Available *online* at : http://jitce.fti.unand.ac.id/

JITCE (Journal of Information Technology and Computer Engineering)

| ISSN (Online) 2599-1663 |

Technical Article

Evaluating IndoGPT for Legal Queries: A Benchmark Against GPT-4 Models

Ade Cahyaning Palupi [1], *Ade Irawan* [12]

[1] *Department of Computer Science, Universitas Pertamina, Jakarta, 12220, Indonesia*
[2] *Center for Data Science and Automation (CDSCAN), Universitas Pertamina, Jakarta, 12220, Indonesia*

ARTICLE INFORMATION

Received: December 24th, 2024
Revised: September 10th, 2025
Available online: September 30th, 2025

KEYWORDS

Chatbot, GPT, indoGPT, RAGAS, RAG

CORRESPONDENCE

Phone: -
E-mail: -

ABSTRACT

This study evaluates a chatbot developed with the Large Language Model (LLM) IndoGPT, focusing on its use of Retrieval-Augmented Generation (RAG) to answer questions about university regulations from legal PDF documents in the Indonesian Language. IndoGPT's performance is benchmarked against the more advanced models, GPT-4 and GPT-4o. The chatbot combines RAG techniques with the LangChain framework, and its effectiveness is assessed using the Retrieval-Augmented Generation Assessment (RAGAS) framework. The dataset includes publicly available legal documents from Universitas Pertamina, with test data created by the authors. IndoGPT consistently underperforms compared to GPT-4 and GPT-4o. GPT-4 achieves superior metrics with Context Precision at 0.9027, Context Recall at 0.8693, Faithfulness at 0.8486, and Answer Relevancy at 0.8142. Similarly, GPT-4o delivers strong results with Context Precision at 0.8896, Context Recall at 0.8594, Faithfulness at 0.8804, and Answer Relevancy at 0.8773. In contrast, IndoGPT shows significant deficiencies, with much lower scores: Context Precision at 0.6687, Context Recall at 0.5711, Faithfulness at 0.0738, and Answer Relevancy at 0.1628. These findings highlight IndoGPT's substantial limitations, especially when compared to GPT-4 and GPT-4o, which excel in providing accurate, contextually relevant answers. The GPT-4-based chatbot demonstrates strong capabilities in understanding user queries and delivering accurate responses while effectively reducing hallucinations through the RAG technique.

INTRODUCTION

Universities implement formal regulations, approved by leadership, to ensure educational quality, consistency in operations, and provide clear guidelines for the academic community. However, accessing university regulations can be particularly challenging due to the scattered nature of documents across different departments, variations in document formats, and the frequent updates that are not centrally communicated. For example, students often struggle to find the latest academic policies regarding course withdrawal deadlines, GPA requirements for scholarships, or internship regulations, leading to delays or misinformed decisions. Similarly, staff members, particularly academic advisors and administrative personnel, face significant difficulties in locating updated human resource guidelines (such as faculty promotion requirements or leave policies), research grant regulations, and procurement procedures. In some cases, staff must manually consult multiple offices or outdated intranet portals to compile a complete understanding of applicable rules, resulting in wasted time and potential administrative errors.

To address this challenge, it is essential to implement a system that facilitates efficient access to regulatory information. One potential solution is the development of a digital platform, incorporating technologies such as chatbots, to provide quick and convenient access to all relevant documents for the academic community [1]. An effective chatbot solution can dramatically improve efficiency by offering instant, centralized access to the most updated regulations. This not only saves time but also promotes better compliance with institutional policies and enhances user satisfaction by reducing the frustration associated with manual searches.

