Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Study on the Performance of a RAG-Augmented Interview Chatbot

Muhammad Rifat Anwar[a], Irsyad Arif Mashudi[b], & Yoppy Yunhasnawa[c]
[a,b,c]*Department of Information Technology, State Polytechnic of Malang, Indonesia*
irsyad.arif@polinema.ac.id

Abstract: Along with the increasing importance of communication skills for the jobseekers, many students face challenges in preparing for interviews, both in terms of knowledge, experience, and self-confidence. This research explores the effectiveness of a RAG-augmented chatbot for job interview preparation, combining information retrieval and large language models (LLMs) to generate contextually relevant responses. Two LLM models, Google/gemma-2-2b-it and deepseek-ai/DeepSeek-V3, were compared across three scenarios: True-True (TT), True-False (TF), and False-False (FF). Results show that DeepSeek-V3 outperformed Google/gemma-2-2b-it in generating more accurate and relevant responses, though both models struggled in False-False scenarios. The System Usability Scale (SUS) testing indicated that the chatbot was perceived as easy to use and effective, with scores above the average threshold of 68. However, feedback highlighted concerns regarding response time and the complexity of some features. Overall, the findings suggest that the RAG-based chatbot is a promising tool for interview preparation, though improvements in model performance, system responsiveness, and interface simplicity are recommended for future development.

Keywords: *AI Chatbot, Retrieval-Augmented Generation (RAG), Job Interview*

1. Introduction

Job interviews remain a critical component of the recruitment process, often determining whether a candidate is deemed suitable for a position (Laiq & Dieste, 2020). Strong communication skills are essential for interview success, yet many students and recent graduates struggle with interview preparation due to limited knowledge, lack of practical experience, and low self-confidence. These challenges hinder their transition into the professional workforce.

In recent years, Artificial Intelligence (AI) has shown remarkable progress, enabling machines to perform tasks that traditionally required human intelligence (Firdaus, 2024). Within the field of human resource development, AI offers opportunities to automate and personalize training and assessment processes. Specifically, AI can streamline job interview preparation by providing students with interactive, on-demand coaching tools (Chaka, 2023; Fraij & László, 2021).

One promising tool is ChatGPT, a conversational agent that leverages natural language processing to simulate human-like dialogue. When combined with Retrieval-Augmented Generation (RAG), ChatGPT can generate more accurate and contextually relevant responses by integrating internal language model knowledge with external data sources (Antico et al., 2024). This hybrid approach overcomes the limitations of standard transformer models, which struggle with incorporating up-to-date external information (Pearce et al., 2023).

Given these developments, AI-powered chatbots have the potential to support students in building confidence and communication skills through simulated interview experiences (Krassmann et al., 2019). This study aims to evaluate the effectiveness of a RAG-augmented chatbot in facilitating job interview preparation for students. The findings are expected to contribute to the design of AI-driven educational tools that bridge the gap between academic learning and career readiness.

2. Research Methodology

2.1 Data Collecting and Processing

The first step in our methodology involves constructing the chatbot model using the Hugging Face interview dataset. This dataset consists of a broad range of commonly used generic interview questions and answers, designed to simulate a variety of job interview scenarios. It covers diverse topics, including work experience, skills, motivations, and personality traits, to ensure that the chatbot can engage users across multiple areas. By incorporating questions and answers from a wide array of themes, the dataset ensures that the chatbot is capable of handling varied user interactions and avoids an overemphasis on any single topic.

Once the dataset is collected, the next step is data cleaning. This involves a series of processes aimed at enhancing the dataset's quality and ensuring its readiness for model training. First, we remove irrelevant information that does not contribute to the interview context. We also address any data duplication, ensuring that the dataset contains unique and diverse content. The data is then converted into a standardized format that is compatible with the model training pipeline. Additionally, we perform text tokenization, which breaks down the text into smaller, manageable units (tokens), making it easier for the model to process. Stopword removal is also carried out to eliminate common, non-informative words that could negatively impact the model's efficiency and accuracy.

After the data has been cleaned and pre-processed, a detailed analysis is performed to gain insights into its variety and structure. This analysis helps to assess the distribution of topics, the depth of coverage in different areas, and the overall balance within the dataset. These insights are crucial for ensuring that the dataset effectively supports the creation of a well-rounded chatbot that can simulate realistic and varied job interview scenarios.

