## 1. INTRODUCTION

LLMs are language models in NLP that have been trained on vast amounts of data using deep learning frameworks with an extremely large number of parameters. These language models are capable to recognize patterns, make inferences, and produce coherent and contextually relevant responses. This results in highly sophisticated language outputs with human-like language style and quality. LLMs are built using deep learning techniques and neural networks, specifically recurrent neural networks (RNNs) or transformers, which are a more recent and powerful architecture.

LLMs can be used for a variety of tasks, including text generation, translation, summarization, question answering, and chatbots. Work [9] summarizes LLMs' emerging abilities in three areas: (1) in-context learning, (2) instruction following, and (3) step-by-step reasoning.

## 2. HISTORY

Language models were first developed as a new path forward for LLMs in the 1960s. A Joseph Weizenbaum chatbot, ELIZA, that used pattern matching and pre-programmed responses in its early stages. Another chatbot utilizing a similar methodology, ALICE was created by Richard Wallace in the 1990s with the addition of machine learning to enhance response quality. Cleverbot, developed by Rollo Carpenter in the 2000s, used artificial neural networks to produce more varied and natural responses. [4]

For the generation of coherent and varied text, OpenAI's GPT-1 (Generative Pre-trained Transformer) released in 2018, released a paper with the title "Improving Language Understanding by Generative Pre-Training." It was trained using a sizable corpus of texts and was built on the transformer architecture. BERT, Bidirectional Encoder Representations from Transformers, is created by Google AI and released in 2018. It is driven by Transformer technology, which enables it to understand text's long-range dependencies. GPT-2 Introduced in 2019, this upgraded version has 1.5 billion parameters and is capable of producing text that is human-like in quality. LaMDA (Language Model for Dialogue Applications) is a LLM created by Google AI and released in May 2022. Its Transformer-Decoder technology, which enables it to produce text that is both fluid and cohesive, powers the system. OpenAI's GPT-3 was released in November 2022. It uses Transformer-XL technology, which enables it to learn even more extensive text dependencies. Pathways Language Model (PaLM) powered by the Pathways technology, which allows it to learn from a variety of different data sources and tasks. Released in January 2023 by Google AI. Microsoft Turing NLG Model, MEENA, powered by the Turing NLG technology, which allows it to generate text that is both fluent and coherent. ChinChilla, low latency and high through put LLM. powered by the Pathways technology. Both MEENA and Chinchilla released in February 2023 by Google AI.

Released on March 14, 2023, GPT4 is a multimodal* large language model developed by OpenAI. As a transformer-based model, GPT-4 was pretrained to predict the next token and was then fine-tuned with Reinforcement Learning from Human Feedback (RLHF) and human alignment and AI feedback for policy compliance.

Google Bard is multimodal* created released on 21 March 2023 by Google AI. It is capable of processing and comprehending a variety of data modalities, including text, images, and audio. As

* Deep learning technique from the combination of various modalities of data. Due to the fundamentally dissimilar statistical characteristics of these modalities, integrating them requires specialized modelling techniques and algorithms.

a result, Bard may draw knowledge from a greater variety of sources and produce more precise predictions.

## 3. CURRENT CHALLENGES AND LIMITATIONS.

Despite LLM's great capabilities, numerous studies have consistently report serious limitations in areas such as factuality, reasoning, math, coding, biases, and so on.
Most challenging one's are described here.

### 3.1 Hallucinations [1]

Hallucination is the phenomena in which LLMs write language that appears cohesive but is not anchored in actual facts or reality. It could involve generating information that is fabricated or is not supported by the training data, resulting in the generation of false or misleading content. The encoding of knowledge in LLMs is inadequate, and knowledge generalization may result in "memory distortion." As a result, these models may generate incorrect information or "hallucinate." Falsie information could be Domain Specific paradigm, common-sense knowledge, irrevelence from main topic or conflict from its previously owned believes.
For instance, ChatGPT responded positively when asked whether it could assist in finding relevant journals to cite in a review paper, but then went on to create a list of five entirely fake publications. On repeat trials, it occasionally listed one or two legitimate publications, but other times it listed a paper that was made up or a legitimate document that had nothing to do with the question other than possibly sharing the same author. Therefore, a response produced by a LLM may be perfectly formatted but may not be factually accurate.

