International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Large Language Models and Rule-Based Approaches in Domain-Specific Communication

Mr Pradeep Nayak[1],Suhas S[2],Sujal Shyam Bandekar[3],Tejaswini A S[4],Thulasi[5],Vidyashree[6]
Faculty,Department of Information Science And Engineering[1]
Students,Department of Information Science And Engineering[2,3,4,5,6]
Alva's Institute of Engineering and Technology,Mijar,Manglore,Karnataka,India

Email:thulasipoojary2005@gmail.com

Abstract: At the moment, artificial intelligence is causing a flurry once more. Across a range of natural language processing tasks, Generative Pre-trained Transformers (GPT) models demonstrate exceptional efficacy. These days, a variety of GPT models are frequently employed to increase productivity. Using services based on the GPT framework, developers create complex software solutions, graphic departments create art designs, and numerous other industries are following suit and incorporating these new toolkits into their operations. But in several fields of natural language processing, a straightforward approach is frequently more appropriate and successful than the large language models that are currently in use. We choose to examine and contrast the real-world applications of J-Large, one of the more well-known GPT solutions, with the straightforward rulebased model in this paper.We put it into practice. We included these two concepts into a private company's internal information system that focuses on customer communication in the gaming sector. The same dataset—a log of verbal exchanges over the previous two years in the system in question—was used to train both models.

I.Introduction

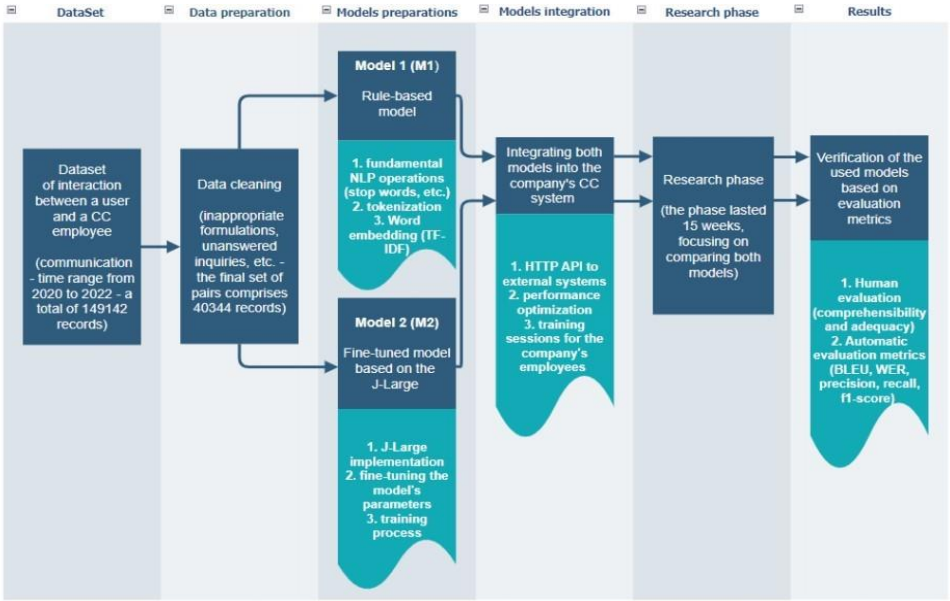
A collection of Generative Pre-trained Transformer (GPT) models that can provide high-quality natural language generation (NLG) has been produced by recent developments in Natural Language Processing (NLP). These models identify and understand the innate patterns and structures of language through a transformer architecture and intensive pre-training on large text corpora. As we move from GPT-1 to GPT-4, we see significant improvements in NLG's quality and capability. Notwithstanding these advancements, GPT models nevertheless have drawbacks that call for comparative research outlining their advantages and disadvantages. Even though there have been many studies comparing the performance of different GPT models [1], these typically focus on certain applications or domains and, because of their high demand, typically require operation via public cloud service.Furthermore, large data sets are needed to train these models due to their significant data requirements. Furthermore, medium- and small-sized institutions cannot maintain the operational costs of these software programs over the long run. These considerations led us to decide to compare the effectiveness of a GPT model and a rule-based model in the actual setting of an information system used by a corporate Customer Care (CC) department. Through this empirical comparison, we may investigate if GPT models are required for all chatbots or whether chatbots may perform particular jobs more efficiently using a different algorithmic concept that requires less data for training.In order to validate and compare a GPT model with a rule-based model in actual CC situations, we ran an experiment. Our six-month experiment involved a number of duties, such as gathering data from the previous two years within the project's host organization. In order to train a customized, fine-tuned model on the J-Large platform and our proprietary rules-based model, this approach produced a corpus of 40,344 question-answer pairs. Employees of the CC department received responses produced by both models during the experiment, and they were given the option to select one suitable response or reject both and provide their own

