Retrieval Augmented Large Language Model Chatbots in Higher Education: A Study on University Open Days

Akash K. Shanthakumar, Farhad Fassihi-Tash, Ahmad Lotfi, and Jordan J. Bird

Department of Computer Science,
Nottingham Trent University,
Nottingham, Nottinghamshire, United Kingdom
Corresponding author: jordan.bird@ntu.ac.uk

Abstract. This study explores the implementation of Retrieval Augmented Large Language Models (LLMs) toward enhancing prospective student engagement during University open days. This work proposes the use of local, 4-bit quantised LLMs such as Microsoft Phi3, Meta's LLaMa3, and Mistral AI's Mistral to facilitate interactive dialogue about the Department of Computer Science at Nottingham Trent University. The proposed approaches are validated through the use of synthetic data generation via the RAGAS framework with additional expert human-in-the-loop oversight. We argue that the current state of the art which often involves ChatGPT as a sole validator is problematic, and we propose the use of an ensemble of multiple local validators that operate when a quorum is present to increase robustness. The results indicate that, while the chatbots are successful in providing the correct information, refining data relevance remains an open issue. Mistral demonstrated the highest performance in terms of information accuracy and coherence of responses, however, it was also the slowest at generating responses.

Keywords: Human-Computer Interaction, Large Language Models, AI in Education

1 Introduction

Engagement strategies during University Open Days are pivotal in shaping perceptions and decisions of prospective students about higher education. Often, especially for schools and colleges in the STEM field, relevant state-of-the-art technology is showcased with the aim to inspire young minds ahead of higher education. Recent technological advancements in Human-Computer Interaction such as Large Language Models (LLMs), including ChatGPT, have emerged as particularly interesting points of interactivity given their ability to generate human-like text. Additional advancements in other fields of Generative AI including neural voice synthesis enable even further layers of natural interaction, forming applications that were once seen as science fiction. Natural interaction

with machines is a key goal of Artificial Intelligence, and thus we argue that this technology should be showcased to prospective students. This is for two main reasons; first, from an institutional point of view to provide demonstrations of recent technological advancements, and second, from a pedagogical point of view to provide exposure to young people on the applications of Generative AI, a field of research that will likely be a growing presence both during and after their education. Given the rapid advancements in open source alternatives to proprietary models, the goal of this study is to assess the capability of local LLMs to engage in Question-Answering regarding queries about Nottingham Trent University's Department of Computer Science, including University programmes, facilities, and staff. The significance of this approach is not only to streamline information dissemination while showcasing novel technology, but also to enhance it by overcoming issues such as the occupation of human resources [1]. The field of research is extremely young, with several open issues and many studies still undergoing peer review. One of the main open issues in the field is validation, wherein the current approach to validate an LLM is to use another LLM for the task [2,3]. Although cases of LLM-based peer review have been noted as promising, we argue that this approach could be improved, and aim to overcome these issues with two proposed approaches. First, human expert oversight regarding the synthetic datasets generated to test question answering capabilities, where University lecturers assess the dataset manually. Second, we propose to use several Large Language Models in the assessment of answers given by the chatbot. In doing so, it was observed during this study that anomalies were produced in some of the validation exercises, which can be overcome by implementing an ensemble approach limited by a quorum.

The remainder of this work is as follows. Section 2 explores the current state-of-the-art in the field by reviewing the relevant literature. Following a review of the literature, Section 3 describes the methodology followed by this work, where the results are discussed in Section 4. Finally, Section 5 concludes this study and proposes future work based on the findings.

