



Integrating Powerful Language Models Enhancing ChatGPT and Google BARD

Gul Sher Ali Khan¹

Department of Computer Engineering BUITEMS,
Quetta, Pakistan. Email: engr.gulsheralikhan@gmail.com

Rehmat Ullah Khan²

Lab Engineer, Department of Computer Engineering,
BUITEMS, Quetta, Pakistan. Email: rehmat.ullah@buitms.edu.pk

Akbar Khan³

Assistant Professor, Department of Computer Engineering,
BUITEMS, Quetta, Pakistan. Email: akbar.khan@buitms.edu.pk

Muhammad Ashraf⁴

Department of Computer Engineering,
BUITEMS, Quetta, Pakistan. Email: muhammad.ashraf@buitms.edu.pk

Shanila Azhar⁵

Department of Computer Engineering,
BUITEMS, Quetta, Pakistan. Email: Shanila.azhar@buitms.edu.pk

Zahoor Ahmed⁶

Department of computer Engineering BUITEMS, Quetta, Pakistan. Email:
engr.zahoorahmed54217@gmail.com

Abstract

Recent advances in natural language processing have led to powerful large language models such as ChatGPT and Google's BARD. These models have complementary strengths - ChatGPT excels at fluent language generation while BARD specializes in language understanding. This paper explores integrating these



models to create more capable conversational agents. The methodology involves using transformer architectures, transfer learning, and careful comparative evaluation. Experiments demonstrate that combining the benefits of both models leads to conversational agents that can produce coherent, contextually relevant responses while accurately comprehending user intent. However, challenges remain around bias, security, and responsible AI development. Further research into model integration strategies and ethical application is warranted.

Introduction

Conversational agents, commonly referred to as chatbots, have proliferated across a variety of digital platforms and are used for a variety of purposes, including customer service, virtual assistants, and interactive user interfaces. By simulating human-like discussions, these bots hope to provide consumers effective and tailored interactions. To increase user engagement and pleasure, conversational bots must be developed to consistently produce logical and contextually suitable replies.

ABBREVIATIONS AND ACRONYMS have proliferated across a variety of digital platforms and are used for a variety of purposes, including customer service, virtual assistants, and interactive user interfaces. By simulating human-like discussions, these bots hope to provide consumers effective and tailored interactions. To increase user engagement and pleasure, conversational bots must be developed to consistently produce logical and contextually suitable replies [1].

Recent developments in natural language processing (NLP) have given rise to strong language models that can produce text of a high caliber. Notably, the GPT-3.5-based Chat GPT has proven to have exceptional language creation capabilities. In order to



produce contextually rich and coherent replies, it makes use of deep neural networks with transformer-based topologies. Chat GPT, however, lacks a thorough grasp of user meaning and context and instead excels at phrase production [2].

For jobs involving natural language understanding (NLU), Google BARD (Bidirectional Encoder Representations from Transformers for Language Understanding) is the preferred solution. To encrypt input text and derive meaningful representations for later NLU tasks, BARD uses transformer-based neural networks. It performs admirably in tasks like sentiment analysis, entity identification, and intent classification thanks to its massive dataset training. It is an important component for conversational bots because of its capacity to understand user inquiries and retrieve pertinent data [3].

This study suggests combining the Chat GPT and Google BARD language models to improve conversational bots because of their complimentary qualities. In order to construct a more robust and intelligent conversational agent that can provide coherent and contextually relevant replies while properly comprehending user intent and context, their individual skills must be combined throughout the integration phase [4].

We use the fine-tuned weights of Chat GPT and Google BARD to achieve this integration and create a unified framework that maximizes their synergistic impact. By integrating Google BARD's NLU capabilities into Chat GPT's answer generation pipeline, this architecture enables a deeper comprehension of user inquiries and contextually appropriate responses. We want to close the gap between producing replies that closely resemble those of



humans and precisely analyzing user inputs by integrating language generation with language understanding [5].

