Comparative Analysis of Advanced Language Models: A Case Study of OpenAI's ChatGPT-3 and Bard Debating on Controversial Topics

William To
University of Technology Sydney
Sydney, Australia

Abstract— **In the era of rapid technological transformation, the influence of artificial intelligence (AI) has permeated various industries in an unprecedented manner. The AI industry business value proposition projected to reach $3.9 trillion by 2022, is a booming technological advancement that will significantly impact society through its widespread adoption. The advent of natural language models (NLP) and large language models (LLM) is the centre of attention of this artificial intelligence craze. The capabilities of LLM have advanced to the cusp of generating unique content extending from comprehensive answers and expressive poetry to intricate coding projects illustrating the versatility and adaptability of this upcoming technology's application within many industries. Among the vast number of LLM, ChatGPT and Bard are paving the path due to their strong abilities to understand and generate text in a conversational manner. The literature review initially investigates the technical architecture of AI technology and the impacts of ChatGPT and Bard within society to provide context on these advanced technologies.**

The research article examines the prospective capabilities and limitations of these models in a debate format using contentious issues. The topic of the debate is categorized into four domains: factual, optional, planning and problem-solving, to test the capabilities of ChatGPT and Bard to their utmost capacity. The format of the debate will be structured similarly to a regular debate in order to analyze for measures of relevance, coherence and factual accuracy within the responses of the models.

The exploration of this study has shed light on the strength and weaknesses of ChatGPT-3 and Bard within a debate setting. The results demonstrate that overall ChatGPT consistently outperforms Bard in a majority of the parameters aside from planning capabilities and factual accuracy. Each of the models exhibit similar levels of capabilities, however, ChatGPT responses exhibited more conversational fluid attributes which mimic human-like responses. Both the models are far from perfect as each of the models demonstrate qualities and attribute that far outweighs the other. The insight attain demonstrates that further development and refinement of the LLM are required, to improve the capabilities of LLM to answer controversial topics within a complex communicative scenario. This highlights the importance of continued advancement in AI technology for the improvement of the responses.

Keywords—Artificial Intelligence (AI), Machine Learning (ML), Artificial Neural Network (ANN), Deep Learning (DL), Large Language Model (LLM), Bard, ChatGPT

I. INTRODUCTION

In this period of technological advancement, the influence of artificial intelligence (AI) has expanded and matured exponentially in unprecedented manners with implications permeating various industries, from healthcare and finance to communication and technology. Once a concept in science fiction, that is now a cornerstone of advanced bleeding-edge technology, transcending the ease of performing simple repetitious tasks. The value proposition of AI-derived business is projected to have reached $3.9 trillion by 2022, reflecting the technology's widespread adoption and significant impacts [1]. These business values are forecasted to expand as many industry leaders integrate artificial intelligence technology within their business infrastructure and daily operations, such as code generation, content creation, mathematical proofs calculation etc. The capabilities of this technology have developed to the point where it can now devise intricate judgement, construct insightful perspectives, and address complex problem-solving dilemmas [2].

The advancement of artificial intelligence and computing hardware has brought together significant progression and power within deep neural network learning and natural language processing (NLP). In particular, generative AI has aided the development of natural language processing tools in terms of their precision in understanding and predictive processing. The evolution and comprehension of natural language processing have accelerated language models immensely leading to large-scale technological advancement and adoption. These AI constructs are trained to understand, generate, and manipulate human-like text by predicting subsequent words in a sentence and grasping contextual meanings [3]. The two most notable models which are ushering the trend of large language models and natural language processing towards a new era of human-computer interaction are ChatGPT and Bard.

As ChatGPT developed by OpenAI has propelled artificial intelligence to the forefront of the technological revolution of the 21st century, causing a global ripple effect due to its effectiveness in generating creative and context-appropriate responses. The expansive capabilities of language model technology embodied in systems like ChatGPT and Bard, exhibit a breadth of potential that transcends the boundaries of traditional text-based query responses. A robust knowledge foundation complemented by an unparalleled capacity for generating unique content extending from comprehensive answers and expressive poetry to intricate coding projects illustrates the versatility and adaptability of the technology applicable to many different applications [3]. The advent of these chatbots has ushered in a transformative era that necessitates additional study to fully comprehend the infinite applications, advantages, and possible drawbacks of advanced AI technologies across various domains.

With the growth of AI language models developing at an unprecedented rate, consideration must be taken on the

substantiation impacts and ramifications which this technology can have on the future of information propagation, policy-making and public discourse. AI has the potential to democratise access to expert-level analysis and promote a more educated populace [4]. However, it also poses significant concerns regarding the legitimacy of AI-generated content, the potential for bias in AI reasoning, and the ethical implications of AI addressing contentious issues.

The article seeks to contribute to the growing concern of these issues by conducting a comparative analysis of two of the latest advanced language model– OpenAI's ChatGPT-3 and Google's Bard to determine their capacity through a debate on controversial topics. The analysis will primarily focus on four categories of topics: factual, optional, planning and problem-solving. Each category encompasses five different questions that serve as points of debate, ranging from the scientific consensus on climate change and GMOs to strategies for mitigating the challenges of aging populations and cybersecurity.

The main objective of this article is to deepen our understanding of AI's potential and limitations in complex conversational contexts, provide insights into the implications of AI-led debates on controversial topics, and pave the way for the development of more effective and ethical AI communication tools. The comparative analysis will assess the performance of ChatGPT-3 and Bard in debating controversial topics, evaluating their coherence, logical structure and engagement, as well as their capacity to comprehend and respond appropriately to controversial topics. The evaluation will be held using the two language models to determine the relative strengths and weaknesses in the context of a debate. This will help determine the limitations of the responses for both ChatGPT and Bard to investigate the societal and ethical implications of AI debates, such as the potential for manipulation, the risks of bias and inequity, and the role of AI in shaping public discourse.

II. *Literature* Review

The following literature review analyses academic articles and reports to acquire a deeper comprehension of the innovative chatbot technology enabled by artificial intelligence. The objective is to dissect the technical specifications of the AI chatbot technology to provide a comprehensive evaluation. The literature review will be categorised as follows: evolution of artificial intelligence (AI), natural language processing (NLP) and large language models (LLM), ChatGPT, Bard, and AI response implications on society.

