ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(Special issue); 378-394

Ask NAEP: A Generative AI Assistant for Querying Assessment Information

Ting ZHANG* Luke PATTERSON** Maggie BEITING-PARRISH*** Blue WEBB****
Bhashithe ABEYSINGHE***** Paul BAILEY****** Emmanuel SIKALI*******

Abstract

Ask NAEP, a chatbot built with the Retrieval-Augmented Generation (RAG) technique, aims to provide accurate and comprehensive responses to queries about publicly available information of the National Assessment of Educational Progress (NAEP). This study presents an evaluation of this chatbot's performance in generating high-quality responses. We conducted a series of experiments to explore the impact of incorporating a retrieval component into GPT-3.5 and GPT-4o large language models and evaluated the combined retrieval and generative processes. This work presents a multidimensional evaluation framework using an ordinal scale to assess three dimensions of chatbot performance: correctness, completeness, and communication. Human evaluators assessed the quality of responses across various NAEP subjects. The findings revealed that GPT-4o consistently outperformed GPT-3.5, with statistically significant improvements across all dimensions. Incorporating retrieval into the pipeline further enhanced performance. The RAG approach resulted in high-quality responses. Ask NAEP reduced the occurrence of hallucinations by increasing the correctness measure from 85.5% of questions to 92.7%, a 50% reduction in non-passing responses. The study demonstrates that leveraging large language models (LLMs) like GPT-4o, along with a robust RAG technique, significantly improves the quality of responses generated by the Ask NAEP chatbot. These enhancements can help users to better navigate the extensive NAEP documentation more effectively by providing accurate responses to their queries.
Keywords: *Generative AI, chatbot, NAEP*

Introduction

The purpose of this paper is to introduce an information retrieval chatbot powered by generative artificial intelligence (GAI). This chatbot aims to enhance public access to the National Assessment of Educational Progress (NAEP) publicly available online sources and facilitate knowledge sharing for the National Center for Education Statistics (NCES). The chatbot, Ask NAEP, answers user queries based on publicly accessible information from the NCES website with relevant web links (see Figure 1). By incorporating cutting-edge Gen AI techniques and ensuring a rigorous evaluation, the chatbot strives to deliver timely, accurate, and comprehensive responses.

This paper begins by describing the context and development of the chatbot, including its design philosophy, framework, and the challenges the project faced along with our corresponding solutions. The subsequent sections cover the evaluation methodology and results. The report concludes with a discussion of the findings and outlines future directions for continuing to develop the chatbot.

Context

NAEP is the longest-standing federally funded U.S. assessment. As an assessment arm of NCES, NAEP's mission is to inform policymakers, educators, researchers, and the public about what the nation's students know and can do in various subjects through comprehensive reports and on-demand

* Senior Researcher, Dr., American Institutes for Research, USA, tzhang@air.org, ORCID ID: 0009-0001-1724-6141
** Data Scientist, American Institutes for Research, USA, lpatterson@air.org, ORCID ID: 0009-0000-2612-0375
***Impact Fellow, Dr., Federation of American Scientists, USA, mbeitingparrish@fas.org, ORCID ID: 0000-0002-3998-8672
**** Researcher, American Institutes for Research,USA, bwebb@air.org, ORCID ID: 0009-0004-4080-9864
***** Researcher, Dr., American Institutes for Research, USA, babeysinghe@air.org, ORCID ID: 0009-0006-4107-8615
****** Principal Economist, Dr., American Institutes for Research, USA, pbailey@air.org, ORCID ID: 0000-0003-0989-8729
******* Acting Branch Chief: Reporting & Dissemination, Assessment Division, Dr., National Center for Education Statistics, USA, emmanuel.sikali@ed.gov, ORCID ID: 0009-0007-5325-0475

To cite this article:.Zhang,T., Patterson,L., Beiting-Parrish, M., Webb, B., Abeysinghe,B., Bailey,P., Sikali,E., (2024). Ask NAEP: A Generative AI Assistant for Querying Assessment Information, *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 378-394. https://doi.org/10.21031/epod.1548128

Received: 10.10.2024
Accepted: 2.12.2024

access to results. NAEP is committed to be transparent about the psychometric, sampling design, instrument design, and other scientific methodologies it uses to produce its assessments, surveys, and estimation procedures. To fulfill the mission, NCES documents the information on two main websites: the main NAEP website under NCES (National Center for Education Statistics, 2024a) and the Nation's Report Card (National Center for Education Statistics, 2024b.

