Exploring Evaluation Methods for Generative AI-Powered Chatbot Output

Magdalen Beiting-Parrish[1,*], Jodi M. Casabianca[2,‡]

Abstract

There is considerable resistance to using Generative Artificial Intelligence (GAI) for educational applications because the inner workings of the models are not well-understood, nor are the outputs consistently accurate, which are threats to validity. This work-in-progress paper explores different popular GAI evaluation methods available and explores which are best for evaluating the output of GAI-based chatbots intended for educational use.

Keywords

Evaluation, Generative AI, chatbot, education

1. Introduction

The use of Generative Artificial Intelligence (GAI) for educational purposes has exploded in recent years. GAI applications have the power to change education and support truly bespoke learning and assessment opportunities for students of all ages and specific needs. GAI-based applications have been found to successfully score open-ended essay responses [1], to support the automated generation of test questions [2], to provide real-time feedback on writing [3] and even to support real-time learning analytics in the classroom [4]. There has also been increased interest in the use of GAI-powered chatbots for educational purposes, which is the focus of the current paper. These chatbots can provide a real-time tool for providing information [5] or as an adaptive tutor that can provide structured tutoring and feedback [6].

These are all incredible potential improvements to the education system; however, more needs to be understood about how these GAI systems work and a more specific set of best practices is needed for ensuring that the outputs of these systems are factually correct, consistent, and avoid bias(es) that may isolate specific group(s) of students. Additionally, since many of the users who typically interact with educational chatbots are K-12 children, it is even more important to ensure that the output is correct because children younger than age 12 frequently struggle with fantasy-reality distinction [7] and may accept incorrect output without thinking critically. Finally, undergraduate students use LLM-based tools to assist in their assignments; however, 16.5% use it as a "knowledge seeking" tool and frequently take the results at face value [8], so it is important that all GAI-powered educational tools be evaluated thoroughly to ensure that the output is of high quality.

In the process of evaluating AI systems and applications, we might perform a series of checks of the system to understand its common errors, and then perform more rigorous testing to capture and quantify these errors. We also wish to measure the system on popular AI constructs such as faithfulness or trustworthiness [9]. Using humans to review system inputs and outputs is the gold standard approach to bringing this type of analysis to scale. However, in quickly iterating environments with limited resources, many teams employ LLM-as-a-Judge techniques to use the same rubrics that human raters would use to evaluate the LLM-based application. Some AI evaluators create customized LLM judges that are aligned to human judgement via their ratings. That process involves using a golden dataset

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop

*Corresponding author.

magdalen.beiting@gmail.com (M. Beiting-Parrish); jodi@broadmetrics.co (J. M. Casabianca)

www.magdalen.beiting-parrish.com (M. Beiting-Parrish); https://broadmetrics.co (J. M. Casabianca)

0000-0002-3998-8672 (M. Beiting-Parrish); 0000-0002-0877-7063 (J. M. Casabianca)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-15

CEUR-WS.org/Vol-4330/paper4.pdf

and including some samples in-context (within the LLM judge prompt) and then testing and refining the prompt until it reaches a minimum alignment threshold. Alternatively, there are many open-source tools that offer generic automated metrics of the AI constructs. A gap in the literature is how these automated metrics perform [10].

This work-in-progress paper seeks to compare some of the commercially-available general GAI evaluation tools that are available on the market and will also explore the strengths and limitations of these for an educational setting that expects much higher thresholds of accuracy than other enterprise output. This paper represents ongoing research to explore and eventually codify a set of metrics for evaluating chatbot output to ensure that content used in educational spaces is accurate and represents the source material.

2. Aims

There are several aims of this research paper. The first is to explore two popularly available methods for evaluating GAI chatbot output to see how well they work compared with each other. The next aim is to see which of these popular method(s)works best for use with educational applications that require high levels of accuracy in output. The third aim is to see which of the popular methods is best aligned to the larger goals of educational measurement, such as culturally and linguistically responsive assessment [11], if any.