**Proposed Solution:**

**I. SelfCheckGPT [3]**

A simple sampling-based technique only relies on sampled responses and can thus be used on **black box models**, while also operating with **no external database**(Zero recourses fashion). SelfCheckGPT's motivating premises is that when an LLM understands a certain topic effectively, the sampled responses are likely to be similar and contain consistent facts. For hallucinated facts, however, stochastically sampled responses are likely to deviate and may contradict one another. By sampling many responses from an LLM, one can assess the consistency of information and determine which statements are true and which are hallucinations. To assess informational consistency, three variations of SelfCheckGPT are considered: BERTScore, question-answering, and n-gram.

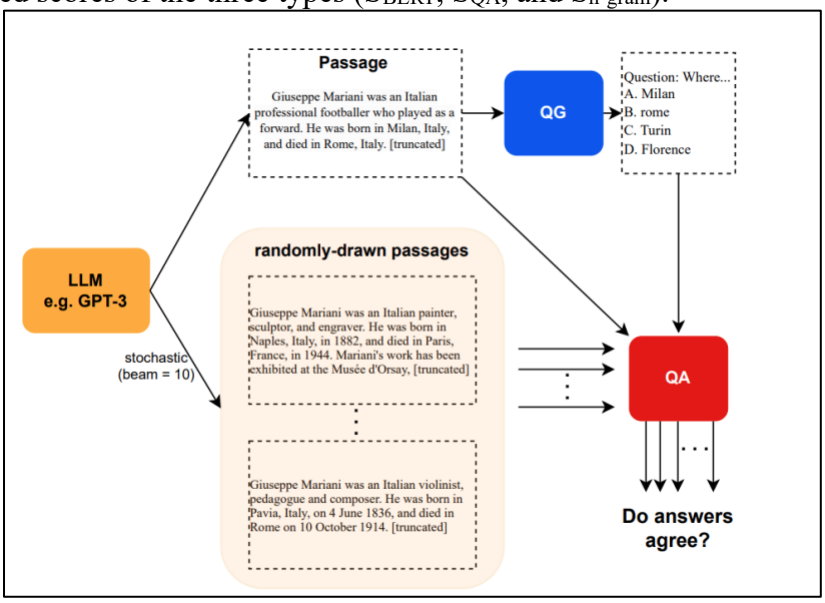
**SelfCheckGPT with BERTScore** finds the averages BERTScore of a sentence with the most similar sentence of each drawn sample.

**SelfCheckGPT with question-answering** uses The MQAG framework (Manakul et al., 2023) that examine consistency by producing multiple-choice questions that an answering system may independently answer given each passage. When facts about consistent concepts are questioned, the replying system should predict similar answers.

**SelfCheckGPT with n-gram**: A new language model could be trained using samples generated by an LLM to approximate the LLM. As N grows larger, this new language model approaches the LLM that generates the answer samples. As a

result, we may use the newly trained language model to approximate the LLM's token probabilities.

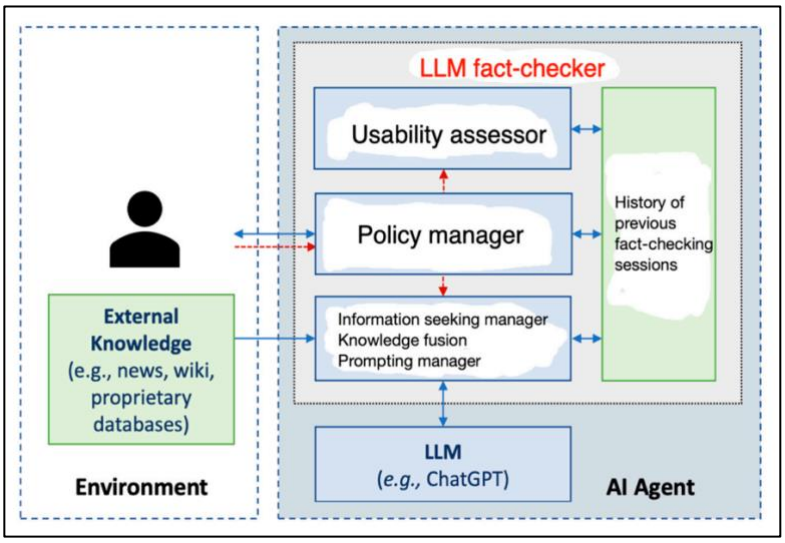
**SelfCheckGPT-Combination**, which is a straightforward combination of the normalized scores of the three types ($S_{BERT}$, $S_{QA}$, and $S_{n\text{-}gram}$).