Conventional chatbots utilize Natural Language Processing (NLP), which is developed using rule-based methods, basic machine learning algorithms, or information retrieval techniques [2]. These chatbots can interact and respond to user queries by recognizing textual patterns and classifying word and sentence meanings based on predefined schemes. However, traditional chatbot models exhibit limitations in comprehending deeper contexts and frequently generate responses that are unnatural or irrelevant, particularly when dealing with complex or open-ended questions [3].

https://doi.org/10.25077/9.1.22-27.2025

On the other hand, Large Language Models (LLMs) offer enhanced context comprehension due to their training on vast and diverse text datasets, allowing them to perform effectively across various domains and datasets [4][5]. LLMs enable chatbots to understand and generate text with greater sophistication and accuracy compared to traditional methods. Furthermore, LLM-based chatbots exhibit improved adaptability, making them more capable of handling complex, open-ended questions. This results in more natural, contextually relevant responses, which enhance the overall user experience and the chatbot's utility in diverse applications.

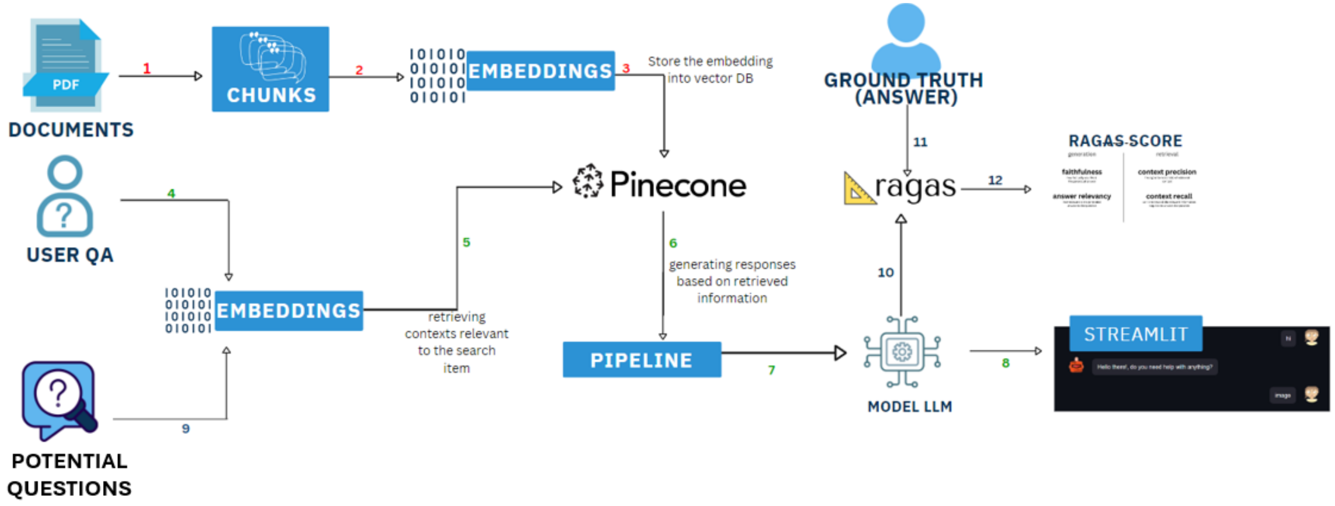


Figure 1. System Diagram of the Chatbot Development

LLM-based chatbots, such as ChatGPT (https://chatgpt.com) and GEMINI (https://gemini.google.com), have recently received significant recognition from researchers across various fields [6-9]. These LLMs, often referred to as transformer-based models, are trained on vast volumes of text data and consist of billions of parameters. The release of OpenAI's Generative Pre-trained Transformer (GPT) in 2018 [10] marked the beginning of widespread public interest in LLMs. Since then, increasingly larger and more advanced models have been developed, including GPT-2 [4], GPT-3 [11], GPT-3.5 [12], and the latest GPT-4 [13].

Previous studies have analyzed the performance of LLMs in question-answering tasks for specific domains, using GPT-4 model. The authors in [14] developed a chatbot system using the GPT-4 model and retrieval-augmented generation (RAG) methods to provide information on cleanliness (taharah) according to different Islamic jurisprudence schools by supplying external data to the model, thereby preventing hallucinations in responses. The study showed that the chatbot using the GPT-4 model with RAG methods increased accuracy in answering questions, achieving an accuracy of 0.9 compared to 0.7 for the GPT-4 model without RAG. The chatbot could provide appropriate responses when the topic was within the database, but returned no results when the query topic was outside the database's scope.