response in its place. We next examined the gathered information to ascertain how well both models produced appropriate client answers.Our experiment's goal was to ascertain whether both strategies were appropriate for realworld use in CC. We sought to comprehend the benefits and drawbacks of the two suggested fixes.

Given the difficulties in putting into practice a reliable large language model (LLM) like GPT and the associated expenses, we aimed to ascertain the degree to which a more straightforward approach, like a rules-based model, could completely replace such a system.

Figure 1 illustrates each step of the experiment, including training, fine-tuning, and model implementation. Section III: Materials and Methods provides a detailed description of each step. The prior correspondence between CC employees and clients throughout the preceding two years served as the foundation for our work.This served as the basis for our two developed models, which we then verified in a commercial company's actual environment.Our paper presents a viewpoint on language models in relation to real-world uses in the field of gaming industry customer communication. There are now a number of articles that compare LLMs to one another. The purpose of our paper is to respond to the query of whether a more straightforward model can take the place of LLMs in real-world applications. We contrast LLMs with the rule-based model, which is a more straightforward method. The outcomes of the actual use of both environments in business settings served as the basis for this comparison.

The paper is organized as follows. The second section provides a summary of the current status of research on the impact of window size and dimension size parameters. The third section describes the spam datasets used in the study, along with associated text pre-processing methods and text vectorization models. The fourth section provides a summary of the most significant findings. The final section of the article consists of the discussion and conclusions.



II.Related Work

These days, it's common for big businesses to aim to automate seemingly straightforward customer service tasks. Customers that frequently call customer lines attempt to obtain the same information, according to multiple analyses, and very few of them require a "custom" strategy to successfully satisfy their needs [2]. Whether a student is trying to find out when they can submit their application or a postal client is unsure of which service is ideal for mailing their parcel, the same formula may be used to apply to a variety of communication channels between an individual and any institution. It is well established that developing a technology tool to meet these demands is more cost-effective [3] than hiring hundreds of customer service representatives to chat or otherwise interact with individuals seeking specialized information.According to 2022 predictions, up to 90% of all customer-bank

communications in 2022 will be conducted by chatbots in certain industries, like banking [4]. Neural networks and their training for NLG purposes enable all of this.

A. Transformer Models' General Operation

Currently, a number of industry leaders generate distinct chat instances for various uses using trained neural networks. Regardless of whether it is Wu Dao 2.0, Jurassic-1, or GPT-4, we are still discussing algorithmically related solutions. The trained set and extra steps taken before the input is routed to the neural network and the result is shown to the user are where they differ most from one another.The term "pre-trained" neural networks refers to the fact that these networks were not designed with a particular situation in mind. Nonetheless, the models in question belong to the Transformer model group. Google Brain created this neural network design in 2017 [5].Transformers models employ a self-attention technique that is suitable for comprehending natural language. It should be noted that the 2015 introduction of the attention mechanism led to a significant breakthrough and made it possible for Google to develop the first models of this kind in the ensuing years, including GPT-1 and BERT. The function that determines the likelihood of another word appearing surrounded by others is called attention.

