Controversia Philosophica

1st Mark Chausov
Innopolis University
Innopolis, Russia
m.chausov@innopolis.university

2nd Artem Bulgakov
Innopolis University
Innopolis, Russia
a.bulgakov@innopolis.university

3rd Anton Kirilin
Innopolis University
Innopolis, Russia
a.kirilin@innopolis.university

***Abstract*—In the modern world chatbots, grounded in natural language processing and machine learning, have evolved beyond basic conversational agents to become dynamic generators of human-like text. One of key aspects of this machine is to create an interaction with the user, by combining 2 or more agents into a group of chatters we can increase the overall quality of the result, but with specific downsides.**

***Index Terms*—generative AI, chatbot, agents, NLP, transformers, text generation**

I. Introduction

The study examines the varied chatbot models for text generation to create a content that would help users in discussions of varied topics. Additionally, the research touches on ethical considerations and challenges associated with deploying generative chatbots. Issues related to bias, misinformation, and the responsible use of AI in decision-making processes are examined to provide a comprehensive understanding of the implications and responsibilities involved.

II. State of the art review

A. Transformers

Transformers have emerged as revolutionary text generation models, fundamentally altering the landscape of natural language processing (NLP). Introduced by Vaswani et al. [1] in 2017, transformers excel at capturing long-range dependencies and contextual information, making them exceptionally well-suited for tasks such as text generation. Unlike traditional recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), transformers leverage attention mechanisms to weigh the importance of different words in a sequence, enabling them to process information in parallel and significantly speeding up training times.

One of the key strengths of transformers in text generation lies in their ability to model relationships between words in a non-linear fashion. By leveraging self-attention mechanisms, transformers can assign varying degrees of importance to different parts of the input sequence, allowing them to capture intricate patterns and dependencies. This not only facilitates the generation of coherent and contextually relevant text but also enables transformers to handle a wide range of language tasks, from translation to summarization, with remarkable proficiency. As a result, transformers have become the backbone of state-of-the-art language models, including widely used variants such as GPT (Generative Pre-trained Transformer) models, showcasing their transformative impact on the field of text generation.

B. AutoEncoders

Autoencoders, originally designed for unsupervised learning tasks, have found application in the realm of text generation by virtue of their ability to learn efficient representations of input data. In the context of natural language processing, autoencoders are employed as text generation models by training them to encode and subsequently decode textual information. The architecture comprises an encoder network that compresses input text into a latent space representation and a decoder network that reconstructs the original input from this condensed representation. Variational autoencoders (VAEs), a specific type of autoencoder, introduce a probabilistic element to the encoding process, allowing for the generation of diverse and novel text samples.

The utility of autoencoders in text generation lies in their capacity to capture semantic features and patterns in the input data. By learning a compact representation of the input text in the latent space, autoencoders can generate new text samples by sampling from this space. This makes autoencoders suitable for tasks such as data denoising, text completion, and even creative text generation. While not as prevalent as transformer-based models in certain natural language processing applications, autoencoders contribute a unique perspective to text generation, particularly in scenarios where the emphasis is on learning informative representations of textual content.

C. Large Language Models (LLMs)

Large Language Models represent a pivotal development in natural language processing, offering a sophisticated approach to understanding and generating human language. These models are designed to capture the intricate patterns and contextual dependencies inherent in linguistic data. Traditional language models, such as n-gram models, rely on statistical techniques to estimate the likelihood of word sequences based on observed data. However, recent advancements in neural network-based LLMs have transformed the landscape of NLP, enabling more nuanced and context-aware language understanding.

One exemplar architecture within the LLM domain is the transformer. Introduced by Vaswani et al. in 2017, transformers have become a cornerstone in NLP. Their unique self-attention mechanism enables them to efficiently capture long-range dependencies in input sequences, making them highly effective

in tasks like language modeling, machine translation, and text generation. The transformer's ability to process input data in parallel contributes to its efficiency and scalability, allowing it to handle vast amounts of textual information and learn complex linguistic structures.