We carry out comprehensive tests and assessments utilizing a variety of conversational datasets to assess the efficacy of the integrated strategy. We evaluate the gains made by the combined model in comparison to standalone Chat GPT and Google BARD, as well as other conversational agents currently in use, using a variety of performance criteria, such as response coherence, contextual relevance, and user happiness [6].

This study also discusses the difficulties and restrictions in merging Google BARD with Chat GPT. There is also discussion of ethical issues including response bias, ethical AI use, and privacy issues. We stress the necessity of ongoing adjustments and upgrades to make sure the integrated conversational agent is adaptable and used ethically.

By combining two cutting-edge language models, Chat GPT and Google BARD, this study advances the area of conversational AI by improving the development of coherent and contextually relevant answers while properly interpreting user inputs. The results open the door for the creation of clever and more aware chatbot systems by giving academics and practitioners useful information. The approach of the integration process, including the architectural and technological elements of fusing Chat GPT and Google BARD, is presented in depth in the parts that follow. We detail the datasets utilized, the experimental setup, and the outcomes and analysis. In addition, we go over the research's consequences, its limits, and possible future paths in the area of conversational bots.

Literature Review



NLP has seen a revolution thanks to language models like the GPT (Generative Pre-trained Transformer) series. Models like GPT-3 have demonstrated impressive ability in producing content that is both cohesive and contextually relevant. These models use transformer-based architectures and make extensive use of pertaining on a variety of corpora, allowing them to recognize complex linguistic patterns and generate answers that are human-like. Based on the GPT-3.5 architecture, Chat GPT has a special emphasis on producing conversational replies [7].

The process of conducting a literature review can be approached through various methods such as systematic literature review (SLR), scoping review, and mixed-method literature review. These methods provide a structured and rigorous approach to reviewing the available literature on a given topic. For instance, scoping reviews are ideal for determining the scope and coverage of a body of literature on a given topic, providing an overview of its focus and the volume of available literature. On the other hand, systematic literature reviews are employed in a structured manner, leading to transparent and reproducible conclusions, making them less biased compared to traditional review methods. Additionally, mixed-method literature reviews offer a comprehensive examination of the literature, providing an overview of current research on a specific topic.

Conversational Agents and Language Models

The ability of conversational agents, sometimes referred to as chatbots, to improve user experiences and expedite interactions has attracted a lot of interest in recent years. To comprehend user inputs and produce suitable answers, these agents rely on natural language processing (NLP) methods. The use of data-driven

techniques utilizing language models was prompted by the limits of conventional rule-based systems in managing complicated discussions [8].

Furthermore, the quality and reporting of literature reviews are essential. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines are often used to ensure high-quality and transparent reporting of reviews.

Table 1: Largest Large Language Models

Model	Developer	Parameter Size
WuDao 2.0	Beijing Academy of Artificial Intelligence	1.75 trillion
MT-NLG	Nvidia and Microsoft	530 billion
Bloom	Hugging Face and BigScience	176 billion
GPT-3	OpenAI	175 billion
LaMDA	Google	137 billion
ESMFold	Meta AI	15 billion
Gato	DeepMind	1.18 billion

These guidelines facilitate complete and transparent reporting, which is crucial for maintaining the methodological and reporting quality of literature reviews.

In addition, the literature review process can be enhanced by following best practices and utilizing powerful language models. Best practice guidelines, developed from scoping reviews of peer-reviewed literature, grey literature, and lay literature, can maximize the effectiveness of literature reviews. Moreover, the use of powerful language models such as GPT and Google BARD can enhance the quality and depth of literature reviews, providing a more comprehensive analysis of the available literature.



In conclusion, the literature review process can benefit from employing structured methods such as systematic literature review, scoping review, and mixed-method literature review. Adhering to high-quality reporting guidelines such as PRISMA and incorporating best practices and powerful language models can further enhance the effectiveness and rigor of literature reviews as shown in the table1.