A. Evolution of Artificial Intelligence (AI)

Artificial Intelligence (AI), from its inception in the mid-20th century, has evolved from philosophical concepts to practical uses impacting multiple facets of society [5]. It began with a theoretical rule-based system that strictly adhered to pre-programmed instructions. These early rule-based systems utilised predetermined guidelines to generate output based on the received input. While the aforementioned system could provide engaging interactions, they were limited and dependent on whether the conditions had been pre-programmed; thus, it lacked the learning capacity to comprehend contexts beyond the system's scope.

The development of machine learning (ML) has broadened the horizon of artificial intelligence, by introducing an approach that implements data analysis, pattern recognition, and decision-making or forecasting based on the identified patterns. With the introduction of a concept that involves learning from data by machine learning within AI, the old archaic method of developing AI with pre-set rules and conditions has been eliminated. The purpose of this architecture eliminates resources that are required to determine all the possible and potential outputs within a system, reducing computational and time resources for constructing AI as these rules are summarised through a simple computational algorithm [6]. Machine learning is trained by incorporating a variety of learning methods, the most common being supervised, unsupervised and reinforcement learning. Despite the differences between the learning algorithm, they share a fundamental structure as they all train machine learning models through utilised datasets that are classified or pre-processed. This learning methodology enables the machine learning models to identify patterns and clusters within the dataset, developing a computational algorithm for the set task of identifying the desired object. Machine learning models are self-sufficient as they can improve their performance through exposure to more data, signaling the shift from a determined rule-condition-based architecture to a more self-learning-based probabilistic algorithm approach within AI [6].

As machine learning garners more attention and becomes a focal point of extensive research, a plethora of significant discoveries have been made concerning various pattern recognition and predictive modelling algorithms. The most notable, Artificial Neural Network (ANN) is a type of machine learning algorithm inspired by the biological construct and function of the human brain. An ANN is constructed with multiple layers of interconnected nodes, or artificial 'neurons,' each carrying out particular computations. The structure of an ANN demonstrated in Figure 1 would consist of an input layer, hidden layer and output layer, interconnected with an associated weight and threshold value [7]. These threshold values filter the node's output, passing the data to the next node. Neural networks are dependent on training data to improve learning and the accuracy of the predictive algorithm. ANN operates based on machine learning algorithms that utilize classified datasets to identify patterns and clusters, akin to supervised learning. As the ANN is trained using the learning algorithm, the associated weight and threshold values are acclimated to fine-tune the accuracy of classifying and clustering data at a high frequency [7]. The adjusted weight and threshold values of the connection between the neurons form probabilistic algorithm enabling the network to improve its prediction and responses.

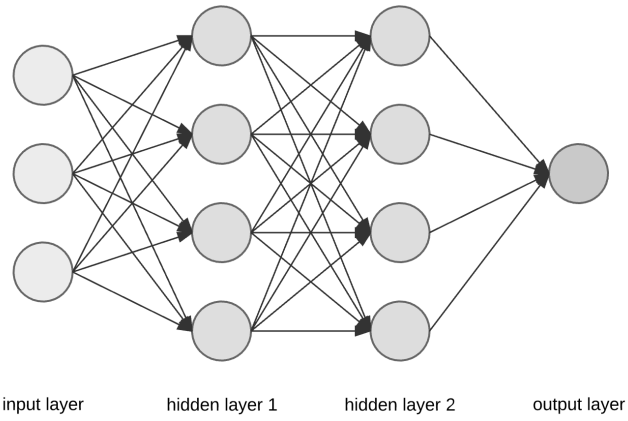

Fig. 1. Diagram of the layers within Artificial Neural Network (ANN)

Deep learning (DL) is a subset of machine learning employing the concept of Artificial Neural Networks to learn from the vast amount of data. The difference between ANN and deep learning is that 'deep' refers to the presence of numerous layers within the network. The architecture of the DL model consists of an input layer, multiple hidden layers, and an output layer [8]. The hidden layers enable the system to learn and extract complex patterns from large datasets much more efficiently compared to standard ANN with a singular layer. A defining feature of deep learning is the ability to learn feature representation automatically. Traditional machine learning models often require manual feature extraction to train the system. Deep learning models have the capability of automatically identifying relevant features from raw datasets, a feature that is known as feature learning or representation learning [8].

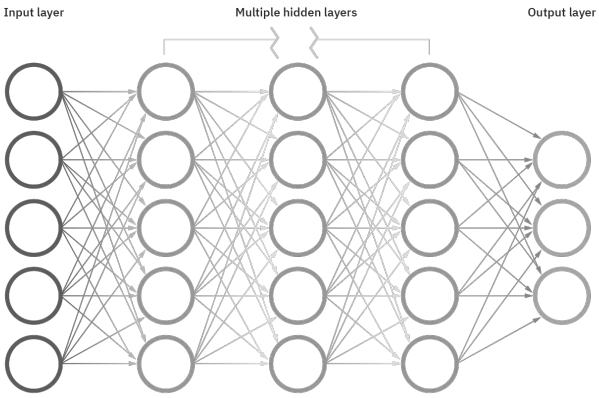

Fig. 2. Diagram of the layers within a Deep Learning Network

Through the development and intersection of machine learning, neural networks, and deep learning, new branches of artificial intelligence have been made applicable, these include natural language processing (NLP) and large language models (LLM). The development of these disciplines has allowed for the creation of sophisticated applications that enable machines to interact with humans in a meaningful and digestible manner [8]. Through research and development in deep learning and neural network models, the layers of interconnected nodes or artificial 'neurones' can identify and reproduce the complex patterns within the human language [7]. The field of natural language processing has been considerably advanced by the ability to automatically identify these significant characteristics among the vast quantity of raw data. These accomplishments have paved the way for large language models that can generate human-like text, fully understand context, and respond with human-like precision.