3. Sample and Chatbot Development

The sample for use in the following study is from an educational chatbot pilot study designed to study an LLM-powered chatbot to support the development of ethical educational LLM-based models and frameworks in a non-judgmental and friendly conversational tone. The original pilot study detailed the development and iteration of a GAI-powered chatbot for ethical model development for educational applications [12] that used OpenAI's 4o-mini model as its foundation. In addition, OpenAI's 4.1 nano model was used to facilitate input sanitization and ensure that no prompt injection or other input issues occurred.

To develop the chatbot, a database containing over 90 journal articles, book chapters, white papers, and frameworks related to ethical LLM model development was created. A complete bibliographic list of these sources is located on the Github associated with this paper [13]. These were vectorized and stored for use in a Retrieval-Augmented Generation (RAG) framework as part of the chatbot design process. In this approach, the internal chatbot prompt is structured such that the chatbot cannot answer without drawing on a reference from the vector database of papers and resources. This is one attempt at mitigating hallucination and facilitating more factual answers, because the chatbot must rely on sources to answer and cannot simply create an unfounded response. Factuality of response is especially important because this chatbot was designed to support researchers and policy makers in educational measurement to think through potential ethical dilemmas and pitfalls in designing LLM-based applications for educational purposes. Consequently, the chatbot output has to be high quality and extremely accurate to avoid leading researchers and policy makers into accidental ethical problems that may impact students and their families, educators, and districts.

To create a golden dataset of question and answer pairs, we generated A set of 30 questions and human-generated answers with the relevant context (that the answer should have come from). These were created alongside the LLM-generated output to the same questions.

To compare human-generated metrics to the evaluation tools, we assigned ratings based on the formulas from the Ragas metrics [14] for Contextual Recall and Faithfulness and on the DeepEval Answer Relevancy metric [15]. Since Contextual Precision involves ranking within the vector database, this cannot be done reasonably by a human rater without being able to understand the "black box" of the LLM's scoring algorithm.

4. Methods

This paper will explore the functionality, usability, and applicable information of popular output evaluation tools that are freely available from the larger field of GAI evaluation to examine the applicability of these tools to educational output. These tools all facilitate an LLM-as-a-Judge evaluation approach, using GPT 4.1 as the judge. The first of these is DeepEval [15], an open-source Python-based tool that contains a number of different metrics including Hallucination, Bias, Toxicity, and additional agentic and conversational metrics. Ragas [14], is another Python-based open-source suite of evaluation metrics, containing RAG-related metrics such as Context Precision, Context Recall, Noise Sensitivity, Response Relevancy, and Faithfulness.

4.1. Metric Details

As mentioned above, DeepEval and Ragas systems contain a number of different metrics for evaluating the relationships between the text of the question, the text of the LLM-generated response and the original content that is supposed to be used in the LLM-generated response. All of these metrics explore how well the LLM is drawing on the vector database in the service of providing correct answers to the user. Each metric is defined differently across both systems (see Appendix A for the system-specific operationalizations of the formulas for each metric); however, all of these taken together can be used to fully understand the current quality of the LLM chatbot output, based on the evaluation questions. We selected four metrics, which are specifically intended to be used with applications using RAG in that there is an intended reference document as well as a question that includes both a human answer, the ethics chatbot-derived answer, and a base GPT-4 answer for comparison purposes to the ethics chatbot.

Firstly, the Contextual Recall score measures the amount of documents and/or their subsequent relevant information retrieved within the output answer. This value is on a scale of 0 to 1; higher scores mean that more relevant documents were included. Contextual Recall must have an intended reference or gold standard document to be calculated.

The Faithfulness score measures how factual the given LLM output is in relationship to the retrieved contextual document(s) from the vector database. This value is also measured on a scale of 0 to 1 with a higher value indicating that the output reliably retrieves and incorporates the information from the vector database; essentially that the statements in the output are supported by the original library of information.

Contextual Precision measures the LLM's ability to more highly rank relevant chunks of information to the user's query than chunks of information that are irrelevant to the user's question; this is on a scale of 0 to 1 with a score closer to 1 indicating that only relevant chunks of information were included. DeepEval also provides a rationale to support the value it awards.