2 Related Work

According to Foster's research [4], university open days benefit from interactivity rather than static display. For example, Sanson [5] showed that the use of Virtual Reality (VR) was promising in inspiring high school students interested in continuing their education in STEM fields. Inspired by these findings, the goal of this study is to explore the initial research required for the development of an AI-based interactive framework for use during Open Days to enhance engagement and shape perceptions. During 2008, when deep learning was of growing interest in the academic community, Bellotto, Rowland, and Hu [6] proposed Lux, a receptionist robot that used human tracking algorithms and provided key information to the host institution through a touchscreen interface. More recently, in [1], authors proposed the use of a bag-of-words preprocessing approach for question classification using a deep neural network for a University open day

chatbot. The purpose of this classification was to provide an appropriate answer using a rule-based approach. The authors note that a rule-based approach is efficient but limited and that robustness could be improved by exploring complex questions beyond those provided. A similar rule-based approach was followed by Nguyen et al. [7], where a deep learning chatbot was able to answer 90.28% of over 50,000 user questions about university admissions.

The limitations of rule-based approaches to question-answering are widespread, such as data scarcity having limitations on the quality, or even mere possibility, of a response being produced given a non-deterministic question [8]. Related work has noted the use of generative AI for alleviating issues of data scarcity [9], and recent studies on the use of Large Language Models [10, 11] often explore solutions to this problem by providing context to the chatbot from a central repository. That is, the use of question/answering based on retrieved context from ground truth data. The goals of this state-of-the-art surround answering a wider range of questions (given adequate NLP ability and source information) to answer a wider range of questions about such source material than a rule-based approach since the provided source material (e.g. a textbook) is contextualised in relation to the question that is posed about it. Given the *reverse engineering* nature of the algorithms, it is also possible to automate the process of building an example set of questions and answers. Although the notion of automatic synthetic data engineering would require human oversight to ensure adherence to ethical standards, it also leads to the viability of being able to generate a much richer dataset than a fully manual rule-based data engineering approach.

3 Method

This section details the method followed by this study in research on university Open Day chatbots with question answering ability. First, data collection and generation of a synthetic dataset are described, followed by validation strategies for the selected approaches.

3.1 Dataset Collection and Generation

Data for this study were collected from two websites from Nottingham Trent University. This includes the homepage for the Computer Science courses offered by the University[1] and the homepage for the Department of Computer Science[2]. Each of the two webpages are scraped using python via the Selenium library, and are not edited towards enabling autonomous compatibility with potential future changes to the department website, such as the offering of new courses or changes in staff roles, etc.

Following the collection of raw text from the two aforementioned webpages, a synthetic Question/Answer dataset is generated through the use of LLMs

[1] https://www.ntu.ac.uk/course/computer-science, accessed: 11/06/2024
[2] https://www.ntu.ac.uk/study-and-courses/academic-schools/science-and-technology/Computer-Science, accessed: 11/06/2024

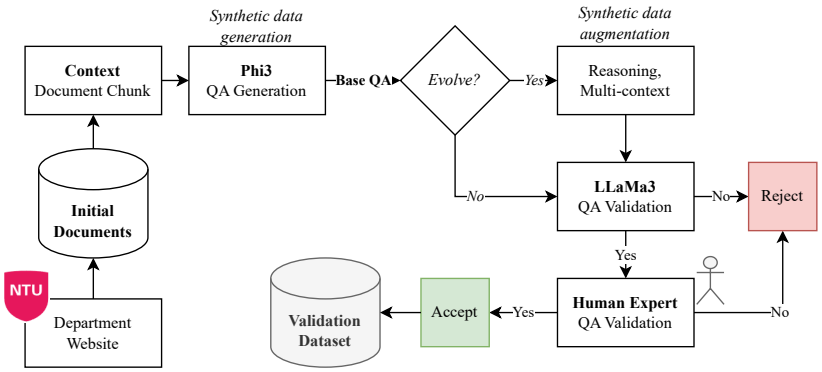

Fig. 1: The approach followed by this work for dataset generation which involves RAGAS synthesis with additional human oversight.

to create validation examples, implemented through the Retrieval Augmented Generated Assessment (RAGAS) framework [2]. As shown in Figure 1, Microsoft Phi3 [12] is first used to generate a synthetic question/answer pair from the given chunk retrieved from the dataset, which is arbitrarily set to 1000 characters. Following this, evolution of the question may take place; 50% of the questions are set to simple, that is, there is no change. 25% of the questions are set to *reasoning* (the question is rephrased to enhance the need for reasoning), and 25% of the questions are set to *multi-context* (the question is rephrased to require information from multiple related chunks to answer). Meta's LLaMa3 is then used to validate the question, with rejected questions being rejected[3]. This process is executed for a maximum of 1000 QA pairs. Following the autonomous process, we then argue for human oversight to both ensure the relevancy of and reduce redundancy within the final dataset. In this study, two Senior Lecturers in Computer Science from the host institution manually observe the dataset, removing all irrelevant QA pairs as well as duplicate questions.