2.2 System development

The system design utilizes a Retrieval-Augmented Generation (RAG) approach, which combines two core components: retrieval and generation. In the retrieval phase, the system searches for relevant information from the pre-processed dataset. Then, the Deepseek-AI/DeepSeek-V3 model generates context-based responses based on the retrieved data. This allows the chatbot to provide more accurate and relevant answers tailored to the user's responses. The use of Deepseek-AI/DeepSeek-V3 in this system design offers a more specialized and advanced retrieval mechanism compared to ChatGPT. Deepseek is designed specifically for handling large-scale datasets and optimizing information retrieval, ensuring that the chatbot's responses are highly relevant and context-aware. This capability allows it to process domain-specific queries more efficiently and generate answers that are closely aligned with the user's needs, offering a higher level of accuracy in comparison to a generalized model like ChatGPT.

The system workflow begins by asking interview questions in Indonesian using text-to-speech (TTS) technology. The questions are selected from a pre-compiled dataset and are specifically designed for students or recent graduates. The user answers the questions via speech-to-text (STT), which converts their voice responses into text. The system then searches for the most relevant information in the dataset using FAISS (Facebook AI Similarity Search), ensuring the answer aligns with the context of the user's input.

Once relevant information is retrieved, the LLM generates a complete and contextually accurate response. The system then evaluates the generated answer based on relevance, completeness, and clarity, providing feedback through voice. After all questions have been answered, the system delivers a final evaluation, including a pass/fail result based on the cumulative scores of all responses, offering users insights into their performance.

2.3 Testing

To evaluate the effectiveness of the chatbot, we conducted a comprehensive testing process focusing on both the system's performance and user experience. The RAG system was assessed through three categories: True-True (TT), True-False (TF), and False-False (FF). In the TT category, the user's answer matched the ideal answer, indicating that the model retrieved and generated relevant and accurate information. In TF, the user's answer was relevant but did not match the ideal, suggesting that the model understood the context but showed some variability in response. The FF category indicated a failure in both relevance and correctness, helping to pinpoint the system's weaknesses. We implement these testing by comparing them to the ideal responses. The system considers three different scenarios as described in Table 1.

Table 1. *Testing Scenario*

Scenario	Description
1 (TT)	The user's answer matches and is relevant to the ideal answer, indicating that the model successfully retrieved and generated accurate and relevant information.
2 (TF)	The user's answer is relevant but differs from the ideal response, suggesting that while the model understands the context, there is some variation in the answer generation.
3 (FF)	The user's answer is neither relevant nor connected to the ideal answer, highlighting a failure in the model's ability to retrieve or generate appropriate responses.

For the LLM side, we compare two different LLM for this model: Deepseek-AI/DeepSeek-V3 and Google gemma. By comparing two different Large Language Models (LLMs), we could analyze their performance in terms of accuracy, consistency, and flexibility, ultimately determining which model was more suitable for real-world application based on the need for precision or adaptability.

In addition to the RAG testing, user experience was evaluated using the System Usability Scale (SUS). Participants interacted with the chatbot and provided feedback on several aspects, including the accuracy of the system's responses, the relevance and context of answers, the fairness of the system's evaluations, the ease of use, and the effectiveness of the feedback. These insights were crucial for understanding the practical usability of the chatbot and how it could be further refined to meet user expectations.

Finally, we implemented a pass/fail evaluation for the interview preparation system. Instead of relying solely on a single final score, the system uses a holistic approach, combining semantic similarity to ideal answers (using cosine similarity) and the quality assessment by the LLM. With weighted scores of 70% from the LLM and 30% from cosine similarity, the final score is calculated for each answer. A passing threshold of 55% was set, and participants were assessed based on both the quality of their answers and the overall learning process. This approach ensures a more comprehensive evaluation, focusing not just on the final outcome but also on the quality of user interactions throughout the interview process, aligning with the view that assessments should support student development rather than just providing a uniform final measurement.

.

