Retrieval-Augmented LLMs with Indonesian Clinical Trials Guidelines: A Comparative Study

Ella Raputri
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University
Jakarta, Indonesia 11480
ella.raputri@binus.ac.id

Ari Jaya Teguh
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University
Jakarta, Indonesia 11480
ari.teguh@binus.ac.id

Saffanah Nur Hidayah
General Practitioner
Zamzam Clinic
South Jakarta, Indonesia 12550
saffanah.nh@gmail.com

Feri Setiawan
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University
Jakarta, Indonesia 11480
feri.setiawan@binus.edu

Nunung Nurul Qomariyah
Computer Science Department
School of Computing and Creative Arts
Bina Nusantara University
Jakarta, Indonesia 11480
nunung.qomariyah@binus.edu

***Abstract*—Artificial Intelligence (AI), especially Large Language Model (LLM) powered chatbots, has emerged as a significant tool in our daily lives. Many industries, including the healthcare industry nowadays use AI chatbot for their operational tasks. One of such tasks is for diagnosing diseases based on the patient's symptoms, so that the medical workers can focus on more complicated tasks. However, in practice, LLM tends to generate incorrect information, so they need to be provided with more specific medical-related knowledge. Therefore, this study aims to enhance LLM diagnostic ability by integrating Indonesian Clinical Trials Guidelines using RAG (Retrieval-Augmented Generation) to give the LLMs more context on medical cases. After that, we compare the diagnostic ability of four different RAG-enhanced popular LLMs (Claude, GPT, Deepseek, and Qwen). As a contrast to the RAG LLM, we also compare their results with a model that is not enhanced using RAG as a representative of non-RAG LLM, which is Deepseek. Our research utilizes context-related metrics with the ground truth from a general medical practitioner to evaluate all five models. The result reveals that Claude has the highest performance with all metrics greater than 0.8 and GPT is the second best with all metrics greater than 0.65, followed by Deepseek with RAG, Qwen, and lastly, Deepseek without RAG, proving that RAG has a significant effect on the increase of diagnostic accuracy.**

***Index Terms*—Large Language Model (LLM), diagnosing diseases, RAG (Retrieval-Augmented Generation), context-related metrics**

I. Introduction

In recent years, Artificial Intelligence (AI) has become a topic that is intensively discussed in both research and real-world applications. Among all variations of AI, the AI chatbot powered by Large Language Model (LLM) has been a revolutionary one as it laid a foundation in creating intelligent agents that can respond to user requests using human-like languages [1]. This technological revolution can not be separated from the release of ChatGPT in 2022 [2], which deeply influenced the general public's view and curiosity in machine intelligence. Therefore, many industries started to take an interest in AI and adopt it in their operation processes.

Adoption of AI technology in various industries has been proven to increase productivity and creativity at the organizational level [3]. Among all industries, the integration of AI into the healthcare system has been a serious topic as many people now rely on AI chatbots for gathering information, including the diagnosis of diseases. According to the study in [4], more than 20 percent of health worker use ChatGPT for their health-related work. Reliance on AI chatbot, such as ChatGPT has become more prevalent these days as they can accelerate the diagnostic process, making it easier for patients and health workers to conclude a simple disease case [5]. Therefore, health workers can focus on more complex tasks that need their attention.

However, as with other technologies, AI also has its limitations. To date, many AI systems have been developed to help the healthcare industry. However, adoption and evaluation are still limited [6]. Besides that, AI chatbots tend to "hallucinate" or generate nonsensical information, so practical deployment of them to diagnose symptoms has also been a problem [7]. Hence, to enhance the reliability of the chatbot's response in a medical context, we combine LLMs with RAG (Retrieval-Augmented Generation) using information from PPPK (*Panduan Praktik Profesional Kedokteran*) or Indonesian Clinical Trials Guidelines. After that, we conduct evaluation to determine the performance of each model in the diagnostic process.

In conclusion, our research aims to compare the diagnostic ability of popular LLMs, such as Claude, Deepseek, Qwen, and GPT. To enhance every models' performance, we provide relevant medical context to them by using RAG. Therefore, it is hoped that each model can provide a more accurate response, and we can benchmark them to decide which one is the best in providing medical diagnoses.