3.2 Model Training and Validation

Following the generation of synthetic QA pairs, autonomous validation, and human expert validation, the final dataset is then used for model validation.

Three state of the art LLMs are selected for the Open Day chatbot: Microsoft Phi3 [12] (3.82B parameters), Meta LLaMa3 [13] (8.03B prameters), and Mistral AI's Mistral (7.25B parameters) [14]. The models are selected for their range of parameters, open source nature, and promising performance in the current state-of-the-art. Figure 2 shows the prompt used to initiate the conversation between

[3] Further details on RAGAS prompts are described at: https://github.com/explodinggradients/ragas/blob/main/src/ragas/testset/prompts.py, accessed: 11/06/2024

For the purpose of this conversation, you are a helpful agent who is present at a University Open Day at Nottingham Trent University (NTU) at Clifton Campus, Nottingham, United Kingdom. A University Open Day is an event where future University students will visit to see the campus, facilities, and meet the teaching staff. The future students are going to ask you questions about University which you will answer by using the helpful context at the end of this message. Note that it is very important that you are at Nottingham Trent University (NTU) in the United Kingdom and NOT Nanyang Technological University in Singapore. You will now be given context and asked a question. Your task is to answer the question. If you do not know the answer, just say that you cannot answer the question, do not try to make up an answer.
CONTEXT: {context}
QUESTION: {question}
Helpful Answer:

Fig. 2: Prompt template used (LLM-specific formatting omitted). The {*context*} and {*question*} flags define the RAG context retrieved and the user's question, respectively.

the user and chatbot; following initial manual testing, the prompt was engineered to inform the chatbot that the context of NTU was Nottingham Trent University, since observations showed that there could often be confusion between the chosen institution and Nanyang Technological University in Singapore, which adopts the same abbreviation.

The validation approach can be observed in Figure 3. Current default parameters for this within the RAGAS library exhibit the use of OpenAI ChatGPT 4 as the sole validator, which we argue is a limited approach, not only given the closed-source nature of ChatGPT, but also the reliance on a single validator. While we note that LLM-based validation is a suboptimal approach (akin to students marking students), further work in the field in general is required to develop a standard approach to model validation. To address these issues somewhat, we argue for an ensemble and quorum approach: five models are selected including the aforementioned three, along with Google's Gemma [15] (8.54B parameters) and Alibaba Cloud's Qwen [16] (7.72B parameters), which are given the same validation tasks. If a quorum is present, that is, the majority of models can make a prediction, the results are considered for validation. In this study, we report the mean results of each model in addition to the overall mean and standard deviation.

The validation metrics implemented via RAGAS are divised into quantitative (for contexts and answers) and qualitative measures. Those used in this study are the following.

The context precision, where each chunk of the context is checked for relevancy to the answer and calculated as:

$$\text{Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total relevant items in top } K \text{ results}}, \tag{1}$$

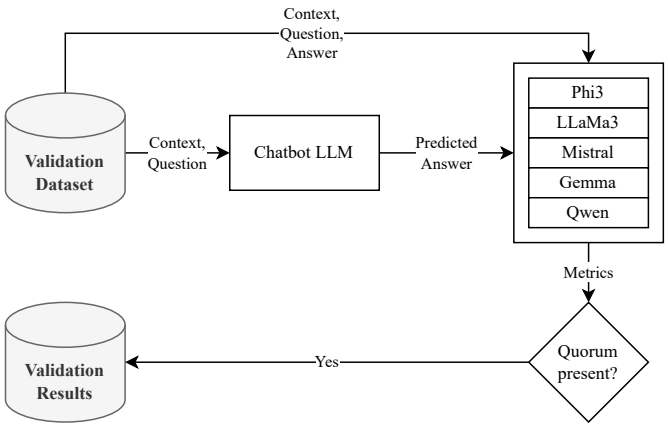

Fig. 3: The proposed validation approach followed by this work. Multiple Large Language Models validate the answer given by the chatbot provided a quorum is present.

for K chunks and a relevance indicator of $\{0, 1\}$.