These NAEP websites provide a wealth of publicly available information, including well-documented assessment frameworks, survey and assessment methodologies, data on participating teachers and schools, student questionnaires, and results from decades of assessments. However, locating information on NCES websites can be particularly challenging for NAEP users due to the vast quantity of documents developed over time by different vendors, with older releases rarely removed. Web pages may contain overlapping information (e.g., sampling designs) and inconsistent details (e.g., the number of plausible values in NAEP). Answers to questions often need to be retrieved from multiple documents or resources and verified for their currency. Some example queries include: What content is in the 2018 NAEP Technology and Engineering Literacy assessment? Can I opt my child out of participating in the NAEP assessment? And What is stratification in NAEP sample design? (see more examples in Table 1).

Development of the Generative AI Chatbot

Large language models (LLMs), such as the GPT(Brown et al., 2020), Llama (Touvron et al., 2023), and Gemini (Anil et al., 2024) models, have demonstrated powerful capacities in language understanding and generation. Most can generate responses to users' queries with patterns of speech that closely resemble those of humans (Gao et al., 2023). However, these models are trained on large datasets that may not be curated exclusively for reliability, and their output is not specifically evaluated for accuracy (Abeysinghe & Circi, 2024). Additionally, some models have limitations in providing up-to-date and content-specific information. Although trained on vast amounts of data, they may still miss specific or niche information, and their knowledge is fixed at the time of training and confined to what they encountered during that training (Gao et al., 2023).

Through this work, we sought to develop an information retrieval chatbot, Ask NAEP, to provide responses to users' queries on NAEP information. We do not claim to have perfect accuracy in all responses, as it would be a claim that is unprovable and inflated. However, in this article, we describe how we worked to increase the quality of responses based on three dimensions: correctness, completeness, and communication.

The RAG Framework and Technology

Retrieval-augmented generation (RAG) is a mechanism that combines the strengths of information retrieval and generative models to produce more accurate and contextually relevant responses. The RAG architecture was introduced to address some of the limitations of purely generative models by incorporating an external knowledge retrieval step before generating a response (Gao et al., 2023).

We used a RAG pipeline that retrieves relevant information from a customized knowledge base. This knowledge base aggregates data from the NAEP application programming interfaces (APIs) and content-related text from web pages under the Nation's Report Card (NRC) subsection of the NCES website as well as under the NRC website. The process is shown in Figure 1 and described in this section, with reference to the steps shown in Figure 1.

Figure 1
Architecture of Ask NAEP.

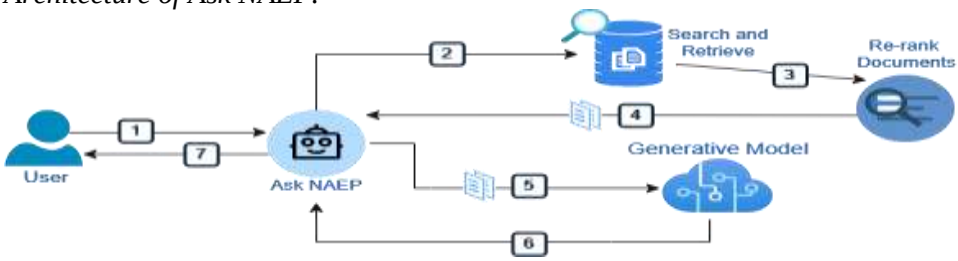


Figure 1 illustrates the workflow of the Ask NAEP Architecture. Upon receiving a user query, relevant documents are retrieved from a vector database and subsequently reranked. The query and documents are then sent to the agent, and an LLM is used to generate the final response, which is then returned to the user.

The information in the knowledge base is projected into numeric vector embeddings using OpenAI's *text-embedding-ada-002* model. When users submit a query, it is converted into a vector embedding using the same model, and documents with the closest vectors to the query (i.e., the most similar vector embeddings) are retrieved (Figure 1, steps 1 and 2). For persistent data storage, we use ChromaDB, which we chose because it is open source, self-hostable, lightweight, and easily integrated into Python applications. It also offers customization options for the parameters used in its search algorithm, Hierarchical Navigable Small Worlds (HNSW) (Malkov & Yashunin, 2018). We measure distance using cosine similarity. Once the top documents have been retrieved using this metric, they are reranked using a cross-encoder model. Cross-encoder models output relevance scores for document query pairs, which are learned through supervised training. Our framework currently uses the *ms-marco-MiniLM-L-6-v2* cross-encoder model from HuggingFace (HuggingFace, 2024).