Finally, Answer Relevancy is a metric that quantifies the extent to which the chatbot's answers were relevant to the questions asked by the user. This is calculated differently between Ragas and Deepeval. In Ragas, the system generates a set of artificial questions that are related to the original chatbot-provided system response, and uses the mean of the cosine similarity between the embeddings of these and the original user question to estimate answer relevancy. By contrast, DeepEval extracts all statements made in the LLM output provided by the chatbot, classifies which of these are relevant to the original input, and divides this by the total number of statements in the output.

5. Preliminary Results

This is a work in progress and the current results compare the same metrics across both DeepEval and Ragas.

Firstly, looking at Contextual Recall, which is the measurement of how many relevant documents/pieces of information were successfully retrieved from the context to craft the output, Ragas and DeepEval demonstrated very different mean scores (0.73, SD = 0.29 vs. 0.66, SD= 0.39); however, looking at the distributions of these scores (Figure 1), Ragas demonstrates a wider distribution, whereas

DeepEval largely has a bimodal distribution in which it gives values clustered around 0 or 1. This was in contrast to the human evaluation, which had a mean score of 0.84 (SD = 0.22). Overall, using human scoring as the baseline, the LLMs tended to mark more information as irrelevant than the human rater did; in this case, the human rating would still be the most accurate metric for system performance.

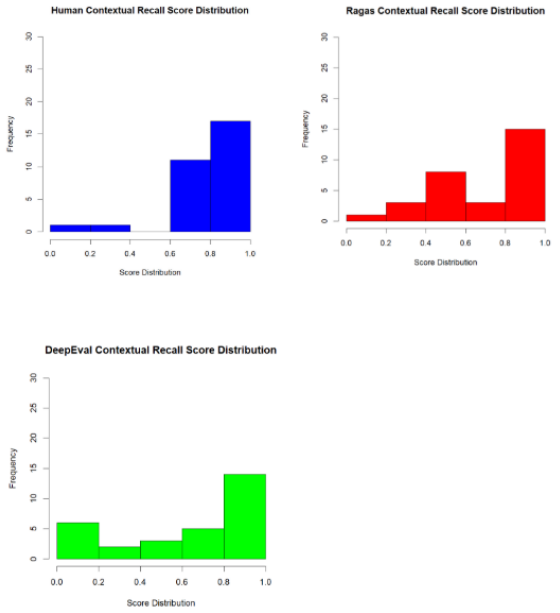


Figure 1: Contextual Recall Distributions for Human, Ragas, and DeepEval

Next, in looking at Faithfulness, which is the metric of how consistent the output is with the context, Ragas had a mean score of 0.86 (SD= 0.24) and DeepEval gave 0.99 (SD= 0.047), meaning that DeepEval rated every output as almost perfectly faithful to the original context. Again in looking at the distributions (Figure 2) , DeepEval gave a 1.00 most of the time, whereas Ragas has more variance within the distribution of assigned metrics. The human score distribution also had a very high mean value of 0.95 (SD = 0.19).

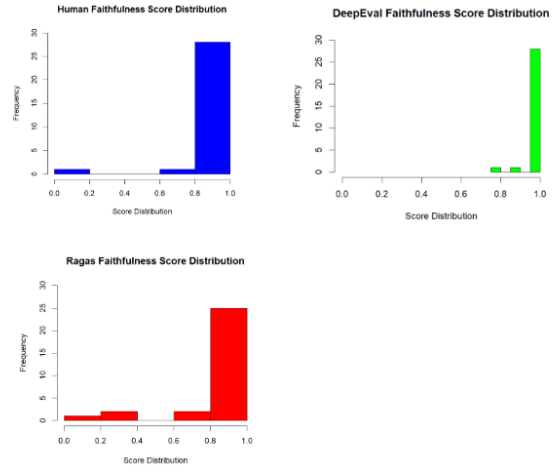


Figure 2: Faithfulness Distributions for Human, Ragas, and DeepEval

Additionally, in looking at Contextual Precision, which is a score that measures the model's ability to retrieve higher ranked relevant chunks of information over irrelevant documents, Ragas had a mean score of 0.66 (SD=0.28) and DeepEval had a mean score of 0.62 (SD=0.31). Again, looking at the

distribution of scores (Figure 3), both systems had similar distributions, so more evaluation questions are needed to see which of these would be better to use in a more standardized fashion.