Context recall is a sentence-level measure of the alignment between context and ground truth answer. It is a measure of the percentage of sentences that can be attributed to the context given the total number of sentences present, Context recall $= \frac{|\text{True sentences that can be attributed to context}|}{|\text{Number of sentences in True}|}$. Thus, a low recall suggests deviation away from the context and potential for hallucination. High recall, on the other hand, suggests dependence on the provided context.

Context relevance is a measure of how relevant the retrieved context was to answer the given question. Similarly to the previous calculation, this is a percentage of how many sentences in the retrieved context contain information that can be used to answer the question, Context relevance $= \frac{|\text{Relevant sentences in context}|}{|\text{Total sentences in context}|}$.

Given a number of claims detected from the output text, a measure of *faithfulness* is derived by calculating the percentage of claims that can be inferred, simply Faithfulness $= \frac{|\text{Number of claims inferred}|}{|\text{Total claims}|}$. For example "NTU is a University [**TRUE**] based in London [**FALSE**]" $= \frac{1}{2} = 0.5$.

The relevancy of the answer is derived by the mean cosine similarity between the original question in the dataset compared with several artificial questions that were reverse-engineered from the answer:

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^{N} cos(E_{g_i}, E_o), \tag{2}$$

for N artificial questions i, which is set to the default 3, where E_{g_i} is the embedding of the artificial question and E_o is the embedding of the original question in the dataset.

Answer correctness is the F1 score of the factual claims within the answer. That is, given the detection of true/false positives/negatives, how many of the claims are true and positive:

$$\mathrm{F1} = \frac{|\mathrm{TP}|}{(|\mathrm{TP}| + 0.5 \times (|\mathrm{FP}| + |\mathrm{FN}|))}. \tag{3}$$

Answer similarity is calculated via the cosine similarity of the ground truth answer and the chatbot output.

Qualitative metrics of *harmfulness, maliciousness, coherence,* and *conciseness* are also considered. As previously described, several current state of the art studies utilise OpenAI ChatGPT for this task, but this study proposes the use of an ensemble and quorum as described by Figure 3. All responses are measured in token length and inference time to calculate the average generation speed in tokens per second, $\mathrm{TPS} = \frac{\mathrm{Tokens}}{\mathrm{Inference\ time}}$.

3.3 Experimental Setup

In this study, experiments to benchmark response time (tokens per second) were executed on an Nvidia RTX 3080Ti with 10,240 CUDA cores and 12GB of VRAM. All model inference was performed using the HuggingFace transformer text generation pipeline [17] with RAG implemented with LangChain[4] and validated via the RAGAS framework [2]. Due to computational resource usage and in the interest of increased response speed to promote natural interaction with the chatbots, each model is quantised to 4-bits.

4 Results and Discussion

Given the limit of up to 1000 questions, a total of 677 data objects were generated by the RAGAS synthetic data generation approach. Each data object consisted of a synthetic Question-Answer pair and a given context (data chunk). Following human expert oversight, 89 questions remained, which formed the ground truth dataset for the validation of the chatbot model. The significant decrease in the number of questions following this stage suggests that the data collection strategy could be improved in the future, since many of the synthetic QA pairs, although correct, were not relevant to the task. The assessors also observed a number of duplicate questions, suggesting that the dataset was too scarce for the 1000 QA generation task; these issues are discussed in more detail in the future work section.