The query, along with a prompt and the reranked documents, is sent to an OpenAI LLM to generate responses (Figure 1, steps 3, 4, and 5). Our RAG framework offers developers the flexibility to choose from various LLMs, including different versions of GPT models. The overall application was developed in Python, with the front end currently deployed in a preproduction environment using a Flask application.

Finally, the user is shown both the chatbot's response and the top documents associated with the response (Figure 1, step 6 and 7).

Knowledge Base

The NCES NAEP websites are a compendium of assessments and results, information for parents, students, researchers, media, school administrators, teachers, and resources for researchers and educators. It also includes a variety of data tools, state and district profiles, etc. To illustrate the complexity of this website, we unpack a small section of the resources for researchers, specifically the NAEP Technical Documentation Website (TDW, National Center for Education Statistics, 2024c). The TDW is the technical description of all the operations that NAEP has used to conduct and assess students since 2000. Prior to this, technical documentation reports were printed. Altogether, there are about 37,000 static and interactive pages on the NRC. The static pages are on the main NAEP website, while all interactive pages are on the NRC.

We conducted an extensive crawl of the NCES websites using the open-source Scrapy(GitHub, 2024a) and Selenium(GitHub, 2024b) modules in Python for crawling, collecting the raw HTML from about 5,000 pages. The purpose of these scrapes was to collect unprocessed HTML to retain in persistent storage, allowing us to experiment with different approaches to processing and splitting the page contents. From the raw HTML, we extracted items such as paragraph text, alt text for figures, and table titles and contents. We separated the page contents into paragraphs, sections, and titles before creating embeddings and adding documents to our vector database. Sections were identified by programmatically splitting the full-page contents at section headers, which were detected by their use of HTML markup

language (e.g., bold text). Within each section, contents were further divided into paragraphs based on the presence of newline characters. As detailed in a subsequent section, the current knowledge base includes documents derived from the paragraph text on these pages, with tables stored as associated metadata in JSON format.

We also sent requests to the data API that powers the state and district profile tools. Each response provided a summary of performance for the specified state or district.

A challenge arose when augmenting our data with content scraped from the Nation's Report Card website, where much of the information is presented in interactive figures or tables that require human interaction to navigate and extract specific results (e.g., the gap between English language learners (ELLs) and non-ELLs in the 2022 grade 8 reading assessment). Unlike static web pages, dynamically rendered content is difficult to scrape programmatically. To address this issue at the state and district levels, we reconstructed API calls to the respective profiles and collected summary texts for each state and district. This data is stored separately and used to answer questions about specific states or districts. Because the NCES website rarely removes pages for older releases, an ongoing challenge is ensuring that the retrieved web page links are both relevant and up to date. In some cases, pages pertain to specific years, grades, and subjects, which we identify through keyword detection in the user's query and apply as a filter. If no such keywords appear in the query, no prefiltering is applied to the knowledge base, and all pages are considered in the similarity search.

Evaluation Approach

Testing Queries

We evaluated Ask NAEP using a bank of expert-generated questions. We selected 55 questions that experts thought most representative of common and important questions that individuals might seek answers to on the NAEP website. These questions were categorized into the topics shown in Table 1 for further analysis:

Table 1
NAEP Questions

Topic	Example Question	Number of Questions
NAEP Content Areas and Assessments	What content is in the 2018 NAEP Technology and Engineering Literacy assessment?	12
NAEP Data Analysis and Statistical Techniques	How are NAEP plausible values used to conduct secondary analysis?	11
NAEP Scores and Achievement Levels	What are the achievement levels for NAEP in general? What was the average 4th-grade math score for NAEP in 2022?	8
NAEP Participation and Accommodations	Can I opt my child out of participating in the NAEP assessment?	7
NAEP Scoring and Assessment Process	When do constructed-response items need to be rescored?	7
NAEP Sample Design and Methodology	What is stratification in NAEP sample design?	5
NAEP Validity and Reliability	How are items treated if the fit is not good in NAEP?	5

Two human raters from the research team evaluated the responses from various versions of the Ask NAEP chatbot using the CCC framework rubric. Interrater reliability was calculated using Cohen's Kappa (Cohen, 1960).

Evaluation Framework and Metrics

Ideally, interacting with a chatbot should feel like a natural conversation, where the chatbot's written responses are as comprehensible as a text message produced by a human author. With this in mind, we created an initial framework based on Grice's Maxims of Conversation (1989), which views conversation as a collaborative product between two parties who share a common aim. In this case, the aim is to gain a better understanding of some aspect of NAEP, whether it involves procedural information or specific test results.