II. Related Works

A. Large Language Models in Healthcare

Large Language Models (LLMs) are emerging in our daily lives, especially in the medical field. An LLM is used to understand the relationship between words to get the context between them. The AI model can determine a diagnosis through learning from large datasets of medical text, thus confirming their potential in the medical field [8]. The emerging LLMs influence is substantiated by the publications of 175 journals [9].

Goh [10] integrated LLMs into diagnostic workflows in random clinical trials, resulting in a more accurate and faster time to diagnose a case. Another study did a slight different way, by comparing two AI (GPT-3.5 and GPT-4) and comparing them on which model can generate a more accurate and reliable diagnosis. The outcome displayed the potential of LLMs in diagnostic support for clinics [11].

Regardless of what perks LLMs given, it still faces with issues with unreliability and accuracy, especially in high-risk scenario, where misdiagnosis can lead to fatality. A survey consisting of 550 studies displays both the possibilities of LLMs and also their drawbacks [12]. According to Chiu [13], in order to overcome the issues of accuracy and reliability of LLM, human oversight is still required to ensure there are no misdiagnoses, especially in a high-risk case.

B. Retrieval-Augmented Generation in Medical Context

Retrieval-Augmented Generation (RAG) is a method that combines external data resources like clinical guidelines, scientific literature databases, electronic health records, etc. Gargari also stated that RAG has become an important weapon to combat the challenges of standalone LLM in healthcare [14]. By providing external data resources that the models have not been programmed with, it also aids in preventing time-sensitive bias and mislabeled information caused by human errors [15].

A study by Miao [16] displayed the benefits, which are improving the accuracy and the relevance of medical outputs, on combining LLM and RAG in nephrology to diagnose a case. A similar study by Bora [17] also exhibited the improvement on the outcome by integrating RAG-based databases and LLMs with medical text, especially where computing resources are scarce.

This has led to the generalizable benchmarking study that, in the field of medical retrieval-augmented generation, provides evaluation suites dedicated to the field that account for over 7,000 sets of clinical queries [18]. Statistically significant gains have been indicated in retrieval-augmented generation following the reinforcement of the latter mechanism with RAG in the context of summarizing tasks up to 6% of the purpose of clinical information extraction as measured by clinical empirical studies [19].

C. Evaluation Metrics for Medical AI Systems

BLEURT, a learned metric based on BERT, was trained on millions of synthetic instances and followed up by fine-tuning on human scores, in order to estimate coherent judgment even in the regime of small and out-of-distribution data [20]. The syntactic and semantic mismatch combined with the surface lexical coverage makes it resilient to change of domain- an absolute necessity of medical AI evaluation where lexicon and phrasings are highly witness altering.

In that sense, BERTScore is superior to n-gram based metrics as the comparison is performed by the tokens of the lens of contextual embeddings, therefore, to the rephrasing and long dependence of the context of meaning-homogeneous sentences. These are the main characteristics that make BERTScore particularly convenient to use in situations where what counts most is the precision of meaning, as in the case of clinical use [21].

In the meantime, Semantic Answer Similarity (SAS) introduces a related concept, semantic assessment to question-answering task context because this task employs a friendlier or harsher definition of the intimate contact between a question and an answer. SAS has the capability of gauging outcomes that are comparable to human beings. Therefore, SAS may consider the same, although it produces various lexical responses to it. This specialty is useful to assist in rating the medical responses, which is most of the time in a free-text form [22].

In more recent times, there is G-EVAL allowing deployment of massive language models (GLM; GPT-4 or the like). Here, the output is assessed according to provided clinical and semantic criteria by assessors independently, and separate from use texts. While, on one hand, G-EVAL was observed closer to them in terms of human judgment for open-ended and fine-prone tasks empirically; in contrast, prompting based on LLM-full texts could be biased in evaluation it may introduce [23].

D. Medical AI Evaluation in Non-English Languages

Even though the use of LLM in medical regions focuses more on the English language, other linguistic contexts are beginning to be recognized. In Indonesia, it is impossible to practice without local language skills. Culturalization of medical AI framework needs specific problems that present language peculiarities in the Indonesian context and cultural factors, as well as the translation process to some names or terms in the medical area.

When the research issues concerning medical AI assessment are considered through the perspective of Indonesia, a gap in the existing literature of a pronounced type is observed. The few existing literature cannot cover the huge field that LLM offers. So, it is critical to design the model so that the AI can still operate like normal even if the model is offered a non-English medical text.