## II. 'Truth-O-Meter' [1]

It identifies incorrect facts by comparing the generated results to the web and other sources of information and suggests Corrections. Text mining and web mining techniques are used for fact finding. additionally, the syntactic and semantic generalization procedure is applied to the content improvement task. Moreover, use a defeasible logic programming approach to argumentation analysis to handle contradictory web sources while fact-checking.

The inverse frequency of incorrect facts that are returned to the user determines the Truth-O-Meter reward R. The next Truth-O-Meter action that produces the best expected reward R is chosen by the policy manager. These steps involve obtaining evidence for user query from external knowledge, calling the LLM to generate a current candidate response, and, if all available knowledge has been cited, sending a response to users as shown in figure below:




The Truth-O-Meter policy blends hand-crafted rules with training data on actual or simulated human-system interactions.

The top-five HTML web pages are then scraped using the combined Google and Bing Search API, and a possible response is then extracted by fuzzy-matching the snippet from Google and Bing. There can be a maximum of five interactions with the search engine. A first answer is generated via chain-of-thought prompting [13] and is then corrected up to three times, ending after two times if the answer doesn't change. During our verification, we evaluate the plausibility and veracity with respect to R.

### 3.2 Biases and Fairness [4]

Biases included in the training data may be unintentionally learned by LLMs. The model can reproduce prejudices in its output text if the training data contains biased or prejudiced content. As Schramowski et al. [10] pointed out, large pre-trained models that try to mimic natural languages, may end up repeating the same unfairness and prejudices. This can lead to discriminatory or inaccurate analyses and recommendations. Moreover, this may lead to public outcry (i.e., political, social, and legal) against the commercial applications.[6]

**Demographic biases** occur when the model's training data leads to biased behavior towards specific genders, races, ethnicities, or other social groupings.

**Cultural biases** occur when large language models learn and perpetuate cultural stereotypes or biases. This could lead to the model's outputs reinforcing or exacerbating already-present cultural prejudices.

**Linguistic biases** develop as a result of the majority of online information being in English or a few other prominent languages, causing large language models to be more adept in these languages. This may result in biased performance and a lack of support for minority languages or low-resource languages.

**Temporal biases** emerge because the training data for these models is often limited to specific time periods. Lack of chronologically representative data may restrict the model's ability to understand past events or out-of-date information.

**Confirmation biases** in training data might occur as a result of individuals seeking information that confirms their pre-existing ideas. As a result, LLMs may unintentionally perpetuate these biases by producing results that confirm or support specific opinions.
LLMs can develop and transmit **ideological and political biases** because to the prevalence of such biases in their training data. This can result in the model producing outputs that favor specific political perspectives or ideologies, exacerbating existing biases.[6]