Another research in [15] discusses the evaluation and optimization of chatbot responses using LLMs, focusing on RAG for a human resource (HR) chatbot. The authors collected a dataset containing potential questions and corresponding articles, performed fine-tuning and retrieval processes, and optimized and evaluated the HR chatbot for response quality. The optimization results of the OpenAI retriever model in the study showed a slight improvement, but the accuracy remained low due to a poor evaluation dataset. Despite this, the GPT-4 model demonstrated good competence in handling user questions, indicating that the model's internal reasoning and domain knowledge were sufficiently robust. These studies have tested the performance of LLMs in question-answering tasks for specific domains.

Although GPT-4 demonstrates strong performance, it is fundamentally trained as a multilingual model. Research in [16] indicates that for certain low-resource languages, such as Indonesian, a decoder-only model like IndoGPT can produce competitive results compared to large multilingual models, i.e., mBART$_{\text{LARGE}}$ [17]. Specifically, IndoGPT achieves this performance while utilizing only 20% of the parameters of the larger model, yet it performs inference 4 times faster on CPU and 2.5 times faster on GPU.

This research evaluates the efficacy of LLMs in answering user inquiries about the regulations at Universitas Pertamina, in the form of a chatbot. A significant challenge in employing LLMs is their tendency to produce hallucinations or irrelevant content, which can be addressed by utilizing techniques such as RAG to enhance contextual accuracy [18][19]. This research assesses three key dimensions: faithfulness (i.e., whether the response is grounded in the retrieved context), answer relevance (i.e., whether the response addresses the user's query), and context relevance (i.e., whether the retrieved context is sufficiently focused) to ensure robust evaluation. These assessments are conducted using the RAGAS metric [20]. Therefore, the primary objective of this research is to evaluate the performance of the IndoGPT model using RAG techniques and to compare its performance with GPT-4 and GPT-4o based on the RAGAS metric.

METHOD

The development of a chatbot to assist the academic community in finding and understanding Universitas Pertamina's regulatory documents were carried out using LangChain (https://www.langchain.com) and RAG. Figure 1 shows the

https://doi.org/10.25077/9.1.22-27.2025

system diagram of the development. The first stage of this research involved collecting data in the form of PDF files containing regulatory documents from Universitas Pertamina. The next step was to divide the documents into smaller chunks to enhance the accuracy of vector-based search retrieval. Subsequently, embedding models were applied to the text chunks, transforming them into vector representations to facilitate information extraction. This process converted textual data into vector representations that captured semantic meaning and word similarity. Each chunk was converted into a vector using embedding models compatible with the selected LLM, such as OpenAI or IndoGPT. The resulting embeddings were stored in a database. Pinecone (https://www.pinecone.io) was employed as the vector store database to store efficiently, index, and retrieve semantically similar documents.

Chatbot implementation involved capturing user questions and obtaining their vector embeddings. This embedding process facilitated an understanding of the contextual meaning and enabled the retrieval of similar contexts. It is important to note that the preprocessing steps such as stopword removal, stemming, or lemmatization were not applied to user input when using the GPT-4 model. This is because GPT-4's self-attention mechanism allows the model to handle informal words and typographical errors effectively. The self-attention mechanism focuses on all words within the input and considered their contextual relationships. This enables the model to understand the meaning of words in a broader context and to correct typographical errors.

Additionally, it is also important to note that models like BERT and GPT-4 were trained on data that had yet to undergo preprocessing steps, such as stopword removal or stemming. Preprocessing the training data could interfere with the model's ability to accurately and contextually understand text by removing or altering critical elements within the data. According to [21], applying preprocessing to user input after a model has been trained may risk altering the text's original meaning, potentially misaligning it with the patterns and contexts learned during training. Therefore, modifications to user sentences or commands were made directly within the chatbot's system prompt without additional preprocessing. This approach ensured that the original meaning of user input was preserved and aligned with the data the model had been trained on.

