation, and dialogue generation.

5. Size and Scale:

- BART: BART models can vary in size, but they are typically designed for specific tasks and might not be as large as the largest GPT-2 models.

- GPT-2: GPT-2 gained attention for its large-scale model with 1.5 billion parameters, which was one of the largest language models at the time of its release.

In summary, while both BART and GPT-2 are based on transformer architectures and are capable of generating text, they differ in their objectives, training approaches, and specific use cases. BART is tailored for sequence-to-sequence tasks, while GPT-2 is a versatile language model with a focus on generative language modeling across a range of applications.

With learning rate = 5e-6, batch size of 2, and 5-fold-cross validation, the average accuracy of the BART model was 0.794095 (the standard deviation of accuracy: 0.022

and the max accuracy: 0.781). However, in the original paper, the accuracy of the FLAN-T5-Base model was 0.8061 which is really close to the accuracy with the BART model.

Notably, GPT-2 outperformed the BART model by approximately 12.95%, showcasing its superior accuracy in the given context. This performance difference can be pivotal in scenarios where precision and efficiency are critical considerations for model selection and deployment.

Table 1

Test accuracy of the models for the Memory dataset. Values are means ± standard deviation of the 5-folds and batch size equal to 2.

Model	**Memory**
GPT 2 - Scenario 1	0.89±0.07
BART - Scenario 1	0.79±0.01
FLAN-T5-Base – Scenario 1	0.8061±0.01

Resources

1- Stuss, D. T., & Alexander, M. P. (2000). Executive functions and the frontal lobes: a conceptual view. Psychological Research, 63(3–4), 289–298.

2- Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. Trends in cognitive sciences, 9(5), 242-249.

3- Chayer, C., & Freedman, M. (2001). Frontal lobe functions. Current Neurology and Neuroscience Reports, 1(6), 547-552.

4- Chollet, F. (2019). On the measure of intelligence.

5- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

6- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022a). Emergent abilities of large language models.

7- Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. New Ideas in Psychology, 34, 22–36. https://doi.org/10.1016/j.newideapsych.2014.03.001 (2014)

8- Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. Verbal lie detection: Its past present and future. Brain Sciences, 12, 1644., https://doi.org/10.3390/brainsci12121644 (2022).

9- Vrij, A., & Fisher, R. P. Which lie detection tools are ready for use in the criminal justice system? Journal of Applied Research in Memory and Cognition, 5, 302–307.

10- DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. Cues to deception. Psychological Bulletin, 129, 74–118. https://doi.org/10.1037/0033-2909.129.1.74.

6 .11-Bond, C. F., Jr., & DePaulo, B. M. Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214–234. https://doi.org/10.1207/s15327957pspr1003_2 (2006).

12- Chen, H. Dark web: Exploring and mining the dark side of the web. In 2011 European Intelligence and Security Informatics Conference, 1-2. IEEE . (2011, September)

12 .13- Daelemans, W. Explanation in computational stylometry. In Computational Linguistics and Intelligent Text Processing, 451–462 Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-37256

14- Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. Are computers effective lie detectors? A metanalysis of linguistic cues to deception. Personality and social Psychology Review, 19, 307-342. https://doi.org/10.1177/1088868314556539 (2015).

15- Tomas, F., Dodier, O., & Demarchi, S. Computational measures of deceptive language: Prospects and issues. Frontiers in Communication, 7. https://doi.org/10.3389/fcomm.2022.792378 (2022).

16- Zhao, W. X., et al. A survey of large language models. arXiv preprint arXiv:2303.18223. (2023).

17- Capuozzo, P., Lauriola, I., Strapparava, C., Aiolli, F., & Sartori, G. DecOp: A multilingual and multidomain corpus for detecting deception in typed text. In Proceedings of the 12th Language Resources and Evaluation Conference, 1423-1430, (2020, May).

18- Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. Recollection versus imagination: Exploring human memory and cognition via neural language models. In Proceedings of the 58th annual meeting of the association for computational linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.178 (2020, July).

19- Kleinberg, B., & Verschuere, B. How humans impair automated deception detection performance. Acta Psychologica,213, https://doi.org/10.1016/j.actpsy.2020.103250 (2021).