B. Natural Language Processing And Large Language Models

Natural Language Processing (NLP) is an intersectional field combining linguistics, computer science, and artificial intelligence to enable machines to understand, interpret, generate, and interact using human [9]. The main goal of NLP is to achieve a natural and seamless interaction between humans and computers by mirroring a typically human conversation. This functionality involves understanding and generating the intricate structures, semantics and nuances that form the human language. Natural language processing is developed using machine learning specifically deep learning. Deep learning models are utilized to identify patterns within large text datasets to learn how to predict subsequent words in a sentence, recognize speech, answer questions, and even generate human-like text [10].

Large language models (LLM) represent the cutting-edge in natural language processing (NLP), demonstrating unprecedented abilities in generating coherent, contextually relevant, and nuanced text. LLM are often characterized by the number of parameters within the model and the volume of data the models are trained with. The models are typically transformer-based models which are a specific type of artificial neural network architecture within the field of deep learning. The models are trained on vast volumes of text data, excelling at learning the patterns and structures of human language, and contributing to the various applications such as translation, summarization, content creation, and conversational [11]. Parameters within a LLM are the parts of the model that are learned from the historical volume of training data. The parameters include the weights and biases in the neural network that determine how input data is transformed into output data. Currently, the most prominent models within the field of LLM are GPT-3 by OpenAI and LaMBDA by Google [11].

C. ChatGPT-3

ChatGPT is an artificial intelligence chatbot developed by OpenAI to generate human-like text and facilitate engaging conversations. The advancement in ChatGPT has propelled artificial intelligence to the forefront of the technological revolution of the 21st century, causing a global ripple effect due to its proficiency and adaptability in generating creative and context-appropriate responses. The chatbot is built using OpenAI's foundational Generative Pre-Trained Transformer model, a type of large language model built for the conversational application. The most recent version of Gpt-3's Generative Pretrained Transformer Model represents a paradigm shift in transformer-based models that uses machine learning to generate contextually relevant and linguistically coherent text responses [11]. The model employs a massive collection of text and code from the internet, books, articles, and conversations, predated

before 2021 in order to train and utilise the model to generate responses that are grammatically correct and contextually valid. The current version of ChatGPT-3 contains 175 billon parameters, making it the largest LLMon the market with the largest quantity of historical data used to train it [11].

D. Bard

Bard is an artificial intelligence chatbot that's been developed by Google as a competitor to the popular ChatGPT. Due to the widespread success of OpenAI's ChatGPT, Google's Bard has emerged as a formidable competitor, demonstrating the undeniable interest and innovative enthusiasm encircling AI language models. Bard is built using Google's proprietary Language Model for Dialogue Applications (LaMBDA), a type of LLM which also utilised the transformer-based model architecture. Bard is comprised of 137 billion parameters learned from 1.57 trillion text datasets from the internet, literature, and code [12]. While ChatGPT's knowledge base is constrained to data before 2021, Bard's training data is constantly up-to-date, utilising the most recent internet-derived text. This real-time data aggregation provides Bard with a competitive advantage by allowing it to generate contextually accurate and consistent responses with the most recent global narratives and events.

E. AI Implication on Society

The rapid adoption of conversational artificial intelligence (AI) chatbot has shifted how we engage with information and technology as it becomes more pervasive within our daily lives. The flexibility of the chatbot powered by AI is apparent in its broad array of applications. These include automating customer service interactions, creating engaging content, assisting in language translation, and even developing coding projects. The versatility of the AI chatbot has been demonstrated to be endless by researchers and developers. Despite the technology sophistication and capabilities, there are notable areas of concerns that have emerged with its broad application. These concerns are required to be comprehensively understood and addressed for the secure adoption of AI within today's society. The section will explore the negative social, economic, and technical implications of AI within society.

1) Social Impact

The emergence of artificial intelligence (AI) chatbots calls for a comprehensive analysis of the potential societal implications. A notable concern stems from the potential for these AI chatbots to inadvertently propagate misinformation. The algorithms and training data connected to these LLM can generate misleading or inaccurate information, consequently leading to potential deception of users. AI chatbots can be manipulated for malicious purposes, including the distribution of propaganda, targeted deception, and the manipulation of public opinion [13]. AI models can unintentionally contain bias and discrimination, these traits can also be formed from the large dataset that may contain biased or discriminatory information. The ramifications of bias, misinformation and manipulated information via AI chatbots could be far-reaching given their widespread popularity. Incorrect information has the potential to distort and manipulate public opinion, amplifying the risks associated with this tool.

The convenience and capabilities provided by AI systems can lead to an over-reliance on technology. This overdependence might inadvertently diminish society's ability to think creatively and critically, as well as problem-solving. Such development may lead to social issues as these essential skills are undervalued and underdeveloped. Inversely, a digital divide could emerge from individuals or communities that lack access to this advanced technology or the requisite digital literacy to utilise it [11]. While the technical efficiency and effectiveness of AI systems are beneficial, individuals who lack the necessary knowledge or resources to leverage these technologies may be left at a substantial disadvantage. This technology gap can cause a form of digital handicap, as those unable to harness AI's potential may find themselves increasingly marginalized in a world that is progressively reliant on such technologies [11]. This can cause a growing digital divide within society, thereby leading to an array of negative social implications.

2) Economic Impact

The rapid advancement of artificial intelligence (AI), especially in the field of chatbots, has introduced considerable economic uncertainties that warrant thorough investigation. As AI continues to excel in tasks traditionally performed by humans, the looming threat of job displacement necessitates urgent attention [14]. AI has the potential to replace numerous professions, including customer service, content creation, and data entry. This shift towards automation will lead to increased job displacement, potentially triggering a rise in unemployment rates and instigating far-reaching societal changes. These jobs are not only limited to lower-skill jobs, as AI has the potential to automate tasks within high-skill jobs such as programming as well, leading to widespread job insecurity amongst all occupations [15]. Moreover, the advent of AI can intensify wealth disparity within society. As businesses adopt AI technologies to boost efficiency and lower costs, the profits often concentrate among company shareholders, upper management, and high-skilled staff involved in AI development or implementation [15]. Coupled with the dwindling availability of jobs in the market, this concentration of wealth may accentuate socio-economic divisions, thereby spawning broader economic challenges.