The vectors obtained from the previous process were connected to a database to perform semantic searches, identifying contexts similar to the given query and retrieving relevant documents. Once the retrieval process produced relevant documents from the data source, these documents were combined with the original prompt and user query as additional context. This combined text was then passed to the model to generate a response, which was prepared as the system's final output.

Subsequently, the LLM utilized patterns and context from prior response histories (if available) to determine the most accurate answer to the given question. By analyzing these patterns and contexts, the LLM produced more relevant responses that aligned with the user's query. The generated answers were displayed through a user interface (UI) built using the Streamlit (https://streamlit.io) framework. The UI facilitated user interaction with the chatbot and provided a clear presentation of the question-and-answer (QA) interactions.

Finally, several aspects were considered based on the RAGAS metrics during the evaluation phase of the answers generated by the RAG system, which are Faithfullness, Answer Relevancy, Context Relevancy, and Context Recall. Faithfulness (F) metric measures hallucinations, which is given by

$$F = \frac{|V|}{|S|} \qquad (1)$$

where $|V|$ is the number of statements that were supported according to the LLM and $|S|$ is the total number of statements. Answer Relevancy (AR) evaluates the relevance of the response to the question, given by

$$AR = \frac{1}{n}\sum_{i=1}^{n} cos\left(E_o, E_{q_i}\right) \qquad (2)$$

where $cos\left(E_o, E_{q_i}\right)$ is a cosine similarity function to calculate the similarity of the original question embedding E_o with n generated potential questions embedding E_{q_i} by the LLM. Context Relevancy (CV) assesses the ratio of meaningful information to noise in the retrieved context, denoted by

$$CV = \frac{|Ne|}{|Sc|} \qquad (3)$$

where $|Ne|$ is the number of extracted sentences, and $|Sc|$ is the total number of sentences in its context. Context Recall (CR) reflects the system's ability to retrieve all relevant information needed to answer the question, denoted by

$$CR = \frac{|GT|}{|Nc|} \qquad (4)$$

where $|GT|$ is the number of ground truth claims that can be attributed to the context, and $|Nc|$ is the number of claims in ground truth. Note that Context Relevancy and Context Recall serve as performance indicators for the information retrieval component of the system. This evaluation process ensured that the RAG system provided accurate and contextually appropriate answers.

RESULTS AND DISCUSSION

Data were collected from legal documents of Universitas Pertamina, which can be publicly accessed through the following URL: https://universitaspertamina.ac.id/download-center. The data were divided into chunks of 1,000 characters each. This process resulted in a total of 2,999 chunks stored in the database. The documents consisted of various regulations and guidelines from Universitas Pertamina, carefully selected for their relevance and reliability in addressing the chatbot's topic. Each user query in the chatbot system underwent a series of stages, including embedding, retrieval, and generation, before being processed by the LLM model to produce responses. The generated answers were then compared against the ground truth to evaluate the system's performance.

The performance of the chatbot was tested using three language models: GPT-4, GPT-4o, and IndoGPT. The evaluation employed four metrics: Context Precision, Context Recall, Faithfulness, and Answer Relevancy. The assessment was conducted within the RAGAS framework on a dataset comprising 123 questions. The results of the evaluation are presented in Table 1.

Table 1. The Average Evaluation of the GPT-4, GPT-4o, and IndoGPT models

Evaluation Metric	GPT-4	GPT-4o	IndoGPT
Context Precision	**0.9027**	0.8896	0.6687
Context Recall	**0.8693**	0.8594	0.5711
Faithfulness	0.8486	**0.8804**	0.0738
Answer Relevancy	0.8142	**0.8773**	0.1628

Context Precision: GPT-4 achieved the highest Context Precision value of 0.9027, demonstrating its ability to consistently select relevant contexts for answering the given questions. The GPT-4o model also performed well, with a Context Precision value of 0.8896. In contrast, IndoGPT showed a significantly lower Context Precision value of 0.6687, indicating that it frequently failed to select relevant contexts and exhibited inconsistency in aligning its selections with the ground truth.