Training LLMs, such as the transformer-based models like GPT (Generative Pre-trained Transformer), involves a two-step process. In the pre-training phase, the model is exposed to extensive and diverse text data, predicting the next word in a sequence based on contextual information. This unsupervised learning phase allows the model to grasp syntax, semantics, and contextual relationships. The subsequent fine-tuning phase narrows the focus to task-specific objectives, adapting the pre-trained model to more specialized applications such as text completion or summarization. LLMs, with their ability to learn from large datasets and generalize across various language tasks, have become integral in advancing the capabilities of natural language processing systems.

III. AI DECISION FALLACY

The AI decision fallacy refers to a misconception or error in judgment that can arise from placing undue trust or reliance on the decisions made by artificial intelligence systems. Despite their advanced capabilities, AI systems are not infallible and can be susceptible to biases, errors, or limitations in their training data. This fallacy often occurs when individuals assume that AI decisions are inherently objective and devoid of human-like errors, overlooking the potential for biases encoded in the algorithms or the data on which they were trained.

One of the key contributors to the AI decision fallacy is the lack of transparency in complex machine learning models. Many AI systems operate as black boxes, making it challenging to understand the underlying processes and factors influencing their decisions. This opacity can lead users to blindly trust AI-generated outcomes without critically examining the potential sources of bias or error. It is crucial to recognize that AI models are not inherently unbiased and that their decisions can reflect and even amplify societal biases present in the data they were trained on.

To ensure social responsibility of AI and address the problem of AI Decision Fallacy, the following steps are crucial:

- Enhance explainability: It is important to develop AI models that can provide clear explanations for their decisions.
- Enforce intermediate conclusions: Implementing a process where AI systems provide intermediate outputs or conclusions can help mitigate errors and biases.
- One proposed solution is the use of NLP GANs. They have the ability to generate diverse counterfactual texts and provide human-understandable text-based justifications. This enables users to understand the underlying reasoning, identify any biased patterns, and make necessary corrections.

A. Assembled learning

Assembled learning, conceptualized as a multiagent debate, presents a novel and promising approach to addressing the AI fallacy problem. The AI fallacy, often rooted in biases and limitations within training data, can be mitigated through the collaborative efforts of multiple AI agents engaged in a constructive debate. Each agent brings its unique perspective, biases, and knowledge, allowing for a more comprehensive evaluation of decisions and reducing the impact of individual biases.

In the context of a multiagent debate, diverse viewpoints and considerations can be systematically integrated into the decision-making process. The collaborative nature of the debate encourages agents to challenge assumptions, identify biases, and collectively refine decision outputs. This not only fosters a more robust decision-making framework but also introduces a layer of transparency as the rationale behind each decision can be scrutinized and discussed.

Furthermore, assembled learning as a multiagent debate promotes continuous learning and adaptation. Agents can update their understanding based on insights gained from the debate, ensuring that the system evolves over time to address emerging challenges and biases. By encouraging dynamic interactions and knowledge exchange among agents, assembled learning offers a pathway to a more resilient and ethically sound AI ecosystem, mitigating the pitfalls associated with the AI fallacy.

Study [2] shows that multi-agent models perform better than singular-agent models. They tested and compared the results of the models on reasoning, factuality, and question-answering tasks. Their experiments were done both with a chain of though and without, in each, since it also affect the performance. The overall conclusion is that to increase the performances of a model it's better to use multiple agents with several rounds of inference, so the result is more accurate.

B. Domain fine-tuning

The provision of domain-specific data is a pivotal strategy for enhancing the accuracy of machine learning models within targeted application areas. By tailoring the training data to the specific domain of interest, models can learn more nuanced patterns and relationships relevant to that particular context. For instance, in medical diagnostics, providing a dataset exclusively focused on healthcare scenarios allows the model to grasp intricate details specific to diseases, symptoms, and treatments, resulting in a higher accuracy when making predictions within the medical domain.

Domain-specific data not only refines the model's understanding but also helps alleviate issues related to out-of-domain challenges. Models trained on diverse datasets may struggle when confronted with scenarios outside their original scope. However, by curating data that aligns closely with the intended application domain, practitioners can significantly improve a model's accuracy, ensuring it performs optimally within the specific context it was designed for. This tailored approach to data provision is particularly crucial in sectors

where precision and reliability are paramount, such as finance, healthcare, or legal applications.