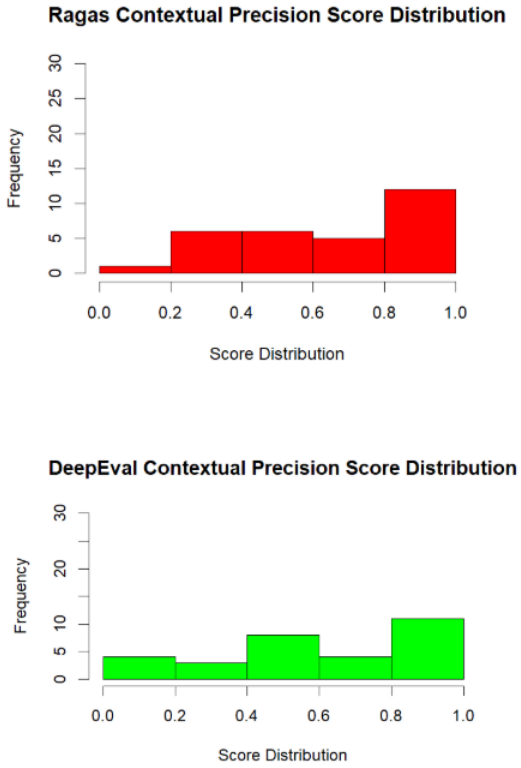


Figure 3: Contextual Precision Distributions for Ragas, and DeepEval (no human rater scores for this metric)

Finally, in looking at Answer Relevancy, which is the measurement of how relevant the output is in relation to the user's question, Ragas had a mean score of 0.70 (SD=0.29) versus DeepEval's mean score of 0.98 (SD=0.062). Much like Faithfulness, Ragas has a distribution that tends to skew towards higher values, whereas DeepEval tends toward awarding perfect scores to every question-answer pair. There is some subjective "judgment" needed by the LLM to decide on Answer Relevancy, and once again, DeepEval tends to over-value Relevancy when it should perhaps be more critical. DeepEval was more comparable to the human score which had a mean score of 0.91 (SD=0.19); however, a larger evaluation set is needed to further explore this relationship. All available values for these metrics are available in Figures 1-4 as well as in Appendix 2.

6. Current and Anticipated Conclusions

In summary, Ragas awarded more distributed scores across each of the metrics, whereas, DeepEval was far more lenient as a grader and awarded a score of 1 most of the time. These experiments were repeated several times for the same metrics and there was a fair amount of variability in scores across metrics for both tools; the best results were presented here. In general, these should be run several times and the average metrics should be derived across runs to ensure that the distributions are consistent across runs and that there is not some anomaly of the text or output. In comparing these metrics with the human ratings, these tended to be more similar to DeepEval; however, based on the RAG-based nature of this system, the human rater also had document(s) in mind that they expected to be referenced based on the questions being asked of the system.

In general, there are a few issues with these metrics. Firstly, these are calculated slightly differently between the two systems; this is important to note and also to be aware when deciding on an evaluation

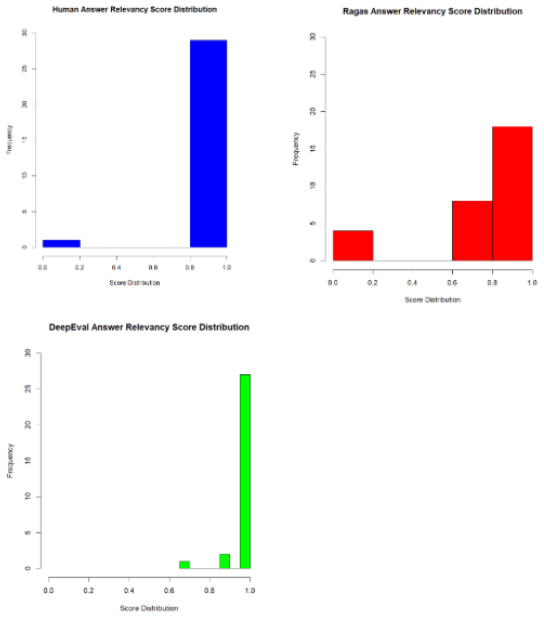


Figure 4: Answer Relevancy Distributions for Human, Ragas, and DeepEval

system to use. Secondly, some of the metrics provide incomplete information. For example, Faithfulness examines the number of claims in the output and measures whether the included claims represent the original source context. In this case, Faithfulness is gameable because a perfect score would say that all claims made in the output are from the original source and there are no additional claims that are irrelevant; however, the output might not include all of the needed information from the original source.