3. Result and Discussion

The RAG testing results revealed clear differences in the performance of the two models across three scenarios. In Scenario 1 (True-True), where the user's answer was relevant and aligned with the ideal response, the deepseek-ai/DeepSeek-V3 model outperformed the Google/gemma-2-2b-it model, scoring 7.34/10 with a cosine similarity of 0.97, compared to 6.33/10 and 0.95 for the latter. In Scenario 2 (True-False), where the user's answer was relevant but did not match exactly, DeepSeek-V3 again scored higher (8.00/10, 0.42 cosine similarity), whereas Google/gemma-2-2b-it scored 5.50/10 with a 0.53 cosine similarity. However, in Scenario 3 (False-False), where the user's answer was neither relevant nor aligned, both models struggled, with DeepSeek-V3 scoring 1.89/10 and Google/gemma-2-2b-it scoring 1.83/10. The comparison can be seen in Table 2.

These results indicate that DeepSeek-V3 consistently provided more relevant and accurate responses across all scenarios, particularly in Scenario 1 and Scenario 2, where precision and relevance were key. However, both models performed poorly in Scenario 3, highlighting challenges in handling responses that significantly deviated from the ideal answers. Overall, DeepSeek-V3 demonstrated superior performance, though further optimization is needed to improve handling of less relevant responses and ensure greater consistency in all scenarios.

Table *2. Testing Results*

Scenario	LLM	Average score	Average Cosine Similiarty
1	Google/gemma-2-2b-it	6.33 / 10	0.95 / 1
1	deepseek-ai/DeepSeek-V3	7.34 / 10	0.97 / 1
2	Google/gemma-2-2b-it	5.50 / 10	0.53 / 1
2	deepseek-ai/DeepSeek-V3	8.00 / 10	0.42 / 1
3	Google/gemma-2-2b-it	1.83 / 10	0.21 / 1
3	deepseek-ai/DeepSeek-V3	1.89 / 10	0.18 / 1

The **System Usability Scale (SUS)** testing revealed that the **RAG-based chatbot** generally performed well in terms of usability. With an average score of **68**, the system falls just above the average SUS threshold, indicating that the application is perceived as relatively easy to use and effective by the majority of respondents. Most participants rated the system above this threshold, suggesting that the chatbot was user-friendly and met their expectations for an interview simulation tool. The score distribution can be seen in Figure 1.

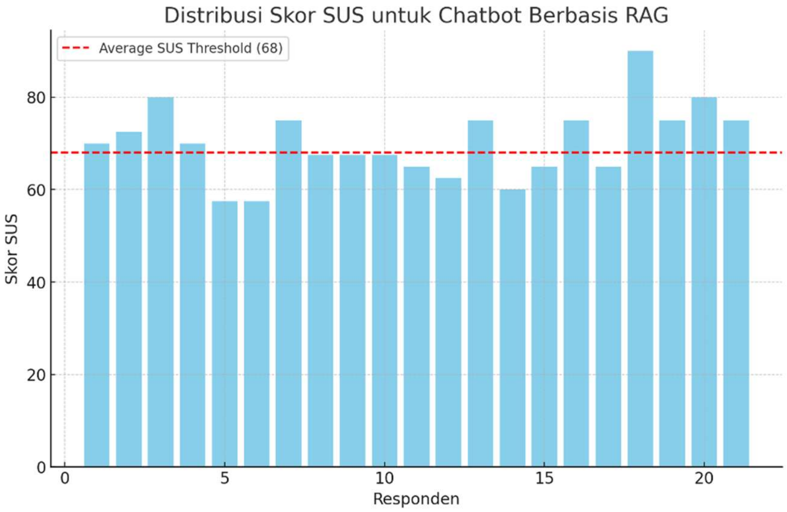

Figure 1. *SUS Result*

However, despite the positive overall score, several respondents provided feedback highlighting concerns regarding the response time and the complexity of certain features. This feedback suggests there is potential for improving the system's interface and processing speed to enhance user experience. Streamlining the navigation flow and optimizing the system's response times could help address these concerns and further improve usability. These results offer valuable insights into the system's effectiveness and accessibility, serving as a solid foundation for further development aimed at refining the interface and overall system performance, ultimately achieving a higher level of user satisfaction.