Context Recall: GPT-4 also demonstrated superior performance in Context Recall, achieving a value of 0.8693, indicating that the selected contexts consistently contained the necessary information to answer the questions correctly. GPT-4o performed slightly lower, with a Context Recall value of 0.8594, but still maintained a strong level of performance. In contrast, IndoGPT exhibited a significantly lower Context Recall value of 0.5711, highlighting its inability to select contexts that effectively contribute to providing correct answers.

Faithfulness: The GPT-4o model achieved the highest Faithfulness value of 0.8804, indicating that the answers generated by this model were highly consistent with the facts presented in the selected contexts. GPT-4 also demonstrated strong performance with a Faithfulness value of 0.8486. In contrast, IndoGPT showed a significantly lower Faithfulness value of 0.0738, suggesting that its answers were often inconsistent with the facts or contexts provided.

Answer Relevancy: GPT-4o achieved the highest Answer Relevancy value of 0.8773, indicating that the answers generated by this model were highly relevant to the given questions. GPT-4 also performed well, with an Answer Relevancy value of 0.8142. However, IndoGPT demonstrated a much lower value of 0.1628, suggesting that its generated answers were often irrelevant or failed to fully address the given questions. This result highlights that IndoGPT is less effective in producing relevant answers to the provided prompts.

Overall, the GPT-4 and GPT-4o models demonstrated superior performance across all evaluation metrics compared to IndoGPT. This indicates that GPT-4 and GPT-4o are significantly more effective in selecting and utilizing relevant contexts to generate factual and relevant answers. The performance evaluation revealed significant disparities between the models, with IndoGPT demonstrating substantial limitations across all metrics. To better understand these performance gaps, we conducted a detailed analysis of specific failure cases that illustrate the underlying issues with IndoGPT's performance in this specialized domain.

Irrelevant Information Retrieval IndoGPT frequently retrieved contextually inappropriate information when processing user queries about university regulations. For instance, when a user asked "Apa syarat untuk mengajukan cuti akademik?" (What are the requirements for applying for academic leave?), IndoGPT retrieved documents related to faculty promotion criteria rather than student academic leave policies. The model appeared to focus on individual keywords such as "syarat" (requirements) without adequately understanding the broader contextual meaning of the query. This resulted in a Context Precision score that was significantly lower than both GPT-4 and GPT-4o, indicating systematic failures in the retrieval component of the RAG pipeline.

Inaccurate Answer Generation Beyond retrieval issues, IndoGPT demonstrated substantial problems in generating accurate responses even when relevant context was available. In one documented case, when provided with correct information about thesis submission deadlines, IndoGPT generated a response stating that students had unlimited time to submit their thesis, directly contradicting the retrieved regulation that specified a maximum period of two years. This pattern of generating factually incorrect information despite having access to accurate source material contributed significantly to the model's extremely low Faithfulness score of 0.0738.

Low Faithfulness and Hallucination Patterns The most concerning aspect of IndoGPT's performance was its tendency to generate responses that were not grounded in the retrieved context. The model frequently produced answers that appeared plausible but contained fabricated details not present in the source documents. For example, when asked about scholarship application procedures, IndoGPT invented specific GPA requirements and application deadlines that did not exist in the actual university regulations. This hallucination pattern was particularly problematic in a legal document context where accuracy is paramount, explaining the dramatic difference between IndoGPT's Faithfulness score and those achieved by GPT-4 and GPT-4o.

These specific failure modes highlight fundamental limitations in IndoGPT's ability to effectively process and synthesize specialized legal and regulatory content, demonstrating the challenges faced by smaller language models when deployed in domain-specific applications requiring high accuracy and reliability.

The chatbot interface was developed using the Streamlit framework, as shown in Figure 2.