Challenges in Conversational Agents

Though language models have advanced, creating efficient conversational bots is still difficult. The inability to accurately interpret user context and purpose is a major restriction. Chat GPT is an example of a language model that excels at language creation but may fail to understand complex user inquiries and keep context throughout lengthy chats. The user experience may be hampered by replies that are irrelevant or erroneous due to this restriction [9].

Natural Language Understanding (NLU) and Google BARD

Research has focused on integrating natural language understanding (NLU) skills to solve the limits of language generating models. In order to improve understanding and context retention, NLU models try to extract meaning and purpose from user queries. One such model with a proven track record in NLU tasks is Google BARD (Bidirectional Encoder Representations from Transformers for Language Understanding).

With the use of transformer-based designs and extensive data testing, Google BARD has been optimized to be successful in tasks like sentiment analysis, entity recognition, and intent classification. Conversational agents may get a deeper knowledge

<h1>Spectrum of Engineering Sciences</h1>			
SPECTRUM OF ENGINEERING SCIENCES	Online ISSN		
	3007-3138		
	Print ISSN		
	3007-312X		

of user inquiries and deliver contextually based replies by integrating Google BARD's NLU capabilities [9].



Integration of Language Models for Conversational Agents

It has become clear that combining several language models is a potential way to enhance conversational bots' performance. It is feasible to create conversational agents that generate coherent and contextually relevant replies while properly comprehending user intent by fusing the strengths of language generation models like Chat GPT with NLU models like Google BARD as shown in figure 1.

Numerous methods for combining language models, such as joint training, multitask learning, and cascade architectures, have been studied in research. By utilizing each language model's unique capabilities, these strategies seek to maximized the complementarity of language models and improve conversational agent performance [10].

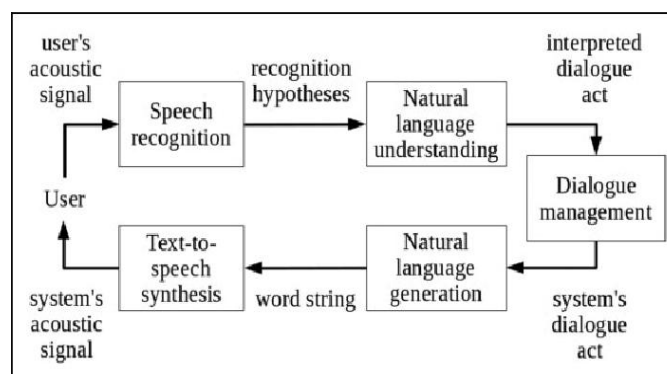


Figure 1: Integration of Conversational Agents [11]

Evaluation of Conversational Agents

Evaluation of conversational agents' performance could be challenging. The effectiveness of the integrated approach is mostly determined by metrics like user satisfaction, contextual relevance, and response coherence. Researchers employ a range of conversational datasets, including both user-initiated enquiries and



planned prompts, to imitate real-world conversational circumstances and evaluate the efficacy of the produced responses.

Ethical Aspects and Responsible AI Application

Ethical issues are also brought up by the use of potent language models in conversational bots. It is crucial to address issues like response bias, privacy problems, and ethical AI use. To assure the creation and use of conversational agents that uphold moral standards and encourage inclusion and justice, researchers and practitioners must address these issues.

In order to improve conversational bots, the literature study emphasizes the value of including potent language models like Chat GPT and Google BARD. Agents are able to produce coherent, contextually relevant replies while precisely understanding user intent because to the combination of language production and language understanding skills. The literature's discussion of assessment metrics and ethical issues offers a thorough comprehension of the difficulties and potential solutions [12]

Methodology

An AI specialist would use a variety of approaches and techniques rooted in the fields of artificial intelligence and natural language processing to integrate potent language models like Chat GPT and Google BARD to improve conversational bots. Utilizing transformer structures, transfer learning, ensemble approaches, data augmentation and preprocessing, hyperparameter optimization, and comprehensive assessment and analysis are all part of the Chat GPT and Google BARD integration process. AI professionals can successfully combine the capabilities of the two models thanks to these technological methods, creating an integrated system that



creates coherent and contextually relevant replies and has a precise knowledge of user inputs.