**Proposed Solution**:

| Solution | Description |
|---|---|
| Regular AI Audits | Regular AI model audits, which involve evaluating performance against predefined fairness, accuracy, and representativeness criteria, aid in the discovery of potential biases and errors. Continuous monitoring enables developers to fix biases before they become a problem. |
| Retraining of biased dataset | By retraining AI models with diverse and balanced curated data, it is possible to reduce the biases in the initial training data. By expanding model learning, this lessens the impact of ingrained biases. |
| Fairness Metrices | It may be feasible to identify and rectify potential inequalities or biases in treatment among user groups by evaluating the performance of AI models using fairness measures like demographic parity and equal opportunity. |
| Algorithmic debiasing techniques | Bias mitigation techniques, such as adversarial training, re-sampling, and re-weighting, are used during or after training to lessen the influence of biased patterns on model predictions. |
| Inclusion of diverse cultural perspectives | The discovery and eradication of biases are improved by diverse and inclusive AI development teams that bring a range of backgrounds and experiences to the table. Fairness and representation are promoted when people from many fields and cultures work together. |
| Human-in-the-Loop Approach | Integrating human specialists into AI model development and decision-making adds an additional layer of quality control by bringing contextual expertise and ethical judgement. By identifying biases and errors, their feedback enhances the fairness and performance of the model. |

### 3.3 Interpretability and Explainability [4]

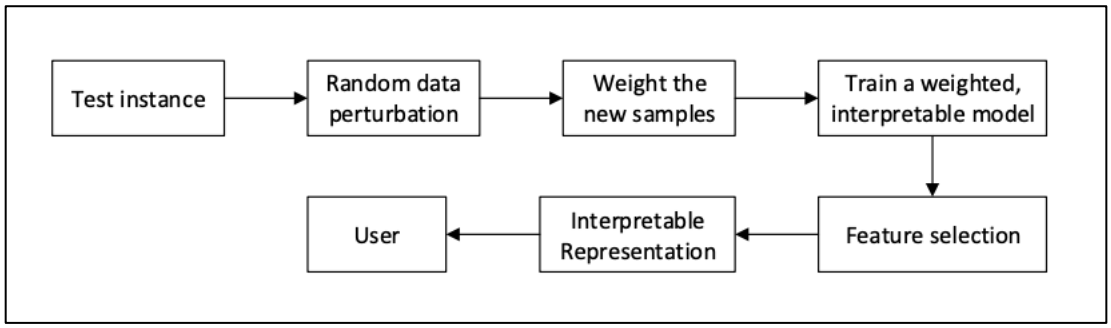
Explainability is the capacity to explain how a model arrived at its output or decision and the capacity to communicate that process to humans. Researchers will be able to understand the reasoning behind the models' decisions, which will also make their output transparent. Failure to address explainability has serious consequences for AI technique adoption and regulatory certification specially in HealthCare. Explainable AI (XAI) is a research area in AI that is gaining popularity because to the real-world deployment needs of AI-based systems.

**Proposed Solution:[4]**

I. Through **Attention**, we can define explainability in an AI system. It allows model to concentrate on certain parts of inputs while generating output results. it has the ability to produce probability distributions that are related to the input and act as indications of the significance of features. it can understandable which portions of the input the model is focusing on and how it uses that knowledge to produce its output by visualizing the attention weights for each component of the input [11].

However, according to Liu et al. [12], suppression effects may prevent attention systems from determining the polarity of the influence of specific features.

II. **LIME**, Local Interpretable Model Agnostic Explanations, is a surrogate model** that is employed to individually explain the predictions of an opaque model. LIME's goal is to locally train surrogate models and provide individual prediction explanations. A high-level block diagram of LIME is shown in Figure below. By randomly permuting the samples surrounding a sample from a normal distribution, it creates a synthetic dataset and then collects predictions using the opaque model that will be explained.




Then, LIME trains an interpretable model, such as linear regression, on this perturbed dataset. By using a regression line with the equation Y = a+bX, where "a" stands for the intercept and "b" for the slope of the line, linear regression preserves links between dependent variables like Y and several independent qualities like X. By using the provided predictor variables, this equation can be used to forecast the value of the target variable. LIME also requires as an input the number of critical features (K) that will be used to create the explanation. The model is simpler to understand the smaller the value of K.