CONCLUSIONS

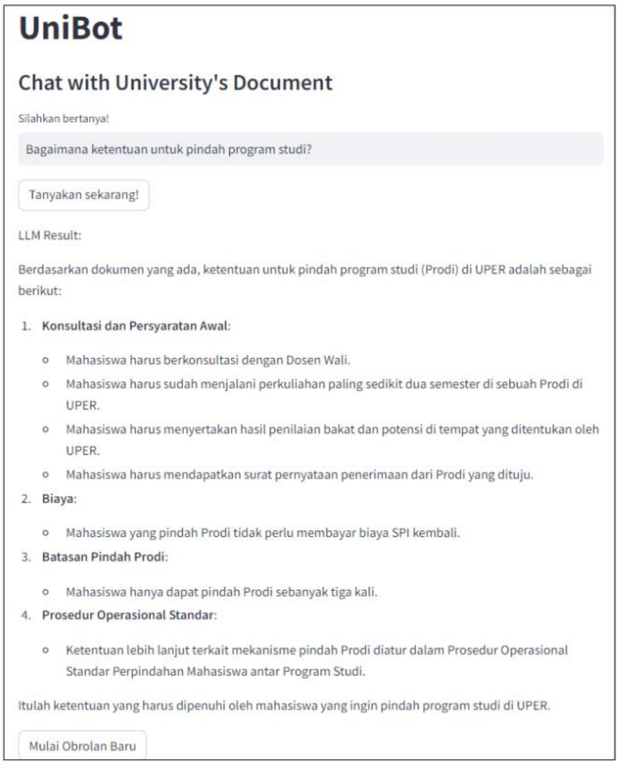
This research evaluated the performance of a chatbot designed for the Universitas Pertamina document domain using the RAG approach, with a particular focus on the IndoGPT model, alongside comparisons to GPT-4 and GPT-4o. The study aimed to explore the potential of IndoGPT as a locally trained language model for answering user queries based on university document data. Key metrics such as context precision, context recall, answer relevancy, and faithfulness were used to assess the models' performance.

The results demonstrated that IndoGPT faced significant challenges in delivering accurate and contextually relevant answers. The model frequently failed to retrieve appropriate contexts and generate responses aligned with the questions, leading to low performance across all evaluation metrics. This highlights a critical gap in IndoGPT's ability to handle specialized document-based tasks compared to larger multilingual models like GPT-4 and GPT-4o, which consistently outperformed IndoGPT in generating accurate and relevant answers.

Despite its limitations, IndoGPT's low-resource nature offers an important avenue for further research and development. Improvements in the embedding and retrieval processes, as well as fine-tuning on domain-specific data, could enhance IndoGPT's performance in such applications. This study underscores the importance of refining locally trained language models like IndoGPT to better serve document-based chatbot systems in specific linguistic and cultural contexts.

The findings have significant implications for other low-resource languages and specialized domain applications. The performance disparity observed with IndoGPT suggests that researchers working with local Indonesian languages such as Javanese, Sundanese, Batak, or Minangkabau may encounter similar challenges when developing domain-specific applications. Organizations seeking to implement specialized chatbots for domains such as regional legal documentation, traditional medicine consultations, or local government services must carefully evaluate trade-offs between computational efficiency and performance accuracy when serving diverse Indonesian linguistic communities.

These broader implications emphasize the critical need for enhanced retrieval mechanisms, improved domain-specific fine-tuning approaches, and sophisticated context integration methods specifically designed for smaller language models. Future work could focus on optimizing IndoGPT through advanced preprocessing techniques and improved training on domain-specific corpora. Such efforts may enable IndoGPT to bridge the gap with larger multilingual models while providing a more accessible and cost-efficient solution for low-resource language applications across diverse specialized domains.



UniBot

Chat with University's Document

Silahkan bertanya!

Bagaimana ketentuan untuk pindah program studi?