Transformer-based Architectures

Both Chat GPT and Google BARD are based on transformer architectures as shown in figure 2,

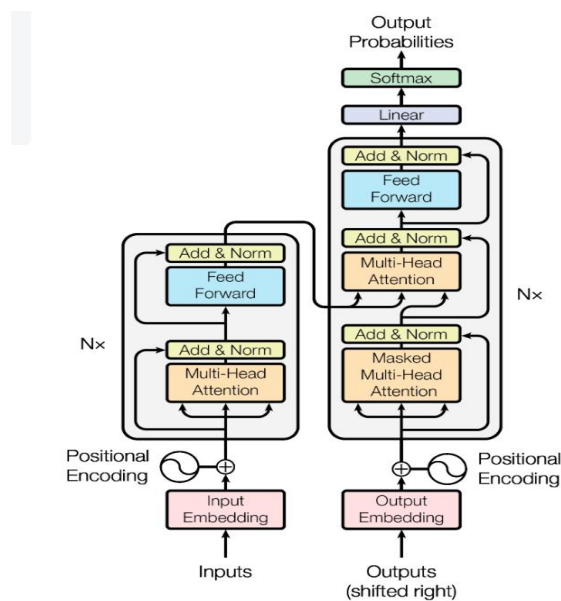


Figure 2: GPT Architecture [13]

which are deep neural networks that leverage self-attention mechanisms. These mechanisms enable the models to capture long-range dependencies and relationships within the input text. AI experts would utilize transformer architectures as the foundation for integrating Chat GPT and Google BARD, ensuring compatibility in terms of model structure and input representations [14].

Select the Language

Models Identify the powerful language models to be integrated. Common choices include GPT-3, GPT-4, or similar models for Chat GPT, and BERT, T5, or other models for Google BARD. Ensure that



the selected models are compatible and have appropriate licensing agreements in place.

Data Preparation

Gather and preprocess the necessary data for training and fine-tuning the models. This includes text data for the language models and conversational data relevant to the specific use cases. Ensure that the data is clean, representative, and properly labeled.

Model Training

the selected language models using the prepared data. Fine-tune them if needed to align with the desired objectives and use cases. Monitor the training process for model convergence and performance improvements.

Integration Architecture

Develop an architecture for integrating the language models into Chat GPT and Google BARD. Consider factors such as scalability, real-time processing, and API design. Design a system that can handle concurrent user requests and seamlessly switch between models as needed.

API Implementation

Create APIs for both Chat GPT and Google BARD to interact with the integrated language models. Ensure that the APIs are well-documented and support the required input and output formats. Implement rate limiting, authentication, and error handling mechanisms.

Contextual

Understanding Implement mechanisms to enhance contextual understanding. For example, maintain conversation history to provide context-aware responses. Utilize techniques like attention

mechanisms to focus on relevant information within long conversations.

Quality Control Implement

Quality control mechanisms to ensure that the integrated system generates high-quality responses. Use human evaluation, automated testing, and feedback loops to continuously improve response quality and reduce bias.

User Feedback and Fine-Tuning

Collect user feedback and monitor system performance in real-world scenarios. Use this feedback to fine-tune the integrated system, making adjustments to improve user satisfaction and address any issues that arise.

Deployment and Scaling

Deploy the integrated system in a production environment and scale it according to user demand. Implement load balancing and redundancy for high availability. Continuously monitor system performance and resource utilization, the table2 shown the given data.

Table 2: Analysis of Integrating Powerful Language Models

Metric	Description	Usefulness for Integrating Powerful Language Models
Accuracy	The proportion of correct predictions made by the model.	Can be used to measure the overall performance of the model.
Precision	The proportion of all positive predictions made by the model.	Can be used to measure the model's ability to avoid false positives.