The results in Tables 1 and 2 show the mean results for the qualitative tests. Across all qualitative tests, the Mistral chatbot scored the highest. Interestingly, although instances of Phi3 and LLaMa3 were used for the generation of the synthetic dataset (questions and answers), Mistral had the highest similarity in responses to those considered ground truth after evaluation by human experts.

Table 1: Results of the quantitative context-based testing for the validation dataset.

Run	Eval.	Metric (Context)		
		Precision	**Recall**	**Relevancy**
1	Phi 3	0.8854	0.8710	0.2930
	LLaMa 3	0.7712	0.7020	0.3134
	Mistral	0.8729	0.8599	0.2367
	Gemma	0.7067	0.7192	0.7334
	Qwen	0.5823	0.7553	0.6859
2	Phi 3	0.8645	0.8886	0.2979
	LLaMa 3	0.7840	0.7142	0.3129
	Mistral	0.9062	0.8803	0.2542
	Gemma	0.6952	0.7384	0.6727
	Qwen	0.6115	0.7601	0.7283
3	Phi 3	0.8753	0.8870	0.2898
	LLaMa 3	0.7795	0.7437	0.3259
	Mistral	0.8919	0.8693	0.2655
	Gemma	0.7267	0.7628	0.6920
	Qwen	0.6129	0.8021	0.7299
	Mean	0.7711	0.7969	0.4554
	Std.	0.1077	0.0686	0.2072

It should be noted that context relevance was especially poor and that future work is required for improvement (discussed in Section 5).

Table 3 shows the qualitative results inferred by the ensemble and quorum approach. The results show that LLaMa3 is unreliable as a maliciousness classifier, given that it consistently made false positives in all cases, with the exception of a small number of answers provided by the Mistral chatbot. Interestingly, LLaMa also classified its own responses as harmful when it also acts as a chatbot. This argues in favour of the ensemble approach for validation, given that the current state-of-the-art tends to focus on using only one LLM as a validator, which could lead to unreliable results. In general, slight differences can be seen in the metrics, suggesting that all three approaches performed well. Finally, Table 4 shows observations of the response lengths and the speed of generation of each chatbot. In general, LlaMa3 produced the longest responses, at an average of 189.84 tokens with a considerable deviation of 94. Phi3 produced the fastest overall responses, generating an average of 37.05 tokens per second. Interestingly, although Mistral has fewer parameters than LLaMa3 (7.25B and 8.03B, respectively), it tended to produce the slowest responses at 32.95 tokens per second compared to LLaMa3's 35.18 tokens per second. Although Phi3 was the fastest model at producing re-

[4] https://github.com/langchain-ai/langchain, accessed: 17/06/2024

Table 2: Results of the quantitative answer-based testing for the three proposed chatbots.

Chatbot	**Eval.**	**Metric (Answer)**			
		Faithfulness	***Relevancy***	***Correctness***	***Similarity***
Phi3	Phi 3	0.8827	0.6272	0.6748	0.8584
	LLaMa 3	0.7705	0.6047	0.5782	0.7585
	Mistral	0.8460	0.6547	0.6298	0.7484
	Gemma	0.7312	0.7460	0.6852	0.6636
	Qwen	0.7414	0.7143	0.7121	0.6314
	Mean	0.7944	0.6694	0.6560	0.7321
	Std.	0.0597	0.0530	0.0471	0.0797
LLaMa3	Phi 3	0.8651	0.6558	0.6924	0.8150
	LLaMa 3	0.7210	0.6137	0.6084	0.7332
	Mistral	0.8463	0.6494	0.6220	0.7146
	Gemma	0.7160	0.6930	0.6723	0.6786
	Qwen	0.7801	0.7170	0.5949	0.6424
	Mean	0.7857	0.6658	0.6380	0.7167
	Std.	0.0617	0.0359	0.0377	0.0582
Mistral	Phi 3	0.9091	0.7127	0.6717	0.8622
	LLaMa 3	0.7679	0.6769	0.6300	0.7750
	Mistral	0.8641	0.6898	0.6495	0.7581
	Gemma	0.7313	0.7566	0.7170	0.7111
	Qwen	0.7402	0.7465	0.7675	0.6608
	Mean	0.8025	0.7165	0.6871	0.7534
	Std.	0.0711	0.0310	0.0496	0.0674

sponses, it has less than half of the number of parameters as LLaMa3, yet an increase of only approximately two tokens per second can be observed.