B. Models of Openai

Like its numerical predecessors, the most recent GPT-4 transformer model was developed by the research group OpenAI and is regarded as the industry standard. The company was established in 2015 and is seen as DeepMind's direct rival [6]. Microsoft announced that exclusive GPT-X licensing has been agreed upon [7].Using deep learning, the third-generation autoregressive language model GPT-3 generates text that is similar to that of a human. In simpler terms, it is a computing system that is intended to produce word, code, or other data sequences from a source input known as the prompt. It is used, for example, in machine translation to statistically predict word sequences. The language model is trained on an unlabeled dataset comprising texts, such as Wikipedia and many other sites, primarily in English and a few other languages. These statistical models must be trained with large amounts of data to produce relevant results. The first iteration of GPT in 2018 used 110 million learning parameters (i.e., the values a neural network tries to optimize during training). A year later, GPT-2 used 1.5 billion of them. GPT-3 used 175 billion parameters. Nowadays, GPT-4 uses 170 trillion parameters, which is a significant increase compared to GPT-3.5. This is expected to significantly improve the model's ability to generate coherent and contextually appropriate responses to text prompts and its overall language understanding and NLP capabilities [8].The more parameters a model has, the more data is needed to train the model. According to the creators, the OpenAI GPT-3 model was trained on 45 TB of text data from several sources. Several data sets which are used to train the model are listed in table1.

are listed in table (Tab 1).

TABLE I
THE RATIO OF DATASETS ON THE GPT-3 TRAINING SET

Dataset	Number of tokens	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60.00%	0.44
WebTex2	19 bilion	22.00%	2.90
Books1	12 billion	8.00%	1.90
Books2	55 billion	8.00%	0.43
Wikipedia	3 bilion	3.00%	3.40

It is trained on the AI supercomputer in Microsoft's Azure [9]. The anticipated cost of the training was $12 million [10]. In addition to chatbots, the chosen method can be used for summarization, grammatical correction, email composing, translation, question answering, and many more applications.Based on the given specifications, GPT-3 wrote an essay that was published in the British publication The Guardian in 2020 [11]. The article was sensationalist and the text was altered. Nevertheless, it must be acknowledged that a wave of criticism regarding the text's presentation surfaced after the piece was published.The Guardian has come under fire from prominent AI experts for deceiving the public. They used terms like "good" and "evil," which are obviously notions that GPT-3 cannot understand, as examples in the article [12].

C. Beijing Academy of Artificial Intelligence Model

Wu Dao 2.0, developed by BAAI, is a multimodal AI with 1.75 trillion parameters, making it ten times larger than GPT-3. It can handle tasks like NLP, image recognition, and protein structure prediction, using FastMoE for greater flexibility. Trained on 4.9 terabytes of data, it is part of a global trend of large-scale AI models developed by countries like Russia, France, and Korea. In contrast, Google's BERT focuses on text-based NLP tasks and has evolved into the chatbot Gemini. Wu Dao 2.0 excels in multimodal applications, while BERT specializes in language understanding.

D. Models Based on Rules

Rule-based models are AI systems that rely on predefined rules to generate responses. When a user's query matches a specific rule, the system provides the corresponding answer. If no match is found, the user is informed that there is no response. The primary advantage of rule-based models is their consistency and accuracy, as they always follow fixed rules. However, they don't scale well because adding new responses requires manually creating additional rules. Google Dialogflow and IBM Watson are popular examples of rule-based systems Some rule-based systems, like retrievalbased chatbots, use a list of predefined sentences and measure the similarity between the user's query and the stored sentences. Cosine similarity is commonly used to determine which sentence most closely matches the query. The chatbot then responds with the most similar sentence from the list.These models are widely used in various fields, including customer service, therapeutic chatbots, and education, to handle frequently asked questions. They are also employed in information retrieval systems, where users input specific criteria (like cities for travel routes), and the system matches the query with predefined responses. While effective for straightforward tasks, rule-based models lack the flexibility and scalability of more advanced machine learning models.

E. Hybrid Models

Because the two previously mentioned approaches—rule-based models and large language models (LLMs)—have both substantial benefits and basic disadvantages in various domains, it has become necessary to combine them in order to maximize their strengths and minimize their weaknesses. The use of rule-based/intent-based model components to determine context or carry out particular activities, followed by the formatting of the output utilizing LLMs to produce syntactically correct responses that mimic human speech, are common characteristics of such combinations.