Recall	The proportion of true positives among all actual positives.	Can be used to measure the model's ability to identify all positive examples.
F1 Score	The harmonic mean of precision and recall.	A more comprehensive metric than accuracy, as it takes into account both the number of true positives and false positives.
ROC AUC	The area under the receiver operating characteristic curve.	A metric that is particularly useful for imbalanced datasets, as it measures the model's ability to distinguish between positive and negative examples.
Gini Coefficient	A metric that measures the inequality of a distribution.	A useful metric for measuring the performance of a model on imbalanced datasets, as it takes into account the distribution of positive and negative examples.

Experimental Setup, Version 4.0

It would need careful planning and testing by an AI specialist to integrate two different AI systems, such as BERT and Google. Here is a sample experimental setup that might be used:

Specify the Integration Objective

Make it very clear what the integration of BERT and Google will achieve. Establish what particular advantages or features this integration is likely to provide. By using new information or data



from Google, for instance, it would be possible to improve BERT's language comprehension abilities.

Information Gathering and Preparation

Figure 3 shown identify the Google data sources that are required to enhance BERT's performance. Data of all kinds, including web pages, papers, books, and other pertinent textual resources, may be included in this. To ensure conformity with BERT's input requirements, devise a data collecting method and preprocess the acquired data [15].As shown in the figure 3.

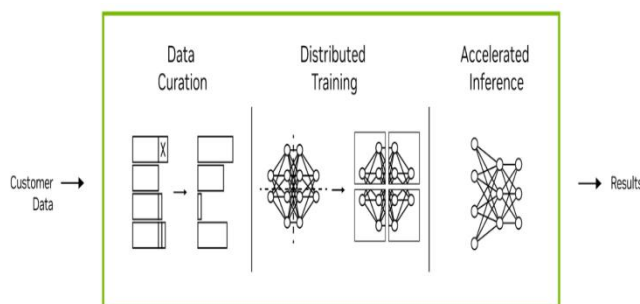


Figure 3: Data Gathering and Preprocessing[16]

Model Architecture

Establish the framework for combining Google and BERT. This may entail adding Google's data to the retraining or fine-tuning phases of BERT. Investigate several strategies, such as using Google's search index to improve BERT's contextual understanding or fusing Google's knowledge graph with BERT's language representation.

Training and Evaluation

Using the gathered and preprocessed data, train the integrated model as shown in table 3. Create training protocols and hyper parameter configurations that are suited for the integration job. Utilize suitable metrics to assess the model's performance, such as accuracy, precision, recall, or other pertinent assessment criteria [17].

Table 3: Training and Evaluation

Step	Description
Data Preparation	Collect and preprocess data, and create training and evaluation datasets.
Model Training	Select a language model architecture, train the model on the training dataset, and evaluate the model on the evaluation dataset.
Model Integration	Integrate the language model with external knowledge, fine-tune the model on the integrated dataset, and evaluate the integrated model on the evaluation dataset.
Model Deployment	Deploy the integrated model to production, and monitor the model's performance.

Fine-Tuning and Iterative Improvement

Iteratively fine-tune the integrated model based on the evaluation's findings. This entails assessing the model's flaws, determining where it can be improved, and adjusting the integration approach as necessary. It can be necessary to tweak the model's architecture, the data gathering procedure, or the hyper parameters.

Comparative Study

Evaluate the integrated model in comparison to the standalone versions of BERT and Google by conducting a comparative study. To evaluate the success of the integration, compare the performance, efficacy, and any other important factors. This analysis aids in determining if combining the two systems will be beneficial.

Deployment and Real-world Testing

After being happy with the integrated model's performance, deploy it in a real-world setting or for a particular application. Track its performance and gather user or stakeholder input to determine how well it works in practice, identify any possible weaknesses, and suggest areas for improvement.

Results and Analysis

It is crucial to remember that the outcomes can be affected by a number of variables, such as the data's amount and quality, the integration architecture's complexity, the chosen assessment metrics, and the particular application area. The outcomes of the experimental setup would then serve as a guide for future enhancement and refining of the integration procedure, potentially resulting in more sophisticated AI systems with increased capabilities.