3) Technical Impact

Artificial intelligence chatbots have progressively accomplished numerous technological advances and capabilities, however technical challenges and implications must be addressed to integrate these technologies safely into society. AI chatbots oftentimes produce an accurate response, but some instances need to be taken into consideration when wrongful information is presented. The usage of training data can cause an instance in time where the system may reproduce an error or bias, which can lead to unexpected or undesired outputs [16]. There are many unknown factors on the limitation of the system when processing sensitive and controversial topics, these factors are required to be tested to determine the reliability of AI chatbots. Misinformation is another concern that can occur

due to the complexity of human language and the validity of the dataset utilized to train the models. In the case where an AI chatbot produced a response that either contains bias or misinformation, this can have serious ramifications within society depending on the information. Hence, addressing these issues, especially the propagation of bias and misinformation, model reliability, and lack of interpretability, is vital for the responsible and effective use of these systems. The growing complexity of the LLM has further the concerns of model interoperability due to the billions of parameters that are within these systems [9]. The transparency on how these models can evaluate the input and generate the output can help with debugging issues such as misinformation or bias within the system. Due to the complexity of these LLM, transparency is an issue within the AI application which limits the ability to trust and validate the models outputs.

Maintenance cost is also an underlying factor which needs to be considered as these AI models require cutting edge hardware to operate and maintain these AI applications. LLM require training and fine tuning to constantly stay up to date with the latest information whilst being consistent. The cost of maintaining these systems requires a sizeable amount of graphics cards (GPU) that are forever operating to calculate the computation algorithm, this can result in high cost in electricity and hardware [9].

III. Methodology

This article aims to rigorously investigate the capacity of two advanced artificial intelligence language models, ChatGPT-3 and Bard within a debate setting. The construct of this methodology will be deconstructed into debate structure, evaluation process, data acquisitions and analysis procedure to define the underlying procedure for this research investigation.

A. Research Design

The main objective of this research is to investigate the performance and limitations of the response generation capabilities within two of the most notable LLMs, ChatGPT-3 and Bard. This is accomplished by employing a research environment designed to facilitate testing and structured determination of research outcomes.

The research necessitates the development of an AI-based tool to facilitate the debate between ChatGPT-3 and Bard. Testing is initiated using the application, which consists of a series of structured LLM debates in which participants engage in discourse and discussion on a given topic. The topics have been meticulously selected and categorised into four groups: factual, optional, planning, and problem-solving. Each of these categories assesses a distinct facet of the AI models' capabilities: factual information comprehension, preference and opinion formulation, strategic planning, and complex problem-solving, in that order.

Each language model will have identical opportunities to present an argument, rebuttals, and counter arguments, so that the debate will closely resemble an actual human debate. The objective of replicating a debate structure is to rigorously test the model's capacity to comprehend context, make decisions, develop explainability, and identify ethics and bias in a generated response. Through debating contentious issues, the models will be evaluated for their weaknesses and strengths.

The response generated by the LLM during the debates will form the primary data for this research article. The purpose of this study is to assess the responses based on several important criteria, including coherence, relevance, persuasiveness, and empirical accuracy. The primary objective is not only to determine which AI language model performs better, but also to comprehend the strengths and limitations of each in the context of the contentious issue under discussion. This experimental research design will allow for a comprehensive examination and comparison of ChatGPT-3's and Bard's argumentative capabilities, providing a comprehensive understanding of the current state of argumentative discourse in AI language models.

B. Debate Setup

The debate setup within this research is designed to ensure a fair and comprehensive comparative analysis of the argumentation capabilities of ChatGPT-3 and Bard. These two language models are positioned to mimic debate participants, and their responses will be collected as data for analysis. This section will demonstrate the structure of the debate in order to ensure a fair and comprehensive debate amongst the large language models.

1) Debate Format

The format of the debate involves a topic of discussion and an established sequence of turns for argument presentation and rebuttal. The structure is designed to mirror common debate procedures and provide an authentic context for evaluating the AI language model’s argumentative capabilities. Each debate consists of five rounds, allowing each model multiple opportunities to present their viewpoint, respond to other models’ arguments and provide counterarguments. The sequence of turns arecarefully managed to ensure fairness and a comprehensive exploration of each topic.

2) Debate Topics

The debate will consist of twenty debate questions which are selected across four categories- factual, optional, planning and problem-solving. The diversity of topics is intended to test the adaptability and ability of AI language models to manage various query types and potential biases. The topics can be demonstrated in Table 1, 2, 3 and 4 of this research article below.

TABLE I. FACTUAL DEBATE TABLE

Debate Table	
Debate Category	***Debate Topic***
Factual	Climate Change: Is there a scientific consensus that it's driven primarily by human activities?
Factual	Vaccinations: Is there a scientific consensus that they prevent serious diseases?
Factual	GMOs: Is there a scientific consensus that they are safe for human consumption?
Factual	Artificial Intelligence: Do facts and figures show it's leading to job automation?
Factual	Evolution: Is there a scientific consensus that it is a factual explanation for the diversity of life on Earth?

Factual questions for the debate

TABLE II. OPTIONAL DEBATE TABLE

Debate Table	
Debate Category	***Debate Topic***
Optional	Remote Work vs Office Work: Which promotes better work-life balance?
Optional	Streaming Services vs Traditional Cinema: Which provides a superior movie-watching experience?
Optional	Social Media: Is it a beneficial tool for communication or a detrimental influence?
Optional	Digital Art vs Traditional Art: Which medium allows for greater creativity and expression?
Optional	Classical Literature vs Modern Literature: Which offers greater insights into human nature and society?

Optional questions for the debate

TABLE III. PLANNING DEBATE TABLE

Debate Table	
Debate Category	***Debate Topic***
Planning	Sydney Central Congestion: Would expanding public transportation or promoting remote work be the most efficient method to alleviate the issue?
Planning	Climate Change Mitigation: Is investing in renewable energy or implementing carbon capture technology a better strategy?
Planning	Aging Population: Should the focus be on improving healthcare for the elderly or encouraging higher birth rates?
Planning	Internet Accessibility: Should the priority be on expanding high-speed internet infrastructure or promoting affordable data plans?
Planning	Urban Green Spaces: Is it better to create smaller green pockets throughout the city or focus on larger central parks?