This approach is employed in many different domains, such as medicine, where Med| Primary AI assistant is utilized [35]. The primary purpose of these tools in the designated field is to increase the efficiency of patient diagnosis, where an LLM presents the results to the patient after a ruleaugmented AI-powered system that includes a rule-based decision system has selected the diagnosis [36]. Although there are many tools available for building chatbots [37], RASA is one of the most often used tools for hybrid models. It started out as a platform for building rule-based chatbots, but as more advanced LLMs became available that could more accurately mimic human responses, the platform was improved to enable the user to integrate LLMs via API [38]

III.Materials and Methods

A. Overview

In this study, the results of two language models with differing algorithmic foundations are empirically compared. By establishing comparable settings for both models, we want to achieve a fair comparison and prevent bias in our assessment of their performance. We anticipate that this comparison will provide us a more thorough grasp of each model's advantages and disadvantages as well as suggestions for future developments.

B. Data Preparation

Initially, a suitable high-quality dataset had to be extracted. We have determined that, despite the fact that the system contained data dating back to 2013, a more pertinent sample would only include the years 2020–2022, based on discussions with CC department staff.With this method, we were able to extract 149142 records in total. Every record reflects a single exchange between a user and a CC employee of the relevant business. However, we used basic pattern matching to eliminate all records that contained words from the given list from the dataset because many of them had offensive language or improper formulations.

C.Model Preparation

It is clear that the model training step is arguably the most important stage for achieving pertinent findings. Our research's findings would be skewed if we made any changes to the aforementioned method between the two models. Therefore, even though employing a single strategy might not be the best for performance, we have chosen to train both models in a comparable manner. Fairness and comparability between the two models can be guaranteed by keeping the training procedure consistent. Although performance optimization may suffer as a result, it offers a strong basis for carrying out objective research and producing trustworthy findings.Following the guidelines in the documentation guarantees that our dataset is ready and compatible with the refined model derived from the J-Large model. By adhering to these rules, we contribute to ensuring that the model operates and is trained appropriately for our particular needs. It's interesting to note that the system automatically split the dataset into 500 test sets. Based on this default behavior, we have implemented the same change in our rule-based model.

D.Model Integration

Through HTTP API calls to external systems that represented the relevant models, both models were integrated into the business's CC system for communication with VIP clients. We were constrained by the capabilities offered by the corresponding system for the model M2, which was the application instance of the optimized GPT model, in both the HTTP request and the HTTP response. The API answer from the M2 model was lengthy and too complicated for our requirements [47].

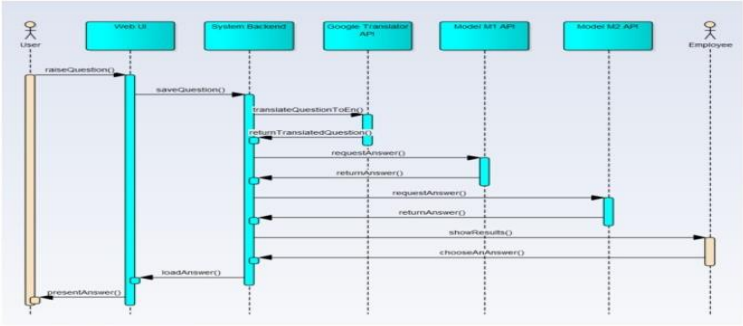


FIGURE 3. System communication representation

Consequently, we chose to reduce the HTTP response for model M1, the rule-based model's application instance, to a simple JSON format with just one property, "answer," which included the desired answer to the query that was sent (fig. 3).We did not have the M1 model using the Flask framework and Python programming language as the user and the company's system did not communicate in real-time. The primary cause was that the model was in the pickle format, which we also used Python tools to compile during the training stage.

E.Data Extraction

Following the system's integration of the aforementioned application interfaces, training sessions were held to teach staff members how to use and communicate with the system. The staff members could use the new feature in a number of ways:

1. Select the whole response that the M1 model offers.
2. Select the complete response that the M2 model offers.

3. Select the response that the M1 model offers and make any necessary modifications.
4. Select the response that the M2 model offers and make any necessary modifications.
5. Write a unique response instead of selecting one of the pre-provided solutions.Staff could customize responses based on expertise, providing optimal support. All activities were logged during the study from September 20, 2022, to January 2, 2023, offering valuable insights into system performance. While staff followed consistent guidelines, individual judgment varied in assessing response accuracy, though this did not significantly affect the results. The study focused on comparing the M1 and M2 models' performance. Staff were unaware of which model provided each response, as the system randomly alternated their order, ensuring an unbiased evaluation of accuracy.