III. Methodology

A. Dataset Overview and Description

The current dataset consists of 45 medical cases that have been carefully selected, in order to reflect a wide range of diagnostic situations of a typical clinical practice. In every case description, there is the exact syndrome of a patient in Indonesian. The dataset also includes the respective ground truth medical diagnoses verified by competent licensed medical practitioners familiar with Indonesian medical language and Indonesian Clinical Trials Guidelines. Besides ground truth, the dataset also consists of the answer from each model. There are two different columns, which are the answer and the full

answer. The answer column contains the diseases that the model predicts or diagnoses based on the medical cases. Meanwhile, the full answer column is the full answer from the model, including the conventional greetings and other unnecessary information. The columns of the dataset can be seen in Table I.

B. Dataset Preprocessing

First of all, we prepared a set of sample questions that represent various medical cases. Then, we gather the answers (diagnoses for the medical cases) from different Large Language Models by prompting them. To increase response accuracy, we use the RAG (Retrieval Augmented Generation) approach using the previously stored Indonesian Clinical Trials Guidelines document inside the AstraDB vector database. Therefore, the models can now use the document as their reference source when giving diagnoses.

After the data gathering step, we obtained the full diagnostic answer from each model. From the full answer of each model, we extract the predicted diseases. Then, we invited a medical practitioner to establish the ground truth for each sample medical case.

The doctor provides the diagnoses for the medical cases based on her experience, but she mainly diagnoses the medical cases using Claude's diagnoses. In other words, she read all the diagnosis results from each model and saw that each of them is similar. Hence, she only read Claude's diagnoses and determined which one of the five diseases in the Claude output was the most likely based on the questions. The result that she provided is the main disease (the most potential to be the correct answer for the question) and the differential diagnosis (other possible diseases based on the symptoms in the question).

C. Model and Techniques

1) Large Language Models Selection: Four well-known LLMs were chosen to be tested under two conditions of the experiment, one being a standalone mode (no external knowledge is supplemented to LLMs) and another being enhanced mode, provided by RAG. The models that were selected are Claude 3.5 Haiku, QWEN 2.5 72B, GPT-4o mini, and Deepseek-V3.

2) RAG Implementation: The RAG system is developed on the basis of incorporating Indonesian medical knowledge contained in the Indonesian Clinical Trials Guidelines as the sources of diagnostic knowledge. The retrieval module, which uses semantic similarity searching, then retrieves the relevant medical information based on the description of the symptoms of patient. A vector database has been developed, in which the processed Indonesian Clinical Trials Guidelines content has served as dense vectors to index the database to identify similarities. The generation part then integrates the retrieved medical information with the original patient case hence providing the LLM with the contextually relevant medical information that helped in the generation of diagnostic responses.

The prompt that was used in the RAG was framed in Indonesian in order to bring out consistency of evaluated models. The Indonesian Clinical Trials Guidelines document and the question for the sample medical case itself is in Indonesian. Therefore, to generate a consistent Indonesian answer, we utilized Indonesian prompt.

D. Evaluation Techniques

Before we evaluate the candidate sentences, we first normalized them by matching them to a standardized synonyms that was listed on the synonym map. Therefore, words with similar meanings, but different wordings, such as "demam dengue" and "demam berdarah" would be standardized to become "demam berdarah dengue".

In practice, "demam dengue" and "demam berdarah dengue" are not the same, although they are both caused by the dengue virus. The difference is in the laboratory results. Based on Indonesian Clinical Trials Guidelines, "demam dengue" can be a separate diagnosis due to the lack of "demam berdarah" criteria. Meanwhile, a patient is diagnosed to have "demam berdarah" if they already satisfy the "demam berdarah" criteria. However, for our current research, we focused on more generalized results that are easier to comprehend by the public, so we standardized them to be the same.