4. Conclusion and Recommendations

4.1 Conclusion

This study evaluated the performance and usability of a RAG-augmented chatbot for job interview preparation, focusing on both the system's accuracy in answering questions and its overall user experience. The RAG testing results demonstrated that the deepseek-ai/DeepSeek-V3 model outperformed the Google/gemma-2-2b-it model in generating more relevant and accurate answers across most scenarios. DeepSeek-V3 showed better precision in Scenario 1 (True-True) and Scenario 2 (True-False), although both models struggled with Scenario 3 (False-False), where the responses deviated significantly from the ideal answers. The system's SUS results indicated that the chatbot was generally well-received by users, with most respondents rating the application as easy to use and effective. However, some concerns were raised regarding response time and the complexity of certain features, which highlight areas for improvement.

Overall, the findings suggest that the RAG-based chatbot is a promising tool for interview preparation, offering users an engaging and effective way to simulate job interviews. While the system performed well in most scenarios, there is still room for improvement, particularly in enhancing the model's ability to handle responses that deviate significantly from the ideal answers and optimizing the system's response time and interface for better usability.

4.2 Recommendations

Future research should focus on the following areas to improve the effectiveness and usability of the RAG-based chatbot:

1. Enhancing Model Performance: Further fine-tuning of the model and optimization of the retrieval strategy could help reduce the frequency of False-False responses, especially in complex or off-topic scenarios. This could involve improving the dataset's coverage and the model's ability to handle diverse user inputs.
2. Interface and Usability Improvements: As indicated by the SUS testing feedback, optimizing the system's response time and simplifying the user interface would improve overall user satisfaction. Future research could explore the integration of more user-friendly features, such as voice command customization and faster processing capabilities, to create a more seamless experience.
3. Real-Time Adaptability: Investigating ways to make the system more adaptable to dynamic, real-time interactions could enhance its effectiveness. For example, incorporating a more flexible feedback mechanism that adjusts based on the user's performance in different interview scenarios could improve the learning process.
4. Expanding Use Cases: Future research could explore the potential for using the RAG-based chatbot for broader applications beyond job interview preparation, such as training for academic presentations, language learning, or other professional scenarios where interview-like simulations are beneficial.
5. User Feedback and Long-Term Testing: Conducting longitudinal studies that track user progress over time could provide deeper insights into how the system affects user learning and confidence in real-world interview situations. This would also help identify additional areas for improvement and system refinement based on long-term user engagement.

By addressing these areas, future research could further optimize the RAG-based chatbot, enhancing its performance and usability to support a wider range of users and applications.

5. References

Antico, C., Giordano, S., Koyuturk, C., & Ognibene, D. (2024). Unimib Assistant: Designing a student-friendly RAG-based chatbot for all their needs (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2411.19554

Chaka, C. (2023). Generative AI Chatbots - ChatGPT versus YouChat versus Chatsonic: Use Cases of Selected Areas of Applied English Language Studies. International Journal of Learning, Teaching and Educational Research, 22(6), 1–19. https://doi.org/10.26803/ijlter.22.6.1

Firdaus, A. (2024). Implementasi Artificial Intelligence dalam Rekrutmen: Manfaat dan Tantangan di Industri 4.0. J-MAS (Jurnal Manajemen Dan Sains), 9(2), 1615. https://doi.org/10.33087/jmas.v9i2.2083

Fraij, J., & László, V. (2021). A Literature Review: Artificial Intelligence Impact on the Recruitment Process. International Journal of Engineering and Management Sciences, 6(1).

Krassmann, A. L., Nunes, F. B., Bessa, M., Tarouco, L. M. R., & Bercht, M. (2019). Virtual Companions and 3D Virtual Worlds: Investigating the Sense of Presence in Distance Education. In P. Zaphiris & A. Ioannou (Eds.), Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration (Vol. 11591, pp. 175–192). Springer International Publishing. https://doi.org/10.1007/978-3-030-21817-1_14

Laiq, M., & Dieste, O. (2020). Chatbot-based Interview Simulator: A Feasible Approach to Train Novice Requirements Engineers. 2020 10th International Workshop on Requirements Engineering Education and Training (REET), 1–8. https://doi.org/10.1109/REET51203.2020.00007

Pearce, K., Alghowinem, S., & Breazeal, C. (2023). Build-a-Bot: Teaching Conversational AI Using a Transformer-Based Intent Recognition and Question Answering Architecture. Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 16025–16032. https://doi.org/10.1609/aaai.v37i13.26903