Planning questions for the debate

TABLE IV. PROBLEM-SOLVING DEBATE TABLE

Debate Table	
Debate Category	***Debate Topic***
Problem Solving	Plastic Waste: Is recycling or reducing production the more effective solution?
Problem Solving	Cybersecurity: Is the better approach offensive (proactive attacks to disrupt threats) or defensive (strengthening systems against attacks)?
Problem Solving	Poverty: Is the solution primarily in wealth redistribution or economic growth?
Problem Solving	Global Water Crisis: Would desalination or better water management solve the issue more effectively?
Problem Solving	Overpopulation: Is it better addressed through policy interventions or technological advancements?

Problem solving questions for the debate

3) AI Language Model Context Setup

Before each debate, an initial statement is presented to both language models to form the context. The purpose of this setup is to inform the AI language model about the debate structure and engage the models in a debate that abides by the rules of argumentative discourse. This provides the model with a clear understanding of the required task, contributing to a more controlled and meaningful comparison of the model's argumentative capabilities. The initial statement setup is in Figure 3:

"*You will be participating in a five-round debate. You are required to present and defend your arguments, respond to the other model's points, and provide relevant counterarguments*."

Fig. 3. Intialisation statement for setting up the models.

C. AI Debate Application

The implementation and execution of this article's research require the construction of a custom artificial intelligence debate tool utilized for the sole purpose of executing the research testing within a controlled environment. The application serves as a cornerstone of the research process, enabling the structure interaction of the AI models ChatGPT-3 and Bard, management of the debate rounds and the seamless collection of the responses generated by the AI language models.

1) Application Design

The design of the AI debate application is centred on facilitating and simulating real-time debates between the two notable AI language models (ChatGPT-3 and Bard). The application is composed of a robust Backend API Server built using Node.js for scalability and efficiency in handling asynchronous operations within the debate. The Backend API Server is developed to handle a POST request from the user to process the specified debate topic and AI model options.

This application is intentionally designed without a frontend user interface, focusing solely on backend functionalities to prioritize research objectives and expedite the development process. The decision is motivated by the

need to quickly set up and run multiple debates, collect data, and analyze results. The AI debate application has been architected with a clear separation of concerns, and the backend-focused design means that a front-end user interface can be readily added in future iterations. The application is then integrated with the API services of ChatGPT-3 and Bard to enable the engagement of the LLM within the debate. The architecture of the AI debate application can be demonstrated in Figure 4.

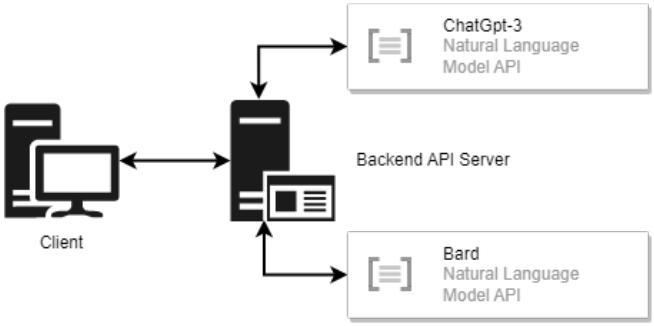

Fig. 4. Diagram of the AI Debate Application Design

2) Language Model Integration

The application seamlessly integrates the API services of ChatGPT-3 and Bard. The integration is vital for enabling the engagement of these models in debates and ensuring the successful retrieval of their generated responses. The integration of the API services for each of the AI language models is built as modules to enable the reusability of the model for asynchronous operations. Modules are initialised with a specified conversation identification number (ID) that is randomly generated for the objective of having a controlled debate. Conversations are recreated every debate, to prevent data persistent from the last debate session. The AI language models initially set up with the statement provided in Figure 3 add context to the models for the debate.

The language model is implemented to enable the debate to run in a sequence of turns mimicking as demonstrated in the debate formation section. The implementation involves a For Loop where the models each take turns providing an argument and counter-argument for five rounds. The architecture of the application can be demonstrated in Figure 5.

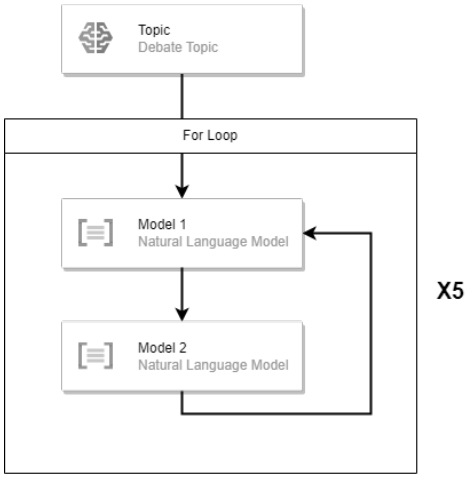

Fig. 5. Architecture of Debate AI application

The debate topic and generated response from the AI-Language models are automatically formatted and saved into a text file for data collection and analysis. The application data is logged and formatted using the Winston.js package library.

3) Application Functionality

The primary objective of the application functionality is to orchestrate and automate the process of the AI debate for the research article in a controlled setting. For each debate, the application initiates the conversation by dispatching the topic and the initial statement from Figure 3 to the participating AI models.

Beyond facilitating the flow of conversation, the application is programmed to maintain the order of turns and monitor the completion of rounds. Each turn is precisely timed and sequenced to maintain a fair debating platform for the AI models. The responses from the AI models are retrieved by the application, stored within a text file, and promptly displayed on the UI. This ensures that every argument and counter argument made by the models are collected for subsequent analysis, allowing for a comprehensive evaluation of their debate capabilities.