Additionally, Contextual Recall had the biggest discrepancy between the human ratings and the automated metrics generated by Ragas and DeepEval. The human rater was able to look at the question and chatbot answer pair and compare these thoughtfully with the intended context to award this score; DeepEval and Ragas do not account for situations in which the chatbot output draws on multiple documents and thus tends to rely on one of the source documents to largely award the score. This could likely be improved with internal prompting or bespoke rubrics.

For next steps, the final presentation will include output from some of the deeper metrics found in the Ragas and DeepEval toolsets and explore the pros and cons of the developer-created rubrics and evaluation tools. Further, additional questions will be added to the evaluation question set to examine whether there are any content areas or types of questions that are especially difficult for these evaluation tools to measure or any that work especially well with these tools. We will also build an LLM judge aligned to human ratings to compare to these systems. In the spirit of collecting validity evidence for the AI ethics chatbot (Casabianca, 2025), we will also explore secondary datasets that are related to educational tutoring or training to see if there are any differences in the educational ethics output and that of output related to tutoring.

6.1. Contributions

Currently, the field of educational measurement does not have a set of best practices for how to evaluate the output of GAI-powered chatbots; however, there is a desperate need for more scholarship in this area to support high quality GAI-powered tools that can be used safely for students and their families. While there are some evaluation tools in the larger LLM development ecosystem that have been widely used, the field of educational measurement does not have a set of codified guidance on which tools to use and which metrics are acceptable for ensuring high quality chatbot output. This paper helps by exploring the use of commercially available evaluation tools, thereby contributing some of the work

toward ensuring valid and reliable educational chatbot output. This paper will help the field to better understand the strengths of existing tools as well as to better understand where additional tools and approaches are needed and where to support the field in where to apply energy next.

6.2. Limitations and Next Steps

One of the biggest limitations of the present study is the relatively small sample size of evaluation questions to explore the quality of chatbot output. There are likely areas of the original ethics vector database that need to be expanded; indeed, a benchmark dataset that has a comprehensive of psychometrically-vetted items that can thoroughly evaluate the chatbot output across all domains of ethical model development is a next step of this research project. Psychometricians are uniquely positioned to design robust benchmarks to ensure that LLM-based systems perform accurately and with high quality. Additionally, a larger sample of users and their questions will also be sought for subsequent research to ensure robust user testing that more reliably represents the population of users for this chatbot system. Finally, neither system emerged as more culturally sensitive or appropriate for educational content, so a benchmark that measures these aspects is also important for further study.

7. Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus in order to perform: grammar and spelling check and citation management. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Whitmer, Author, Results of NAEP math item automated scoring data challenge & comparison between reading & math challenges, 2023. URL: https://doi.org/10.35542/osf.io/uwymf. doi:10.35542/osf.io/uwymf.

[2] R. Circi, J. Hicks, E. Sikali, Automatic item generation: foundations and machine learning-based approaches for assessments, Frontiers in Education 8 (2023) 858273. doi:10.3389/feduc.2023.858273.

[3] S. Lottridge, A. Burkhardt, C. Ormerod, Woolf, Write On with Cambi, Technical Report, Cambium Assessment, 2025. URL: https://files.portal.cambiumast.com/corporate-site/documents/CAI-CSSC-NCME-Write-on-with-Cambi-White-Paper.pdf.

[4] K. Holstein, B. M. McLaren, V. Aleven, Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity, Journal of Learning Analytics 6 (2019) 27–52. doi:10.18608/jla.2019.62.3.