Authors of [3] propose to use a special system for answering question that are especially specific their domain. It increases the model result even with decreased number prompt tokens. Therefore in case of debate chatbots it's highly recommend to add this specific context to the debating agents, so the overall performance of the model is better.

IV. Methodology

For this assignment we decided to create 2 agents working as antagonists debaters with a 3rd agent - Iudex that analyses their argument and chooses a better one, so it doesn't require to get a feedback from a human. For this purposes we created a special class that initiates a debating agent using already pretrained models. There are 2 main sources for them:

1) OpenAI allows to get acsees to GPT models
2) Hugging Face to get access to a wider range of models, but not that complex

Initially user gives a topic for discussion as the initial prompt for context. After that each agent is initiated with a specific starting prompt for acting like a debater and information for the conversation. Finally, agents start to exchange their arguments untill a specific number of rounds is passed. The result of the model is a generated text which is the whole conversation of the agents itself. Except for the basic version models, that were trained on big corpus of wiki data. We also used models that were specifically pretrained for argument discussion, so by design they are more preferable for fine-tuning over other models.

V. Github

https://github.com/lcensies/Controversia-Philosophica/tree/dev

VI. Experiments and Evaluation

For experimenting we choose to use different initial prompt for describing the purpose of each agent. We had to ask them to not repeat their previous arguments and to not just agree or disagree with their opponent. Those recommendation help to avoid an "argument convergence" when both models didn't provide enough points that are different from the initial prompt and just repeat the info that is already presented in the whole conversation history. For grading the soundness of an argument a 3rd agent -Iudex is created, that can judge how valid a given argument is. However the analysis of the results of the whole system is yet to be done.

VII. Conclusion

In conclusion, the fusion of a multi-agent debate model with domain-specific data emerges as a powerful synergy, promising substantial enhancements in performance across various applications. The multi-agent debate framework introduces a collaborative environment where diverse agents contribute unique perspectives, fostering a dynamic exchange of insights. By incorporating domain-specific data, this approach tailors the learning process to the intricacies of a particular field, ensuring that the model gains a deep understanding of the nuances and context relevant to that domain. The amalgamation of these two elements leads to a robust decision-making mechanism that not only considers a variety of viewpoints but also leverages specialized knowledge for more accurate and informed outcomes.

Providing domain-specific data plays a pivotal role in amplifying the efficacy of the multi-agent debate model. This focused data provisioning allows the model to delve into the specifics of a given domain, capturing intricate patterns and relationships that might be overlooked in a more generalized dataset. Consequently, the model becomes adept at navigating complex scenarios within its designated domain, translating into heightened accuracy and reliability. Whether applied in healthcare, finance, or any other specialized field, the multi-agent debate model, fortified with domain-specific data, emerges as a potent tool for decision-making, capable of addressing the nuanced challenges and intricacies unique to each domain.

The symbiosis of a multi-agent debate model and domain-specific data not only elevates the model's accuracy but also contributes to transparency and adaptability. Through collaborative discussions among agents and the incorporation of tailored data, the model gains a more comprehensive understanding of the domain, making it better equipped to navigate uncertainties and evolving scenarios. This holistic approach holds immense promise in advancing the capabilities of artificial intelligence systems, paving the way for more precise, context-aware, and ethically sound decision-making processes across diverse domains.

Finally the overall structure of our system is proposed as show on the fig.1. This system should increase the accuracy of the generated text by using a domain-specific data and multiple agents

References

[1] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Łukasz, and Polosukhin Illia. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010.

[2] Du, Yilun and Li, Shuang and Torralba, Antonio and Tenenbaum, Joshua B and Mordatch, Igor. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv preprint arXiv:2305.14325

[3] Md Adnan Arefeen, Biplob Debnath , and Srimat Chakradhar. 2023. LeanContext: Cost-Efficient Domain-Specific Question Answering Using LLMs. arXiv preprint arXiv:2309.00841v1

Fig. 1. Proposed system structure