Next, four metrics of evaluation were complemented to make a complete analysis of performance:

- BERT (Bidirectional Encoder Representations from Transformers) Score: Calculate the similarity score between the candidate sentence and ground truth based on the sum of cosine similarities between the pre-trained BERT contextual embeddings of their tokens [24].
- BLEURT: a trained evaluation metric based on BERT that conveys human-level judgment of reference versus candidate sentence. BLEURT is not calibrated (normalized), so we plotted a distribution graph based on BLEURT and used it to compare the model's performance [25].
- SAS (Semantic Answer Similarity): A cross-encoder-based evaluation metric that estimates the semantic meaning of the answer or candidate sentence and compare it with the reference sentence [22].
- LLM-based Judge or G-Eval: prompt-based evaluator that utilizes the GPT LLM family as the judge to simulate human quality judgment [23]. In our evaluation process, we utilized the GPT-4-Turbo.

E. Research Flow Overview

The research was conducted using an experimental research design with a systematic procedure that compared the conditions of all models consistently. All four models are presented with uniform RAG setups and the same sample questions. However, we also include an additional model, which is Deepseek without using RAG as a baseline to compare RAG and non-RAG LLM diagnostic capability. The instructions involved in carrying out the assessments were also uniform, and the prompts were designed in a similar manner, so that the evaluation process is considered 'fair' for all models. For the overview of our research flow, we can see it in the flowchart in Figure 1.

TABLE I: Dataset's Details

Columns	Description	Data Type
No	the serial number of the medical case	float
Question	the medical case statement in the form of question	object (string)
Dr Answer	the ground truth for that medical case	object (string)
Claude Answer	diseases diagnosed by Claude for that medical case	object (string)
Qwen Answer	diseases diagnosed by Qwen for that medical case	object (string)
GPT Answer	diseases diagnosed by GPT for that medical case	object (string)
Deepseek RAG Answer	diseases diagnosed by Deepseek for that medical case with RAG	object (string)
Deepseek Non RAG Answer	diseases diagnosed by Deepseek for that medical case without using RAG	object (string)
Claude Full Answer	full answer from Claude	object (string)
Qwen Full Answer	full answer from Qwen	object (string)
GPT Full Answer	full answer from GPT	object (string)
Deepseek RAG Full Answer	full answer from Deepseek using RAG	object (string)
Deepseek Non RAG Full Answer	full answer from Deepseek without using RAG	object (string)

TABLE II: LLM Performance

Evaluation Metrics	Claude	Qwen	GPT	Deepseek with RAG	Deepseek without RAG
BERT's Precision	0.860	0.814	0.820	0.812	0.791
BERT's Recall	0.937	0.850	0.874	0.875	0.843
BERT's F1	0.896	0.831	0.846	0.842	0.816
SAS	0.962	0.494	0.688	0.653	0.414
LLM-based Judge	0.859	0.629	0.758	0.771	0.687

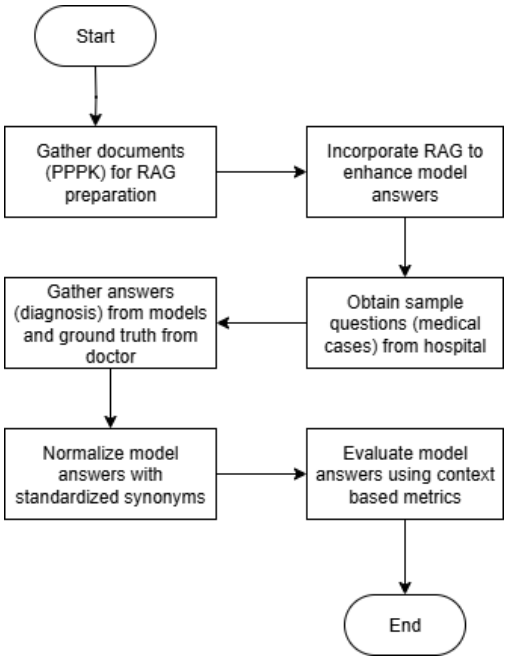

Fig. 1: Research Flow Flowchart

IV. Results and Discussion

Table II shows the mean of the evaluation score for each model when they are benchmarked with the medical case dataset.

From the table, we can see that overall, Claude's performance is the best, followed by GPT, Deepseek with RAG, Qwen, and lastly, Deepseek without RAG. The reason why Claude has the best performance is highly related to the ground truth establishment method that the medical practitioner used. The doctor determined the ground truth mostly by evaluating Claude's diagnosis. Therefore, it is not surprising that Claude has the highest performance among all LLMs. Meanwhile, Deepseek without RAG is as expected to rank as the last one, as it does not have any additional context given by RAG.