[5] T. Zhang, L. Patterson, M. Beiting-Parrish, B. Webb, et al., Ask NAEP: A generative AI assistant for querying assessment information, Journal of Measurement and Evaluation in Education and Psychology 15 (2024) 378–394. doi:10.21031/epod.1548128.

[6] W. A. Sahlman, A. M. Ciechanover, E. Grandjean, Khanmigo: Revolutionizing learning with GenAI, 2023. Harvard Business School Case 824-059, Revised April 2024.

[7] J. D. Woolley, M. E. Ghossainy, Revisiting the fantasy-reality distinction: children as naïve skeptics, Child Development 84 (2013) 1496–1510. doi:10.1111/cdev.12081.

[8] A. Stojanov, Q. Liu, J. H. L. Koh, University students' self-reported reliance on ChatGPT for learning: A latent profile analysis, Computers and Education: Artificial Intelligence 6 (2024) 100243.

[9] J. M. Casabianca, Psychometrics is all you need, EdArXiv preprint, 2025. URL: https://doi.org/10.35542/osf.io/7w6pz_v1. doi:10.35542/osf.io/7w6pz_v1.

[10] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. Parameswaran, I. Arawjo, Who validates the validators? aligning LLM-assisted evaluation of LLM outputs with human preferences, in:

Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 2024, pp. 1–14.

[11] G. Solano-Flores, Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation, Frontiers in Education 4 (2019) 43. doi:10.3389/feduc.2019.00043.

[12] M. Beiting-Parrish, Introducing a generative AI chatbot that supports ethical model development, 2025. Paper presentation at the annual meeting of the National Council on Measurement in Education, Denver, CO, United States.

[13] M. Beiting-Parrish, Ethics_bot, 2025. URL: https://github.com/mathgie/ethics_bot.

[14] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, 2023. URL: https://arxiv.org/abs/2309.15217, arXiv preprint arXiv:2309.15217.

[15] Confident AI, DeepEval, 2025. URL: https://github.com/confident-ai/deepeval?tab=readme-ov-file.

A. Formulas for Metrics

A.1. Contextual Recall Metrics

$$\text{Context Recall} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$$

Ragas

$$\text{Contextual Recall} = \frac{\text{Number of Attributable Statements}}{\text{Total Number of Statements}}$$

DeepEval

Figure A.1: Contextual Recall Metric Formulas for Ragas and DeepEval

A.2. Faithfulness Metrics

$$\text{Faithfulness Score} = \frac{\text{Number of claims in the response supported by the retrieved context}}{\text{Total number of claims in the response}}$$

Ragas

$$\text{Faithfulness} = \frac{\text{Number of Truthful Claims}}{\text{Total Number of Claims}}$$

Deepeval

Figure A.2: Faithfulness Metric Formulas for Ragas and DeepEval

A.3. Contextual Precision Metrics

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

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Ragas

$$\text{Contextual Precision} = \frac{1}{\text{Number of Relevant Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Number of Relevant Nodes Up to Position } k}{k} \times r_k \right)$$

DeepEval

Figure A.3: Contextual Precision Metric Formulas for Ragas and DeepEval

A.4. Answer Relevancy Metrics

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

$$\text{Answer Relevancy} = \frac{1}{N}\sum_{i=1}^{N}\frac{E_{g_i}\cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

Where:
- E_{g_i}: Embedding of the i^{th} generated question.
- E_o: Embedding of the user input.
- N: Number of generated questions (default is 3, configurable via `strictness` parameter).

Ragas

$$\text{Answer Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

DeepEval

Figure A.4: Answer Relevancy Metric Formulas for Ragas and DeepEval

B. Metric Frequency Statistics

	Human Score M (SD)	Ragas M (SD)	DeepEval M (SD)
Answer Relevancy	0.91 (0.19)	0.70 (0.29)	0.98 (0.062)
Faithfulness	0.95 (0.19)	0.86 (0.24)	0.99 (0.047)
Contextual Precision	NA	0.66 (0.28)	0.62 (0.31)
Contextual Recall	0.84 (0.22)	0.73 (0.29)	0.66 (0.39)

Figure A.5: Metric Frequency Statistics