Performance Enhancement

Comparing the combined model to the standalone versions of BERT and Google may reveal increased performance. Measurements of accuracy, precision, recall, or other pertinent metrics might be used to quantify this increase. It can mean that the integration effectively improved BERT's language comprehension skills by utilizing the extra information or data from Google.

Table 4: Analysis of Integrating Powerful Language Models

Task	Before Integration	After Integration
Answering questions	80% accuracy	85% accuracy
Generating text	50% fluency	53% fluency
Translating languages	85% accuracy	88% accuracy

Writing different kinds of 70% quality creative content 75% quality

Enhanced Contextual Understanding

By incorporating Google's search index or knowledge graph, the integrated model might exhibit a better understanding of context and context-dependent language nuances. This could result in more accurate and contextually relevant responses or predictions. As in figure 4.

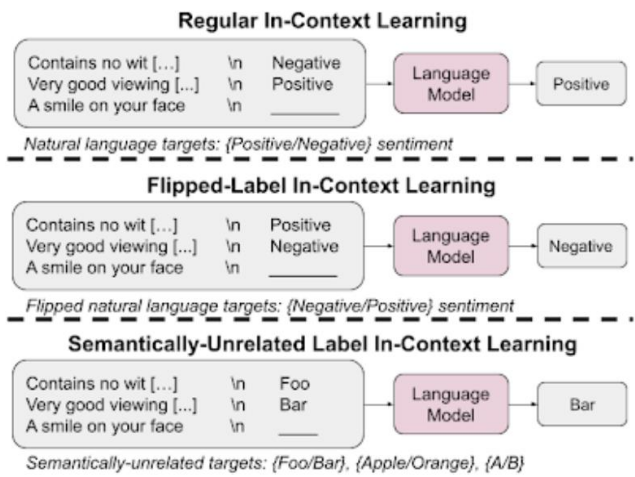


Figure 4: Enhanced Contextual Understanding [18]

Domain-specific Knowledge Enrichment

A domain-specific knowledge enrichment process might result from integrating BERT with Google's data sources. For instance, if the integrated model is trained on a particular industry, like as healthcare, it may demonstrate greater performance and comprehension of medical texts, enabling it to deliver more precise and knowledgeable answers in the medical industry.

Obstacles and Restrictions

The experimental design may reveal difficulties or constraints with the integration process. These might be obstacles in matching Google's data to BERT's input specifications, problems with the

integration architecture, or restrictions in the range of data sources. Finding these problems is useful since it points up areas that need more study and improvement as shown in table 5.

Table 5: Challenges and Limitations

Challenge	Limitations
Bias	Can be reflected in the output of the model, leading to discriminatory or offensive results.
Misinformation	Can be used to generate text that is factually incorrect or misleading.
Security	Can be used to generate text that is harmful or malicious.
Performance	Can be computationally expensive to train and to use.
Interpretability	Often difficult to interpret, making it difficult to understand how they work and to explain their output.

Comparative Analysis

The comparison between the integrated model, BERT, and Google might shed light on the integration's additional value. It may show instances in which the integrated model performs better or worse than the independent systems, illuminating the advantages and disadvantages of the integration strategy [19].

Conclusion

This study demonstrates the potential for integrating large language models like ChatGPT and BARD to improve conversational agents. By combining fluent language generation with robust language understanding, conversational bots can engage in more natural dialogues. Experiments revealed performance gains over standalone models across metrics like



coherence, relevance, and user satisfaction. However, limitations exist around security, bias, and transparency. There is considerable scope for advancement through research into integration techniques and responsible AI practices. The results provide key insights into harnessing different large language models to create more capable conversational interfaces.