Within this conversational exchange, there are four maxims that ensure a quality response: quantity, quality, relation, and manner (Grice, 1989). These are especially relevant to the presentation of statistical chatbot responses, which should ideally be long enough to include all necessary information without being burdensome (quantity), be truthful (quality), include only relevant information (relation), and be as concise and clear as possible (manner). Since several of these criteria are specific to individual users, for our purpose, we simplified the system to include three criteria—Correctness, Completeness, and Communication—as outlined in Table 2 below.

Table 2

Framework for Generative Component Evaluation

Construct	Annotation	Description
Correctness	$Q_{correct}$	Is the content of the chatbot's answer factually correct?
Completeness	$Q_{complete}$	Does the chatbot's answer include the relevant facts and information needed to answer the question?
Communication	Q_{comm}	Is the chatbot's answer written in a clear and concise fashion?

The following weights are applied to generate a composite score from the three constructs. Since the primary goal of this chatbot is to deliver accurate, complete, and reliable responses to user queries, we prioritize correctness and completeness over communication by assigning greater weight to the first two dimensions. It is worth noting that these weightings are not based on prior studies or established theories.

$$Q_{composite} = \frac{2}{5}Q_{correct} + \frac{2}{5}Q_{complete} + \frac{1}{5}Q_{comm}$$

Table 3 below describes how these dimensions were graded by human reviewers. Some evaluation analyses are based on "pass/fail" grading. The rubric was constructed so that grades of 3, 4, and 5 represent qualitatively acceptable answers for a published chatbot, while grades of 1 and 2 do not. This is why the threshold for a passing answer is 3 or higher for all dimensions.

Table 3
Dimension Scoring Rubric

Grade	Pass/Fail	Correctness	Completeness	Communication
1 (Poor)	Fail	Significant factual errors or misinformation	Incomplete, missing crucial information	Unclear, convoluted, or difficult to follow
2 (Below Average)	Fail	Some inaccuracies or lack of precision	Lacks relevant details or fails to address all aspects	Lacks coherence and may confuse the reader
3 (Average)	Pass	Several minor inaccuracies, but generally correct	Covers the basics but could benefit from more depth	Clear but could be more concise
4 (Good)	Pass	Generally accurate with 1-2 minor inaccuracies	Sufficiently complete, addressing the main points	Generally clear and concise
5 (Excellent)	Pass	Completely correct with no errors	Comprehensive with thorough information	Succinct, well organized, and easy to understand

A major concern in the present chatbot evaluation process is that, since this chatbot represents the interests of a federal statistical agency, it is imperative that it does not hallucinate—that is, it should not produce any answers that are partially or completely incorrect. These three criteria were applied in various forms to all of the human evaluation work conducted.

Research Questions

One could argue that the only component that needs to be evaluated properly is the generative component and how effective the generated responses are. In a RAG bot, however, the retrieval is an important intermediary that can help diagnose why a chatbot responds correctly or incorrectly to queries. If receiving the correct retrieval is unimportant, RAG is not providing a significant improvement over unaltered ChatGPT, so testing the retrieval is one of the evaluation's research questions.

Our evaluation of Ask NAEP centers around four research questions (RQ):

RQ1. How satisfied are users with Ask NAEP?
RQ2. Which LLM performed better in the RAG chatbot?
RQ3. How important is good retrieval at generating a quality answer?
RQ4. Does the Ask NAEP retrieval process and bot configuration produce quality answers?

Method

To answer RQ1, we conducted the user testing when Ask NAEP was using GPT-3.5 as its generative model. We consider user testing to be an important component to ensure that the chatbot effectively meets real-world user needs and satisfaction. This method allows for iterative improvements that align the chatbot's performance with actual user behavior and preferences.

Participants included NAEP users from various states across the United States who used Ask NAEP and recorded any unsatisfactory interactions; the focus of this round of human evaluation was negative experiences with the chatbot. Among these users, seven interacted with the chatbot and filled out 13 forms, representing a total of 58 problematic interactions with the chatbot out of a much larger pool of interactions. Users also provided feedback on why the output they received was problematic and answered multiple-choice questions regarding why they flagged the output, whether it was easy to understand (correctness), whether it contained relevant information (completeness), and how the output was communicated (communication). The feedback from this user testing was used to improve the performance of Ask NAEP. The current paper presents the results from this round of user testing.