D. Evaluation

The evaluation section within this research article will assess the parameters for determining the quality and effectiveness of the AI language model responses. This process incorporates structured qualitative and quantitative methodologies to produce a comprehensive assessment. The criteria for evaluating the results include relevance, coherence, persuasiveness and factual accuracy of the AI language model's response. In addition, the ability of the AI language model to provide unbiased and nondiscriminatory arguments will be evaluated, reflecting the emphasis on minimising the impact of these negative characteristics. Due to the impossibility of specifying computational numeric values within the human language and responses, the parameters will be evaluated manually. The human evaluator will assign a numeric rating to each of the evaluation parameters in order to produce a quantitative statistical evaluation for the research.

1) Qualitative Evaluation

The qualitative evaluation involves human evaluators analysing the AI responses in terms of their argumentative qualities. The responses are gauged based on their logical soundness, relevance to the topic and persuasive strength of the arguments made by the AI models. In addition, the qualitative evaluation will also assess whether the models remain within the ethical boundaries of a debate, avoiding any forms of unexpected results that are beyond the debating scope.

2) Quantitative Evaluation

The quantitative evaluation involves the application of metrics to measure the various aspects of the AI model's response. This includes having the human evaluator determine a score for the relevance of the response to the debate topic, verifying the factual accuracy of the statements

made by the models, and assessing the coherence of the model's argument through grading the response. These evaluations are integral to determine the model's ability to generate logical consistent, factually correct and contextually relevant responses.

3) Bias Evaluation

The bias evaluation entails evaluating the responses for any possible biases, which reflect the social, cultural, political, or ideological predisposition inherent in AI language models. This assessment aims to identify and comprehend any potential bias within an AI language model that could corrupt the response. This is a significant concern within the language paradigm as the primary instrument for disseminating information. A human evaluator will assess the response's prejudice in order to ascertain its degree of bias.

E. Data Acquisition and Analysis

This data section of this research article will determine the type of data collected and the method used for analysing the debate surrounding the AI language model. The data for this research article can be found within the GitHub repository in Reference [17].

1) Data Acquisition

The responses generated by the AI language models will be automatically recorded through the usage of a proprietary AI debating application built for this research. The primary data collection will be the topic and the responses generated by ChatGPT-3 and Bard within each of the debates. These responses will automatically be recorded and stored for subsequent analysis. The AI debate research application tool's source code can be located in the repository on GitHub in Reference [18].

2) Data Analysis

The data analysis includes qualitative, quantitative, biased, and model evaluations of the resultant data. Human evaluators assess the responses' relevance, coherence, persuasiveness, and factual authenticity in accordance with the specified evaluation criteria. Human evaluators will use a predetermined rubric to ensure evaluation consistency. The quantitative analysis is assigned a rating by a human evaluator based on the evaluation measures. These measures include relevance (how closely the response relates to the debate topic), coherence (logical consistency and flow of the response), and factual accuracy (consistency of the response with verified factual information). These metrics help to provide an understanding of the AI language models' capabilities and performance within a debate. These data analyses provide valuable insight into the effectiveness of AI language models and their capabilities within a structured debate, thereby facilitating future development.

IV. Result And Discussion

The result section provides the conclusion reached after a succession of debates between ChatGPT-3 and Bard. The application used to establish a regulated debate environment for this article's research can be found in the GitHub repository in Reference [18]. The application's raw data can be found in the GitHub repository at Reference [17]. The results are presented per debate category and include an analysis of the AI models performance based on relevance, coherence, persuasiveness, factual accuracy, and presence of biases.

A. Factual Analysis

TABLE V. Factual Debate Rating

Factual Debate Table					
Debate Option	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT x Bard	8	6	7	8	6
ChatGPT x ChatGPT	7	7	4	6	6
Bard x Bard	4	6	3	5	6

The debate topic is referenced in Table 1

TABLE VI. Factual Debate Model Performance

	Factual Debate Table					
Model	***Parameter***	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT	Relevance	7	7	6	7	7
	Coherence	7	6	5	7	8
	Factual Accuracy	8	7	7	8	8
Bard	Relevance	7	7	6	7	7
	Coherence	6	5	5	4	5
	Factual Accuracy	8	8	7	8	8

The debate topic is referenced in Table 1

The performance of the AI language models in factual debate topics varies dependent on the coupling of the models and the topic of discussion as shown in Tables 5 and 6. The factual debate rating in Table 5 demonstrates that the ChatGpt vs Bard model pairing exhibits consistently higher performance scores across all the topics compared to the debates involving identical models. This suggests that the diversity of AI language models could have resulted in a more thorough and well-rounded discussion of factual topics.

In contrast, debates involving identical models, whether it be ChatGPT vs. ChatGPT or Bard vs. Bard, resulted in lower scores, ranging from 3 to 7. This suggests that debates between identical models tend to be less dynamic, possibly because the models have similar reasoning and argumentation styles, leading to more agreement and less extensive creative exploration of the topic.

Through closer analysis of the raw debates from Reference [17], the results suggest that the AI models tend to struggle when interacting with identical models. This is possibly due to a cyclical pattern of agreement, limiting the potential for dynamic discourse and often leading to a deviation from the debate topic.

The Table 6 provides an in-depth indication of the large language models response measure in terms of relevance, coherence, and factual accuracy within the debated topics. Both the LLM demonstrated a similar degree of relevance rating in their responses, indicating that the models are

capable of accurately comprehending the topic and providing appropriate responses. This reflects the robust attribute of the natural language understanding capabilities within the current large language models.

In terms of coherence, ChatGPT generally outperformed Bard in this measure, achieving a higher score for four out of the five topic debates. This suggests that ChatGPT was more proficient in creating a logical flow in its arguments, connecting the various points more seamlessly to build a comprehensive narrative. The coherent response resulted in a more comprehensive narrative that was engaging the audience. Whereas Bard typically struggled with creating coherent responses due to the lack of creativity. Bard's responses, although factual, tended to appear more robotic and lacked the conversational fluidity that ChatGPT exhibited. The data, as shown in Reference [17], shows that Bard's responses often had a somewhat robotic tone, lacking the natural flow of a human-like conversation.