IV.Outcomes

The M1 and M2 models were deployed in a customer care (CC) environment for 15 weeks, with staff evaluating response adequacy and comprehensibility. The majority of M1 responses received a high understanding score (5), but it often reused previous human-generated responses. The M2 model, however,peformed better with high adequacy and comprehensibility scores (median and upper quartile both 5), indicating it generated clearer and more relevant responses.During the study, 57% of responses suggested by the models were not selected by employees, who preferred to modify or select alternative responses. Staff rated responses for comprehensibility (with 5 being fully understandable) and sufficiency (relevance to the issue). Due to time limits, only selected responses were evaluated.While the M2 model was chosen less frequently (7% of the time), its responses were found to be more suitable and comprehensible than those from M1, which was selected 36% of the time. The study highlights the importance of further evaluation methods, as human assessors only reviewed a subset of responses.

V.Conclusion

This research compares GPT models and rule-based models for NLP tasks, focusing on human evaluations and automatic metrics. Human evaluation, while subjective and time-intensive, effectively assesses comprehensibility, contextual relevance, and naturalness. Automatic metrics provide scalability and complement human assessments.Findings show that rule-based models excel in domain-specific tasks, especially when tailored datasets are used, making them cost-effective and reliable for structured environments like customer care. However, they struggle with complex, unpredictable user inputs. In contrast, GPT models demonstrate superior comprehensibility, natural language generation, and contextual adequacy, enhancing user experience in open-ended interactions. Their drawbacks include high computational costs and extensive data
requirements.The study highlights the strengths and limitations of both approaches, emphasizing that model selection should align with task complexity, dataset quality, and resource availability.Future research could focus on hybrid models that combine the efficiency of rule-based systems with the adaptability and language proficiency of GPT models for optimized chatbot performance.

References

[1] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction Tuning With GPT-4. Accessed: Mar.7,2024.[Online]. Available: https://instruction-tuning-with-gpt-4.github.io/
[2] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju,‘‘Touch your heart: A tone-aware chatbot for customer care on socialmedia,’’ in Proc. CHI Conf. Human Factors Comput. Syst., Apr. 2018,pp. 1–12.
[3] A. A. Georgescu, ‘‘Chatbots for education—Trends, benefits and chal-lenges,’’ in Proc. 14th Int. Conf. eLearning Softw. Educ., Apr. 2018, vol. 14,no. 2, pp. 195–200.
[4] J. Chu, ‘‘Recipe bot: The application of conversational AI in home cookingassistant,’’ in Proc. 2nd Int. Conf. Big Data Artif. Intell. Softw. Eng.(ICBASE), Sep. 2021, pp. 696–700.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,Ł. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in Proc. 31stConf. Neural Inf. Process. Syst., vol. 30, Dec. 2017, pp. 5999–6009.