LLM Result:

Berdasarkan dokumen yang ada, ketentuan untuk pindah program studi (Prodi) di UPER adalah sebagai berikut:

1. **Konsultasi dan Persyaratan Awal:**
   - Mahasiswa harus berkonsultasi dengan Dosen Wali.
   - Mahasiswa harus sudah menjalani perkuliahan paling sedikit dua semester di sebuah Prodi di UPER.
   - Mahasiswa harus menyertakan hasil penilaian bakat dan potensi di tempat yang ditentukan oleh UPER.
   - Mahasiswa harus mendapatkan surat pernyataan penerimaan dari Prodi yang dituju.
2. **Biaya:**
   - Mahasiswa yang pindah Prodi tidak perlu membayar biaya SPI kembali.
3. **Batasan Pindah Prodi:**
   - Mahasiswa hanya dapat pindah Prodi sebanyak tiga kali.
4. **Prosedur Operasional Standar:**
   - Ketentuan lebih lanjut terkait mekanisme pindah Prodi diatur dalam Prosedur Operasional Standar Perpindahan Mahasiswa antar Program Studi.

Itulah ketentuan yang harus dipenuhi oleh mahasiswa yang ingin pindah program studi di UPER.

Figure 2. Chatbot Implementation using Streamlit

ACKNOWLEDGMENT

This work was supported by Universitas Pertamina under UPEResearch Grant 2024. The authors would like to express their gratitude for the financial support provided, which made this research possible.

REFERENCES

[1] S. Panda and N. Kaur, "Exploring the viability of chatgpt as an alternative to traditional chatbot systems in library and information centers," *Library hi tech news*, vol. 40, no. 3, pp. 22–25, 2023.

[2] S. Nithuna and C. Laseena, "Review on implementation techniques of chatbot," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 0157–0161.

[3] N. Dolbir, T. Dastidar, and K. Roy, "NLP is not enough – contextualization of user input in chatbots," 2021. [Online]. Available:https://arxiv.org/abs/2105.06511.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[5] I. L. Alberts, L. Mercolli, T. Pyka, G. Prenosil, K. Shi, A. Rominger, and A. Afshar-Oromieh, "Large language models (llm) and chatgpt: what will the impact on nuclear medicine be?" European journal of nuclear medicine and molecular imaging, vol. 50, no. 6, pp. 1549–1552, 2023.

[6] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," Healthcare, vol. 11, no. 6,

2023. [Online]. Available: https://www.mdpi.com/22279032/11/6/887.

[7] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266734522300024X.

[8] J. H. Choi, K. E. Hickman, A. B. Monahan, and D. Schwarcz, "Chatgpt goes to law school," J. Legal Educ., vol. 71, p. 387, 2021.

[9] F. C. Kitamura, "Chatgpt is shaping the future of medical writing but still requires human judgment," p. e230171, 2023.

[10] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[12] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen et al., "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," arXiv preprint arXiv:2303.10420, 2023.

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[14] R. Abdurrohman, "Uji performa chatbot dengan retrieval augmented generation dan model gpt-4 untuk domain taharah berdasarkan empat imam mazhab fikih (studi kasus kitab rahmah al ummah fi ikhtilaf al a'immah)," Master's thesis, Universitas Islam Negeri (UIN) Syarif Hidayatullah Jakarta, 2024, accessed on 1 May 2024. [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/77195

[15] A. Afzal, A. Kowsik, R. Fani, and F. Matthes, "Towards optimizing and evaluating a retrieval augmented qa chatbot using llms with human-in-the-loop," in DaSH workshopNaacl, 04 2024.

[16] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. L. Khodra et al., "Indonlg: Benchmark and resources for evaluating indonesian natural language generation," arXiv preprint arXiv:2104.08200, 2021.

[17] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," Transactions of the Association for Computational Linguistics, vol. 8, pp. 726–742, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.47.

[18] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," 2023.

[19] P. Chung, "Specializing llms for domains: RAG vs fine-tuning," Towards AI, 2024, accesed on 14 April 2024. [Online]. Available: https://towardsai.net/p/machine-learning/specializing-llms-for-domains-rag-vs-fine-tuning

[20] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," arXiv preprint arXiv:2309.15217, 2023.

[21] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, "Understanding the behaviors of bert in ranking," 2019. [Online]. Available: https://arxiv.org/abs/1904.07531NRK. "Medieval helpdesk with English subtitles," YouTube, Feb. 26, 2007 [Video file]. Available: http://www.youtube.com/watch?v=pQHX-SjgQvQ. [Accessed: Jan. 28, 2014].