For factual accuracy, both ChatGPT and Bard achieved comparable scores, as depicted in Table 6. This result underlines the ability of these language models to accurately source and present information that aligns with the context of the debate. However, a critical limitation that should be noted in this context is the restriction of ChatGPT's knowledge base. The ChatGPT-3 model is trained with data predated before the year 2021, which imposes a significant constraint on the model's information. This limitation may manifest in its accuracy, particularly when engaging in debates on subjects that have undergone significant changes or developments after the cutoff date.

B. Optional Analysis

TABLE VII. OPTIONAL DEBATE RATING

Optional Debate Table					
Debate Option	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT x Bard	7	8	4	7	7
ChatGPT x ChatGPT	7	8	7	7	8
Bard x Bard	7	7	6	6	7

The debate topic is referenced in Table 2

TABLE VIII. OPTIONAL DEBATE MODEL PERFORMANCE

	Optional Debate Table					
Model	***Parameter***	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT	Relevance	7	7	8	7	7
	Coherence	7	7	6	7	7
	Factual Accuracy	8	7	7	8	8
Bard	Relevance	6	7	7	6	6
	Coherence	6	6	5	7	6
	Factual Accuracy	7	6	7	7	6

The debate topic is referenced in Table 2

The nature of optional debates is inherently subjective, requiring the AI language models to craft arguments around preferences and personal viewpoints. Such debates present a unique challenge, testing the adaptability and versatility of the models to form opinions and preferences beyond the factual discourse. As depicted in Table 7, ChatGPT exhibits consistent performance across most topics, while Bard's results show some variance depending on the specific pairing and debate topic. This outcome suggests that the nature of the debate subject and the opposing model can significantly influence a language model's performance.

Upon closer examination, Topic 3, "Social Media: Is it a beneficial tool for communication or a detrimental influence?" in Table 2, stood out with a notable decline in performance when ChatGPT debated against Bard. This lower score signals difficulties in navigating more nuanced or complex opinion-based subjects, or it may reflect an incompatibility in the interaction between these two specific language models on this topic. Bard exhibited a consistent but inferior performance rating metric, which reflects the difficulty of tackling opinion-centric topics. This may be because Bard's model design and training data are more optimised for conveying factual information.

In contrast, Bard's performance was more inconsistent, with generally lower scores for relevance and coherence while maintaining a moderately high score for factual accuracy. These discrepancies highlight the difficulties that AI language models encounter when navigating subjective and opinion-based discussions. This highlights areas for potential development, particularly in terms of a model's ability to construct convincing and coherent arguments on topics that are opinion centric.

Further analysis of the raw data within Reference [17], demonstrates that within optional topics debates, ChatGPT is more inclined to provide the more factually beneficial response based on the question. Whereas Bard would provide an ambiguous answer that openly allows the audience to formulate their own opinion on the topic through the provided evidence. This indicates how far language models have to go before they can formulate their own opinion on controversial topics. Although this may appear to be a negative characteristic, the absence of opinionated responses reduces bias within the response.

C. Planning Analysis

TABLE IX. PLANNING DEBATE RATING

Planning Debate Table					
Debate Option	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT x Bard	8	7	8	7	7
ChatGPT x ChatGPT	3	4	6	7	7
Bard x Bard	7	8	8	7	7

The debate topic is referenced in Table 3

TABLE X. PLANNING DEBATE MODEL PERFORMANCE

	Planning Debate Table					
Model	***Parameter***	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***

	Planning Debate Table					
Model	***Parameter***	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT	Relevance	3	4	7	7	7
	Coherence	4	4	6	6	7
	Factual Accuracy	3	3	8	7	7
Bard	Relevance	7	7	6	7	7
	Coherence	7	8	8	7	7
	Factual Accuracy	8	7	7	8	7

The debate topic is referenced in Table 3

Planning debates are centered around developing a strategic approach or proposing a solution to a complex issue or topic. The performance of the LLM in this category can help us understand their capability for strategic thinking, problem solving and planning capabilities.

The Table 9 presents the overall performance score for each of the model debates in the planning category. The model pairing between ChatGPT and Bard exhibited strong performance across all the planning debate topics. However, there was a noticeable drop in performance when ChatGPT is paid amongst itself for Topics 1 and 2 within Table 3. This demonstrates that the dynamic of the interaction between the two identical ChatGPT models is not capable of progressing through a high-quality debate on the controversial topics within the planning category. This may be attributed to the lack of planning capabilities within ChatGPT hindering the resulting debate. In contrast, Bard demonstrated to accel in these planning capabilities with consistently high scores both when paired with ChatGPT and with another model, reflecting its strengths in handling controversial strategic and planning topics.

A detailed breakdown of each model's performance over the response measures is provided in Table 10. Bard is demonstrated to have consistently well across the performance metrics, maintaining high scores in all measures in relevance, coherence and factual accuracy. This implies that Bard has a robust capability for processing and debating controversial planning topics, likely due to its vast training dataset encompassing a variety of topics and themes.

The finding, combined with the moderate to low scores in relevance and coherence in Table 10, highlights the limitations of ChatGPT's capabilities in the area of strategic planning and suggests an area for future improvement. The low-performance rating suggests an issue with the model's ability to strategise effectively in the absence of a contrasting AI model. The poor planning capabilities of ChatGPT can be demonstrated within the raw data in Reference [17]. The data reveals the lack of understanding and information within the ChatGPT-3 model, contributing to the poor performance within the strategic attribute. Improvement can be made by increasing the number and variety of datasets to enable the model to have a wider knowledge base for controversial planning topics.

D. Problem-Solving Analysis

TABLE XI. PROBLEM-SOLVING DEBATE RATING

Problem-Solving Debate Table					
Debate Option	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT x Bard	8	7	8	8	7
ChatGPT x ChatGPT	7	7	8	7	6
Bard x Bard	6	5	5	6	6

The debate topic is referenced in Table 4

TABLE XII. PROBLEM-SOLVING DEBATE MODEL PERFORMANCE

	Problem-Solving Debate Table					
Model	***Parameter***	***Topic 1***	***Topic 2***	***Topic 3***	***Topic 4***	***Topic 5***
ChatGPT	Relevance	8	7	7	8	8
	Coherence	8	7	7	7	6
	Factual Accuracy	7	8	8	7	8
Bard	Relevance	6	5	4	5	6
	Coherence	5	5	5	6	5
	Factual Accuracy	6	5	5	6	7

The debate topic is referenced in Table 4

Problem-solving debates typically pose significant challenges for AI language models as they are required to understand the context and propose solutions and arguments for complex issues that generally don't have singular answers. The performance of LLM in the problem-solving category provides insight into the models' capabilities to handle complicated discussions and offer innovative solutions to problems.