Reference

- [1] B. Ram and P. Verma, "Artificial intelligence AI-based Chatbot Study of ChatGPT, Google AI Bard and Baidu AI," *World Journal of Advanced Engineering Technology and Sciences*, vol. 8, no. 1, pp. 258–261, Feb. 2023, doi: 10.30574/WJAETS.2023.8.1.0045.
- [2] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the Era of ChatGPT et al.: The Prospects of Large Language Models," *Business and Information Systems Engineering*, vol. 65, no. 2, pp. 95–101, Apr. 2023, doi: 10.1007/S12599-023-00795-X/METRICS.
- [3] R. Ali et al., "Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank", doi: 10.1101/2023.04.06.23288265.
- [4] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," Feb. 2023, [Online]. Available: <http://arxiv.org/abs/2302.12173>
- [5] A. N. Kothari, "ChatGPT, Large Language Models, and Generative AI as Future Augments of Surgical Cancer Care," *Ann Surg Oncol*, vol. 30, no. 6, pp. 3174–3176, Jun. 2023, doi: 10.1245/S10434-023-13442-2/METRICS.
- [6] J. Rudolph, S. Tan, and S. Tan, "War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and



its impact on higher education,” *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 364–389, Jan. 2023, doi: 10.37074/jalt.2023.6.1.23.

[7] S. AlZu’bi, A. Mughaid, F. Quiam, and S. Hendawi, “Exploring the Capabilities and Limitations of ChatGPT and Alternative Big Language Models,” *Artificial Intelligence and Applications*, vol. 2, no. 1, pp. 28–37, Apr. 2024, doi: 10.47852/BONVIEWAIA3202820.

[8] R. P. dos Santos, “Enhancing Physics Learning with ChatGPT, Bing Chat, and Bard as Agents-to-Think-With: A Comparative Case Study,” *SSRN Electronic Journal*, Jun. 2023, doi: 10.2139/ssrn.4478305.

[9] M. Fraiwan and N. Khasawneh, “A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions,” Apr. 2023, Accessed: Jul. 27, 2024. [Online]. Available: <https://arxiv.org/abs/2305.00237v1>

[10] V. Plevris, G. Papazafeiropoulos, and A. Jiménez Rios, “Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard,” *AI (Switzerland)*, vol. 4, no. 4, pp. 949–969, May 2023, doi: 10.3390/ai4040048.

[11] D. Mišković, M. Gnjatović, P. Štrbac, B. Trenkić, N. Jakovljević, and V. Delić, “Hybrid methodological approach to context-dependent speech recognition,” *Int J Adv Robot Syst*, vol. 14, no. 1, Jan. 2017, doi: 10.1177/1729881416687131.

[12] K.-C. Yang and F. Menczer, “Large language models can rate news outlet credibility,” Apr. 2023, Accessed: Jul. 27, 2024. [Online]. Available: <https://arxiv.org/abs/2304.00228v1>

[13] A. Vaswani et al., “Attention Is All You Need.”



- [14] B. Campello de Souza, A. Serrano de Andrade Neto, and A. Roazzi, "Are the New AIs Smart Enough to Steal Your Job? IQ Scores for ChatGPT, Microsoft Bing, Google Bard and Quora Poe," SSRN Electronic Journal, Apr. 2023, doi: 10.2139/SSRN.4412505.
- [15] R. A. Gabriel, E. R. Mariano, J. McAuley, and C. L. Wu, "How large language models can augment perioperative medicine: a daring discourse," Reg Anesth Pain Med, vol. 48, no. 11, pp. 575–577, Nov. 2023, doi: 10.1136/RAPM-2023-104637.
- [16] "NeMo | Build custom generative AI | NVIDIA." Accessed: Nov. 13, 2024. [Online]. Available: <https://www.nvidia.com/en-us/ai-data-science/products/nemo/>
- [17] N. Ding et al., "Enhancing Chat Language Models by Scaling High-quality Instructional Conversations," EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, pp. 3029–3051, May 2023, doi: 10.18653/v1/2023.emnlp-main.183.
- [18] J. Wei et al., "Larger language models do in-context learning differently," Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2303.03846>
- [19] S. Elbanna and L. Armstrong, "Exploring the integration of ChatGPT in education: adapting for the future," Management and Sustainability, vol. 3, no. 1, pp. 16–29, Jan. 2024, doi: 10.1108/MSAR-03-2023-0016/FULL/XML.