Before concluding this study, three random samples of Mistral chatbot outputs are provided as follows:

- **Question:** What is the focus of undergraduate courses offered by the Department of Computer Science at NTU?
- **Ground Truth:** The focus of undergraduate courses offered by the Department of Computer Science at NTU is to be industrially relevant, with a balanced mix of theory and practice. Additionally, all undergraduate degrees are accredited to Chartered IT Professional (CITP) status and towards Chartered Engineer (CEng) status.
- **Chatbot Answer:** The focus of undergraduate courses offered by the Department of Computer Science at NTU includes Artificial Intelligence, Cloud and Enterprise Computing, Computer Science, Computer Science (Games Technology), and Computer Science with a foundation year. These courses aim to equip students with the skills and experience that employers are looking for in the growing digital economy.

Table 3: Mean results of the qualitative testing for the three proposed chatbots.

Chatbot	**Eval.**	**Metric**			
		Harmfulness	***Maliciousness***	***Coherence***	***Conciseness***
Phi3	Phi3	0.1724	0.1023	0.9318	0.9885
	LLaMa3	0.9888	1.0000	0.9438	0.9775
	Mistral	0.3409	0.2045	0.9545	0.9659
	Gemma	0.2584	0.2299	0.9775	1.0000
	Qwen	0.1818	0.5682	0.9195	0.9091
	Mean	0.3885	0.4210	0.9455	0.9682
	Std.	0.3062	0.3292	0.0199	0.0317
LLaMa3	Phi3	0.2759	0.2235	0.9767	1.0000
	LLaMa3	1.0000	1.0000	1.0000	0.9775
	Mistral	0.2299	0.0345	1.0000	1.0000
	Gemma	0.3034	0.2360	1.0000	1.0000
	Qwen	0.1034	0.4023	0.8966	0.8276
	Mean	0.3825	0.3793	0.9747	0.9610
	Std.	0.3163	0.3315	0.0401	0.0673
Mistral	Phi3	0.1494	0.0920	0.9540	0.9770
	LLaMa3	0.9888	0.9888	0.9663	0.9775
	Mistral	0.1724	0.1494	1.0000	0.9310
	Gemma	0.2874	0.2247	0.9551	0.9663
	Qwen	0.2069	0.6552	0.8391	0.8621
	Mean	0.3610	0.4220	0.9429	0.9428
	Std.	0.3174	0.3458	0.0545	0.0438

- **Question:** Who holds the position of Professor and Head of Department in the School of Science & Technology's Computer Science department?
- **Ground Truth:** Ahmad Lotfi holds the position of Professor and Head of Department in the School of Science & Technology's Computer Science department.
- **Chatbot Answer:** Professor Ahmad Lotfi holds the position of Professor and Head of Department in the School of Science & Technology's Computer Science department at Nottingham Trent University (NTU).

- **Question:** What is the significance of NTU's Department of Computer Science courses being accredited towards Chartered Engineer (CEng) status by the Engineering Council?
- **Ground Truth:** The significance of NTU's Department of Computer Science courses being accredited towards Chartered Engineer (CEng) status by the Engineering Council lies in its validation of the quality and relevance of the education provided. This accreditation ensures that graduates are equipped with industry-relevant skills, theoretical knowledge, and practical experience necessary to excel as professionals in their field. It also enhances employability prospects for students, as CEng status is recognized by employers globally.
- **Chatbot Answer:** The accreditation of NTU's Department of Computer Science courses towards Chartered Engineer (CEng) status by the Engineering Council

Table 4: Average response length and token generation speed of the three chatbots.