From the BERT scores, it can also be inferred that all LLMs have better recall than precision in diagnosing diseases. In other words, the models tend to include many relevant (positive) terms in their diagnosis, but are inclined to over-predicting positives. Hence, resulting in a lower precision score than the recall score. However, in medical diagnosis case, higher recall score is more favorable than higher precision score, because it is better to have a false positive than a false negative, so that there is no disease that is missed in the diagnosis.

Next, we explore the BERT F1 and LLM-based Judge result. As stated before, the order of LLMs based on their performance is Claude, GPT, Deepseek with RAG, Qwen, then Deepseek without RAG. BERT F1 score reveals a similar score for each model, implying that although some models are better, the difference is not that significant. This conclusion is also supported by the LLM-based Judge result, where the scores for each model are close. The scores are still in the range of 0.6 to 0.9, although the discrepancy between each score is higher. However, as we can see in Table II, unlike other metrics, LLM-based Judge gives a higher score to Deepseek without RAG than Qwen, even though Qwen is already enhanced with RAG. A possible cause for this behaviour is due to the bias or hallucination of the LLM judge, which gives a higher score to answers that match more with the evaluation prompt.

Meanwhile, SAS also has a similar pattern to BERT F1, but shows a greater difference between each model's score. Claude has a relatively high score, while models such as Qwen and Deepseek without RAG obtained a low score of 0.4. This distinction indicates that cross-encoder approach that is used in SAS evaluation metrics

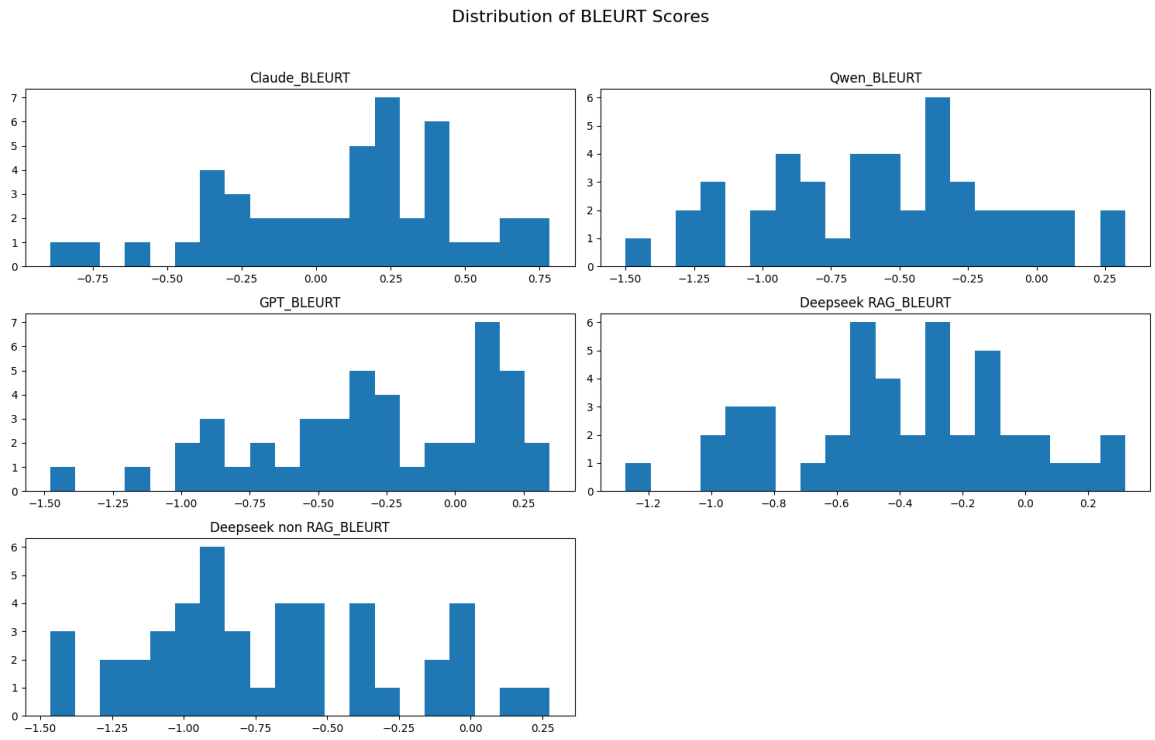

Fig. 2: BLEURT Distribution Histogram for Each LLM

impose stricter criteria when evaluating similarity. Cross-encoders process sentence pairs jointly to capture the complex interaction between them, thereby making them more sensitive when penalizing subtle mismatches. As a result, the score discrepancy is higher.