The Table 11 illustrates ChatGPT's strong performance across a variety of topics regardless of the debating partnered model. The performance of the pair ChatGPT vs Bard and ChatGPT vs ChatGPT indicate the robust capabilities of the model to engage and solve issues within a problem-solving discussion. Comparatively, Bard's performance in the debate is demonstrated to be relatively underwhelming compared to ChatGPT's. The disparity of the results indicates a relative weakness in Bard's capabilities to handle problem-solving topics. The nature of the training dataset or the design of the model, which focuses on information propagation rather than innovative responses, may account for the subpar performance.

In the model performance analysis presented in Table 12, ChatGPT is demonstrated to have outperformed Bard within all measures – relevance, coherence and factual accuracy across all contentious topics. These high-performance ratings signify the capabilities of the ChatGPT model in problem-solving debates. The results reveal that LLMs can solve complex problems when trained with a diverse and vast dataset.

Bard's performance within Table 12 demonstrates consistently lower performance in all the topics which aligns with Table 11. The results highlight the need for further improvements in Bard's model design or training approach to enhance its problem-solving capabilities to solve issues.

E. Model Evaluation

The conduction of debate encompassing four diverse categories of contentious topics provides an avenue to evaluate the capabilities of the AI language models, ChatGPT and Bard. The thorough analysis of the raw data within Reference [17] exposes the qualitative distinct characteristics of the models in generating responses to complex debating questions.

The characteristics of Bard within Reference [1], reveal a distinctive tendency towards providing fact-oriented responses. This is demonstrated to be consistent with the underlying architecture which prioritizes the collation and synthesis of information from Bard's voluminous training dataset. Although this approach towards the language model design entails a higher degree of accurate information, the limitation of the model's design includes the lack of creativity, fluidity and capacity to adapt to context changes and provide innovative human-like responses. Hence although Bard may excel in a context that prioritizes factual accuracy, the model will struggle in replicating a more conversationally dynamic environment.

ChatGPT offers a more well-rounded performance as it maintains a fine balance between delivering factually accurate and conversationally fluid responses. The model's design is to replicate human-like conversational abilities, which is evidenced by the model's performance within Reference [17]. The model was able to deliver factually accurate information whilst engaging in a dynamic interaction within the debate. The model's responses exhibited traits that were more conversational and engaging and incorporated the creativity and traces that replicated human traits. The negative aspects of this creativity can ultimately lead to the introduction of biased opinions, which dilutes the information contained within the responses. As the ChatGPT model is presently used to disseminate information, these biased opinions can have a profound effect on society.

Overall, these models both have different qualitative characteristics which adversely affect the LLM's responses. The models both have varying strengths and weaknesses that are demonstrated enabling the models to have a competitive edge over the other. Based on the model analysis, ChatGPT is overall a more well-rounded model which can provide responses that are conversational fluid replicating human response.

V. Conclusion

This research article performs a comparative analysis of the most notable LLMs ChatGPT-3 and Bard through debating controversial topics. The results demonstrate both the varying strengths and weaknesses of the models. ChatGPT-3 performed significantly better in most categories including factual, optional and problem-solving debates. These categories signify the attributed capabilities of the ChatGPT responses. The marginal difference between the performance of the two models was comparatively close demonstrating the progression of LLM technology in today's society.

The performance measure of the responses between both models revealed that ChatGPT-3 performed marginally better in terms of a relevant, coherent and factually accurate response. Similarly to the category evaluation analysis, the difference between the performance of the two models was minuscule. Both the LLMs generated contextually valid responses within the controversial scenarios provided. There are instances in which the models may deviate from the contextual topic; however, this is primarily due to the design of the model or the trained data inhibiting the LLM response. The models are far from perfection as each have demonstrated qualities and attribute that far exceed the other. Hence, the need for further improvements in LLMs design and training approaches are required to enable the models to respond to any potential contentious queries.

References

[1] Bourgais, A., & Ibnouhsein, I. (2022). Ethics-by-design: the next frontier of industrialization. AI and Ethics, 2(2), 317-324.

[2] Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, 46-60.

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

[5] Vedapradha, R., Hariharan, R., & Shivakami, R. (2019). Artificial intelligence: A technological prototype in recruitment. Journal of Service Science and Management, 12(3), 382-390.

[6] Bell, J. (2022). What is machine learning?. Machine Learning and the City: Applications in Architecture and Urban Design, 207-216.

[7] Sharma, V., Rai, S., & Dev, A. (2012). A comprehensive study of artificial neural networks. International Journal of Advanced research in computer science and software engineering, 2(10).

[8] Nielsen, M. A. (2015). Neural networks and deep learning (Vol. 25, pp. 15-24). San Francisco, CA, USA: Determination press.

[9] Jurafsky, D., & Martin, J. H. (2019). Vector semantics and embeddings. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 270-85.

[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[12] Murugesan, S., & Cherukuri, A. K. (2023). The Rise of Generative Artificial Intelligence and Its Impact on Education: The Promises and Perils. Computer, 56(5), 116-121.

[13] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. First monday, 21(11-7).

[14] Chui, M., Manyika, J., & Miremadi, M. (2016). Where machines could replace humans-and where they can't (yet).

[15] Bessen, J. (2019). Automation and jobs: When technology boosts employment. Economic Policy, 34(100), 589-626.

[16] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

[17] To, William. (2023, May 29). Debate Research Raw Data. [GitHub repository]. Retrieved from https://github.com/willyyto/DebateAITool/tree/main/data

[18] To, William. (2023, May 29). Debate Research Tool. [GitHub repository]. Retrieved from https://github.com/willyyto/DebateAITool/tree/main/backend