III. In the context of Explainable AI, **counterfactual explanations** are hypothetical scenarios or instances that show how a model's prediction might change if some of the input features were different. These explanations emphasize the significance of various attributes and how their values affect the outcome in order to shed light on why a particular prediction was made.
Counterfactual explanations, to put it simply, assist people understand why an AI model made a particular decision by showing how changing key input variables would have produced a different result. These explanations, which provide concrete insights into the model's decision-making process, are helpful for improving transparency, interpretability, and confidence in AI systems.

**A surrogate model is an interpretable model that is trained to approximate the predictions of a black box model.

3.4 Transferability [4]

Transferability in LLMs refers to an LLM's ability to be applied to new task or data sets without having to be retrained from scratch. This is significant because it enables LLMs to

be utilized for a broader range of applications while saving time and costs. However, a variety of issues can impact the transferability of LLMs, including:

**Domain Shift**: LLMs may struggle to transfer knowledge across disciplines due to differences in language, themes, and writing styles. It's possible that the model's pretraining on a general dataset failed to account for all the subtleties of a new domain.

**Biased Amplified**: When applied to a new task or domain, biases that the pretrained LLMs picked up from their training data could be amplified, thereby producing biased or unjust results.

**Fine-Tuning challenges:** LLMs must be carefully handled when being fine-tuned for new jobs. Transferability may be impacted by the degree of fine-tuning, the selection of data, and hyperparameters. Performance on new data can suffer from overfitting on the target task.

**Data Distribution Mismatched:** The LLM may not be able to generalize successfully if the training data and the data being used are sufficiently dissimilar from one another. It's important to have enough labelled data for the target job or area in order to transfer information successfully. The model's performance might decrease if labelled data is inadequate or nonexistent.

**Proposed Solution:**

Large Language Models (LLMs) are one machine learning domain where transfer learning is an effective method to address transferability difficulties. The process of transfer learning entails pretraining a model on a large diverse dataset before optimizing it for a particular target task or domain. This approach can help mitigate transferability challenges by leveraging the knowledge gained during pretraining to improve performance on the target task. Transfer learning is a key strategy in the context of LLMs. Pretraining on a vast corpus of text helps the model learn grammar, syntax, and general language patterns. The model can be fine-tuned using task-specific data to tailor its knowledge to the current context, reducing the impact of domain shifts and biases.

It's important to remember that it might not entirely resolve all transferability problems. To achieve successful adaptation of the model to the target task or domain, significant fine-tuning, careful data selection, and appropriate evaluation are still needed.

Reference:


1. B. A. Galitsky, "Truth-O-Meter: Collaborating with LLM in Fighting its Hallucinations," 2023.
2. J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P. C. Peng, and J. H. Moore, "ChatGPT and large language models in academia: opportunities and challenges," BioData Mining, vol. 16, no. 1, p. 20, 2023.
3. P. Manakul, A. Liusie, and M. J. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," arXiv preprint arXiv:2303.08896, 2023.
4. M. Fraiwan and N. Khasawneh, "A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions," arXiv preprint arXiv:2305.00237, 2023.

5. Z. Liu, Z. Yao, F. Li, and B. Luo, "Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT," arXiv preprint arXiv:2306.05524, 2023.
6. E. Ferrara, "Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models," arXiv preprint arXiv:2304.03738, 2023.
7. X. Hu, Y. Tian, K. Nagato, M. Nakao, and A. Liu, "Opportunities and challenges of ChatGPT for design knowledge management," arXiv preprint arXiv:2304.02796, 2023.
8. M. R. Zafar and N. M. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," arXiv preprint arXiv:1906.10263, 2019.
9. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, and J. R. Wen, "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.
10. P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," Nature Machine Intelligence, vol. 4, no. 3, pp. 258-268, 2022.
11. J. Choo and S. Liu, "Visual analytics for explainable deep learning," IEEE Computer Graphics and Applications, vol. 38, no. 4, pp. 84-92, 2018.
12. Y. Liu, H. Li, Y. Guo, C. Kong, J. Li, and S. Wang, "Rethinking attention-model explainability through faithfulness violation test," in International Conference on Machine Learning, pp. 13807-13824, PMLR, June 2022.
13. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, pp. 24824-24837, 2022.