[6] Microsoft Invests in and Partners With OpenAI To Support Us Build-ing Beneficial AGI. Accessed: Mar. 7, 2024. [Online]. Available:https://openai.com/blog/microsoft-invests-in-and-partners-with-openai
[7] Microsoft Teams Up With OpenAI To Exclusively License GPT-3 LanguageModel—The Official Microsoft Blog. Accessed: Mar. 7, 2024. [Online].Available: https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/
[8] OpenAI. GPT-4 Technical Report. Accessed: Sep. 15, 2023. [Online]. Available:https://arxiv.org/abs/2303.08774v#
[9] OpenAI's Massive GPT-3 Model is Impressive, but Size Isn't Everything | VentureBeat. Accessed: Mar. 7, 2024. [Online]. Available:https://venturebeat.com/ai/ai-machine-learning-openai-gpt-3-size-isnt-everything/
[10] A Robot Wrote This Entire Article. Are You Scared Yet, Human? |GPT-3 | The Guardian. Accessed: Mar. 7, 2024. [Online]. Available:
https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3
[11] The Guardian's GPT-3-Written Article Misleads Readers About AI.Here's Why—TechTalks. Accessed: Mar. 7, 2024. [Online]. Available:https://bdtechtalks.com/2020/09/14/guardian-gpt-3-article-ai-fake-news/
[12] Models—OpenAI API. Accessed: Mar. 7, 2024. [Online]. Available:
https://platform.openai.com/docs/models/overview
[13] AI21 Labs Makes Language AI Applications Accessible To BroaderAudience | Bus. Wire. Accessed: Mar. 7, 2024. [Online]. Available: https://www.businesswire.com/news/home/20210811005033/en/AI21-Labs-Makes-Language-AIApplications-Accessible-to-Broader-Audience
[14] Language Models Are Unsupervised Multitask Learners. Accessed:Sep. 15, 2023. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
[15] Jurassic-1: Technical Details and Evaluation. Accessed: Sep. 15, 2023.[Online]. Available:https://cdn.prod.website-files.com/60fd4503684b46
6578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf
[16] Introducing J1-Grande! Accessed: Mar. 7, 2024. [Online]. Available:
https://www.ai21.com/blog/introducing-j1-grande
[17] L. Reed, C. Li, A. Ramirez, L. Wu, and M. Walker, "Jurassic is (almost)all you need: Few-shot meaning-to-text generation for open-domain dia-logue," in Proc. Int. Conf. Conversational AI Natural Hum.-CentricInteract., in Lecture Notes in Electrical Engineering, vol. 943, 2021,pp. 99–119.
[18] A. Przegalinska and D. Jemielniak, Strategizing AI in Business and Edu-cation. Cambridge, U.K.: Cambridge Univ. Press, Apr. 2023.
[19] The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bringan African Nation to the Brink. Accessed: Sep. 15, 2023. [Online]. Avail-able:
https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/
[20] J. He, J. Qiu, A. Zeng, Z. Yang, J. Zhai, and J. Tang, "FastMoE: A fastmixture-of-expert training system," 2021, arXiv:2103.13262.
[21] AI Weekly: China's Massive Multimodal Model Highlights AI ResearchGap | VentureBeat. Accessed: Mar. 7, 2024. [Online]. Available:
https://venturebeat.com/business/ai-weekly-chinas-massive-multimodal-model-highlights-ai-research-gap/
[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc.Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Tech-nol., vol. 1, Oct. 2018, pp. 4171–4186.
[23] Introducing Google Research Football: A Novel Reinforcement Learning Environment. Accessed: Sep. 15, 2023. [Online]. Available:
https://research.google/blog/introducing-google-research-football-a-novel-reinforcement-learning environment/
[24] Google Bard is Now Gemini: How to Try Ultra 1.0 and New Mobile App. Accessed: Jul. 9, 2024. [Online]. Available:

https://blog.google/products/gemini/bard-gemini-advanced-app/
[25] S. A. Thorat and V. Jadhav, ‘‘A review on implementation issues of rule-based chatbot systems,’’ in Proc. Int. Conf. Innov. Comput. Commun.
(ICICC), 2020, doi: 10.2139/ssrn.3567047.
[26] R. Agarwal and M. Wadhwa, ‘‘Review of state-of-the-art design techniques for chatbots,’’ Social Netw.Comput. Sci., vol. 1, no. 5, pp. 1–12, Sep. 2020.
[27] N. V. Shinde, A. Akhade, P. Bagad, H. Bhavsar, S. K. Wagh, andA. Kamble, ‘‘Healthcare chatbot system using artificial intelligence,’’ in Proc. 5th Int. Conf. Trends Electron. Informat. (ICOEI), Jun. 2021, pp. 1–8.
[28] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner,and M. Househ, ‘‘An overview of the features of chatbots in mentalhealth: A scoping review,’’ Int. J. Med. Informat., vol. 132, Dec. 2019,
Art. no. 103978.
[29] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir,D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, ‘‘Conver-sational agents in healthcare: A systematic review,’’ J. Amer. Med. Inform.
Assoc., vol. 25, no. 9, pp. 1248–1258, Sep. 2018.
[30] K. Moore, S. Zhong, Z. He, T. Rudolf, N. Fisher, B. Victor, and N. Jindal,
‘‘A comprehensive solution to retrieval-based chatbot construction,’’ Com-put. Speech Lang., vol. 83, Jan. 2024, Art. no. 101522.
[31] H. Akkineni, P. V. S. Lakshmi, and L. Sarada, ‘‘Design and development of retrieval-based chatbot using sentence similarity,’’ in Proc. Int. Conf.IoT Anal. Sensor Networks, in Lecture Notes in Networks and Systems,
vol. 244, 2022, pp. 477–487.
[32] Z. Chen, Y. Lu, M. P. Nieminen, and A. Lucero, ‘‘Creating a chatbot for and with migrants: Chatbot personality drives co-design activities,’’ in Proc. ACM Designing Interact. Syst. Conf., Jul. 2020, pp. 219–230.
[33] B. Thomson, Statistical Methods for Spoken Dialogue Management.London, U.K.: Springer, 2013, doi: 10.1007/978-1-4471-4923-1.
[34] R. Dsouza, S. Sahu, R. Patil, and D. R. Kalbande, ‘‘Chat with bots intel-ligently: A critical review & analysis,’’ in Proc. Int. Conf. Adv. Comput.,Commun. Control (ICAC3), Dec. 2019, pp. 1–6.
[35] D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, ‘‘Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis,’’ Electronics, vol. 13, no. 2, p. 320, Jan. 2024.
[36] D. P. Panagoulias, F. A. Palamidas, M. Virvou, and G. A. Tsihrintzis,‘‘Rule-augmented artificial intelligence-empowered systems for medical diagnosis using large language models,’’ in Proc. IEEE 35th Int. Conf.Tools with Artif. Intell. (ICTAI), Nov. 2023, pp. 70–77.
[37] J. S. Cuadrado, S. Pérez-Soler, E. Guerra, and J. de Lara, ‘‘Automating the development of task-oriented LLM-based chatbots,’’ in Proc. 6th ACM Conf. Conversational User Interfaces (CUI), 2024, pp. 1–10.
[38] Using LLMs With Rasa. Accessed: Jul. 9, 2024. [Online]. Available:
https://rasa.com/docs/rasa/next/llms/large-language-models/
[39] M. Y. Helmi Setyawan, R. M. Awangga, and S. R. Efendi, ‘‘Comparison of multinomial naive Bayes algorithm and logistic regression for intent classification in chatbot,’’ in Proc. Int. Conf. Appl. Eng. (ICAE), Oct. 2018,pp. 1–5.
[40] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, ‘‘Benchmarking natural language understanding services for building conversational agents,’’ in Increasing Naturalness and Flexibility in Spoken Dialogue Interaction
(Lecture Notes in Electrical Engineering), vol. 714. Singapore: Springer,2019, pp. 165–183.
[41] D. Braun, A. Hernandez-Mendez, F. Matthes, and M. Langen, ‘‘Evaluat-ing natural language understanding services for conversational question answering systems,’’ in Proc. 18th Annu. SIGdial Meeting Discourse Dialogue, 2017, pp. 174–185.

[42] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" in Proc. Deep Learn. Inside Out (DeeLIO), 3rd Workshop Knowl. Extraction Integr. Deep Learn. Architectures, 2022, pp. 100–114.
[43] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul,P. Chaovalit, and T. Theeramunkong, "A survey on evaluation methods for chatbots," in Proc. 7th Int. Conf. Inf. Educ. Technol., Mar. 2019,pp. 111–119.
[44] S. Pandey and S. Sharma, "A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning," Healthcare Anal., vol. 3, Nov. 2023, Art. no. 100198.
[45] P. Qin, W. Xu, and J. Guo, "A novel negative sampling based on TFIDF for learning word representation," Neurocomputing, vol. 177, pp. 257–265,Feb. 2016.
[46] C.-H. Chen, "Improved TFIDF in big news retrieval: An empirical study,"Pattern Recognit. Lett., vol. 93, pp. 113–122, Jul. 2017.
[47] J2 Complete API. Accessed: Mar. 7, 2024. [Online]. Available:
https://docs.ai21.com/reference/j2-complete-api-re