Besides scores such as BERT, SAS, and LLM-based Judge, we also evaluate each model's performance with BLEURT. However, as stated in the Methodology part, the BLEURT score is not normalized, so we plotted the histogram to visualize the distribution of the BLEURT score of each model's diagnosis for each question. The histogram can be seen in Figure 2.

The histograms in Figure 2 reveal the similar outcome as the other evaluation metrics. The more left-skewed a distribution, the better the result is because it means that there are many positive scores. From the histograms, we can see that Claude has the best result because the distribution is skewed towards a positive score. Then, we have GPT which is the second best, it is left skewed, but still produces a high negative score (about -1.50). Next are Deepseek with RAG and Qwen. Both histograms are similar, but Deepseek with RAG is slightly better than Qwen because it does not have a strongly negative value. Lastly, there is Deepseek without using RAG. The histogram shows that the model prediction is unstable, with some predictions being good (positive score), but some are strongly negative. This result demonstrated that RAG can also improve the stability of the answers for LLMs.

V. Conclusion and Future Work

In conclusion, this study analyzes the medical diagnosis performance of each popular LLM by collaborating with a medical practitioner (doctor) for the ground truth establishment, then uses it as a base to benchmark each model. Each model is also enhanced by RAG using the *Panduan Praktik Profesional Kedokteran* or Indonesian Clinical Trials Guidelines to provide context related to disease symptoms and diagnosis. Then, we evaluate each model using context-based metrics.

After reviewing the metrics, the result showed that the order of models based on their diagnosis performance is Claude, GPT, Deepseek with RAG, Qwen, and then followed by Deepseek without RAG. Claude has the best performance because the doctor used it as the main evaluation reference. Meanwhile, Deepseek without RAG scored the lowest because it is not supported by the context given in RAG.

Looking ahead, future research could focus on evaluating more diverse models and datasets. We plan to gather more data to be used as the medical cases in the benchmark dataset. Besides that, collaborating with more medical practitioners to establish the ground truth for each medical case is also beneficial, so that we can obtain a more representative ground truth. Choosing a specific medical dataset for evaluation, such as a dataset that only contains heart diseases, can also be used to benchmark each model to determine which model is the best in diagnosing diseases in a specific medical field. Additionally, we can also incorporate more evaluation metrics to provide a more comprehensive view of each model's performance, or more LLMs to be tested, so that we can see their performance.

Supplementary Codes

All codes that were used in our research can be accessed in our GitHub repository through the link: https://github.com/Ella-Raputri/LLMDiagnosisPPPK.

CONTRIBUTORSHIP

Ella Raputri: Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. **Ari Jaya Teguh**: Methodology, Software, Investigation, Writing - Original Draft. **Saffanah Nur Hidayah**: Validation, Resources. **Nunung Nurul Qomariyah**: Conceptualization, Writing - Review & Editing, Supervision, Project Administration. **Feri Setiawan**: Conceptualization, Writing - Review & Editing, Supervision, Project Administration.

REFERENCES

[1] S. Paliwal, V. Bharti, and A. K. Mishra, "Ai chatbots: Transforming the digital world," in *Recent trends and advances in artificial intelligence and internet of things*. Springer, 2019, pp. 455–482.

[2] F. Sun, "Chatgpt, the start of a new era," *A Bright and Gloomy Future. In https://feisun. org/2022/12/23/-chatgpt-the-start-of-a-new-era (ultima consultazione: 27/03/2023)*, 2022.

[3] H. Taherdoost and M. Madanchian, "Artificial intelligence and knowledge management: Impacts, benefits, and implementation," *Computers*, vol. 12, no. 4, p. 72, 2023.