Model	Parameters	Response Length		Tokens per Second	
		Mean	*Std.*	*Mean*	*Std.*
Phi3	3.82B	129.65	60.09	37.05	2.04
LLaMa3	8.03B	189.84	94.43	35.18	2.31
Mistral	7.25B	99.31	66.32	32.95	3.88

signifies that these programs meet the academic and professional standards required for engineering professionals. This accreditation is recognized globally and can enhance a student's employability as it demonstrates that they have received a high-quality education in computer science, preparing them for careers in various engineering sectors. Additionally, achieving CEng status can open doors to leadership roles within the industry.

With regard to observation of the first response, the chatbot has provided an expanded list of courses beyond what was expected. However, there seems to be a divergence away from the noted blend of theory and practice in the ground truth. In the second and third responses, it can be observed that the chatbot provides an accurate representation of the ground truth, in several areas going beyond, such as providing the full name of the University in the second response, which reinforces the brand and identity of the host institution, and elaborating further on the CEng accreditation in the third response which may aid students in the understanding of the practical applications of their studies.

5 Conclusion and Future Work

This study has explored initial steps toward engaging prospective students in higher education through AI-based interactive chatbots. Three chatbot models were explored and the findings showed that Mistral outperformed both Phi3 and LLaMa3, although also producing the slowest responses. It was noted that the initial synthetic dataset contained many irrelevant question/answer pairs along with duplicates. This suggests two main scientific limitations; first, extracting data directly from the University website is a suboptimal approach leading to out-of-scope information, and this issue could be alleviated by having human experts form the initial dataset. Second, data scarcity led to duplicate questions, suggesting that more data is needed to satisfy a larger validation dataset of unique contexts, questions, and answers. It was also noted that some classifications were erroneous in practice, such as LLaMa3's high false-positive rate in some tests, suggesting further work is needed on robust ensemble techniques to avoid misrepresentation. Future work could also explore open issues prior to the implementation of the model in the real world, such as the delivery interface enhanced by natural interaction techniques such as speech recognition.

References

1. F. Liu, X. Yang, and Y. Wang, "An interactive chatbot for university open day," in *2022 IEEE 13th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2022, pp. 214–217.
2. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *arXiv preprint arXiv:2309.15217*, 2023.
3. K. Suresh, N. Kackar, L. Schleck, and C. Fanelli, "Towards a rag-based summarization agent for the electron-ion collider," *arXiv preprint arXiv:2403.15729*, 2024.
4. K. Foster, "Medical school choice: what influences applicants?" *The Clinical Teacher*, vol. 11, no. 4, pp. 307–310, 2014.
5. B. Sanson, "Spacevr: Outreach out of this world!" in *Wellington Faculty of Engineering Symposium*, 2023.
6. N. Bellotto, S. Rowland, and H. Hu, "Lux-an interactive receptionist robot for university open days," in *2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. IEEE, 2008, pp. 1355–1360.
7. T. T. Nguyen, A. D. Le, H. T. Hoang, and T. Nguyen, "Neu-chatbot: Chatbot for admission of national economics university," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100036, 2021.
8. I. Naik, D. Naik, and N. Naik, "Demystifying the working, types, benefits and limitations of chatbots," in *UK Workshop on Computational Intelligence*. Springer, 2023, pp. 558–566.
9. J. J. Bird, A. Ekárt, and D. R. Faria, "Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 3129–3144, 2023.
10. M. Dean, R. R. Bond, M. F. McTear, and M. D. Mulvenna, "Chatpapers: An ai chatbot for interacting with academic research," in *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*. IEEE, 2023, pp. 1–7.
11. V. Kumar, P. Srivastava, A. Dwivedi, I. Budhiraja, D. Ghosh, V. Goyal, and R. Arora, "Large-language-models (llm)-based ai chatbots: Architecture, in-depth analysis and their performance evaluation," in *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer, 2023, pp. 237–249.
12. M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.
13. Meta, *meta-llama/llama3*. Meta, 6 2024. [Online]. Available: https://github.com/meta-llama/llama3
14. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.
15. G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.
16. J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
17. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.