[4] M.-H. Temsah, F. Aljamaan, K. H. Malki, K. Alhasan, I. Altamimi, R. Aljarbou, F. Bazuhair, A. Alsubaihin, N. Abdulmajeed, F. S. Alshahrani, R. Temsah, T. Alshahrani, L. Al-Eyadhy, S. M. Alkhateeb, B. Saddik, R. Halwani, A. Jamal, J. A. Al-Tawfiq, and A. Al-Eyadhy, "Chatgpt and the future of digital health: A study on healthcare workers' perceptions and expectations," *Healthcare*, vol. 11, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2227-9032/11/13/1812

[5] I. Altamimi, A. Altamimi, A. S. Alhumimidi, A. Altamimi, and M.-H. Temsah, "Artificial intelligence (ai) chatbots in medicine: A supplement, not a substitute," *Cureus*, Jun. 2023. [Online]. Available: http://dx.doi.org/10.7759/cureus.40922

[6] S. Reddy, W. Rogers, V.-P. Makinen, E. Coiera, P. Brown, M. Wenzel, E. Weicken, S. Ansari, P. Mathur, A. Casey, and B. Kelly, "Evaluation framework to guide implementation of ai systems into healthcare settings," *BMJ Health amp; Care Informatics*, vol. 28, no. 1, p. e100444, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1136/bmjhci-2021-100444

[7] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating LLM hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1827–1843. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.123/

[8] S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, Y. Fang, S. Ding, J. Wang, K. Xu, L. Xia, J. Yeung, D. Zha, D. Cai, G. B. Melton, M. Lin, and R. Zhang, "Large language models for disease diagnosis: a scoping review," *npj Artificial Intelligence*, vol. 1, 6 2025.

[9] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, 11 2024.

[10] E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. Olson, A. Rodman, and J. H. Chen, "Large language model influence on diagnostic reasoning: A randomized clinical trial," *JAMA Network Open*, vol. 7, 10 2024.

[11] A. Ríos-Hoyo, N. L. Shan, A. Li, A. T. Pearson, L. Pusztai, and F. M. Howard, "Evaluation of large language models as a diagnostic aid for complex medical cases," *Frontiers in Medicine*, vol. 11, 2024.

[12] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang, J. Gao, Y. G. S. Wang, J. ming Ji, Z. Qiu, M. Li, C. Qian, T. Guo, S. Ma, Z. Wang, Z. Guo, Y. Lei, C. Shao, W. Wang, H. Fan, and Y. D. Tang, "The application of large language models in medicine: A scoping review," *iScience*, vol. 27, 5 2024.

[13] W. H. K. Chiu, W. S. K. Ko, W. C. S. Cho, S. Y. J. Hui, W. C. L. Chan, and M. D. Kuo, "Evaluating the diagnostic performance of large language models on complex multimodal medical cases," *Journal of Medical Internet Research*, vol. 26, 2024.

[14] O. K. Gargari and G. Habibi, "Enhancing medical ai with retrieval-augmented generation: A mini narrative review," 1 2025.

[15] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, J. C. L. Ong, D. S. W. Ting, and N. Liu, "Retrieval-augmented generation for generative artificial intelligence in health care," *npj Health Systems*, vol. 2, 1 2025.

[16] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. G. Valencia, and W. Cheungpasitporn, "Integrating retrieval-augmented generation with large language models in nephrology: Advancing practical applications," 3 2024.

[17] A. Bora and H. Cuayáhuitl, "Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications," *Machine Learning and Knowledge Extraction*, vol. 6, pp. 2355–2374, 12 2024.

[18] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," *Association for Computational Linguistics*, pp. 6233–6251, 8 2024. [Online]. Available: https://github.com/Teddy-XiongGZ/MedRAG https://aclanthology.org/2024.findings-acl.372/

[19] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, vol. 156, 8 2024.

[20] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *Association for Computational Linguistics*, pp. 7881–7892, 6 2020. [Online]. Available: http://github.com/google-research/

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *International Conference on Learning Representations*, 2 2020. [Online]. Available: http://arxiv.org/abs/1904.09675

[22] J. Risch, T. Möller, J. Gutsch, and M. Pietsch, "Semantic answer similarity for evaluating question answering models," in *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, A. Fisch, A. Talmor, D. Chen, E. Choi, M. Seo, P. Lewis, R. Jia, and S. Min, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 149–157. [Online]. Available: https://aclanthology.org/2021.mrqa-1.15/

[23] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/

[24] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[25] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," in *Proceedings of ACL*, 2020.