To answer the second RQ, the research team evaluated the Open AI generative models (e.g., GPT-3.5 and GPT-4o) within the RAG framework to identify the best-performing model. Due to our institutes' security and efficiency concerns, only OpenAI's GPT models were tested for powering the generative answers that Ask NAEP produces. Development of the chatbot began when GPT-3.5 was the latest

OpenAI LLM available. However, GPT-4 and GPT-4o have since been released. As part of our evaluation, we assessed whether GPT-4o performed better than GPT-3.5 as the underlying chatbot model. We did this by generating answers to all 55 test questions using GPT-3.5 with no retrieval augmentation, then repeating this process for GPT-4o. We then performed a round of human evaluation of each answer across all dimensions.

In RQ 3, we assessed how well the bot answered questions given proper context, as well as its effectiveness in retrieving relevant content to support its answers. Finally, we combined the two components to address the last RQ. Does the chatbot produce quality answers?

Given the relative novelty of these applications, evaluation methods for a RAG chatbot are still emerging, and the research community has not yet reached a consensus on the most effective evaluation approaches (Abeysinghe & Circi, 2024; Celikyilmaz et al., 2021; Howcroft & Rieser, 2021; Iskender et al., 2021; Schoch et al., 2020; Smith et al., 2022; van der Lee et al., 2019). Consequently, it is evident that a variety of evaluation methods exist, some of which are domain-specific while others are more general in nature (Abd-Alrazaq et al., 2020; Chan et al., 2023; Fu et al., 2023; Sellam et al., 2020; Zhang et al., 2020). To evaluate chatbots, researchers use both human evaluation and automated methods. For this research, we adopted the human evaluation approach.

Results

How Satisfied Are Users with Ask NAEP, the GPT-3.5 Version?

We conducted the user testing when GPT-3.5 was the latest OpenAI LLM available. After users interacted with Ask NAEP and provided feedback between May and June 2024, the data were analyzed to better understand the typical issues with that version of the chatbot needing improvement. These responses were further coded using the Correctness-Completeness-Communication (CCC) framework. One of the authors also investigated why these chatbot responses were flagged as problematic.

In general, out of the 58 problematic responses flagged by the experts, 10 (17.2%) occurred because this version of the chatbot declined to answer. For example, when asked, “How many were below basic?” the chatbot provided responses such as: “I'm sorry, but I cannot provide an answer to your question as it cannot be answered using information about NAEP. Is there anything else I can assist you with?” In this case, the chatbot may have declined to answer because the question wasn’t specific to a particular subject area, grade level, or year.

Additionally, 9 responses (15.5%) were flagged because they included some form of hallucination, meaning that part or all of the answers consisted of completely incorrect information. For example, when asked, “Can I find out how my student did on NAEP?” the chatbot responded, “Yes, you can find out how your student did on NAEP. The NAEP Data Explorer allows users to search for state results by student demographic groups and other variables....” The first sentence is considered a hallucination because NAEP never provides scores down to the individual student level.

At first glance, these values seem very high; however, the experts were instructed to report any problematic chatbot responses, so these values are likely higher than they would be for the entire set of chatbot interactions.

In addition, we asked the experts overall why they flagged each question. The two most common reasons were that the chatbot only partially answered the question (38.5%) or did not answer it at all (16.9%). Despite these issues, most respondents found the chatbot easy to understand (51.6%). When asked whether all responses included relevant information, 36.1% of the respondents agreed that they did. Finally, the experts found that the chatbot communicated in a logical manner with a beginning, middle, and end 71.4% of the time. This feedback suggests that while the chatbot’s information may need refining, its communication style is generally accessible.

Finally, the chatbot output was also scored by one of the authors using the CCC rubric. The results are presented in Table 4, which includes both averages and medians. However, it is important to note that human evaluations often treat rubric scores as continuous values, which may not always be appropriate,

as they are ordinal categories (Howcroft & Reiser, 2021). Given this distinction, the scores from this analysis are much lower than those from the larger set of sample questions; however, they remain consistent with the types of response values that were flagged.

Table 4

Average and Median Correctness-Completeness-Communication (CCC) Scores for Flagged Chatbot Output

Question	Average Correctness	Median Correctness	Average Completeness	Median Completeness	Average Communication	Median Communication
1	2.33	3	2.46	3	2.85	3
2	3.25	3	3.17	3	3.33	3.5
3	2.82	3	2.91	3	3.27	3
4	2.18	3	2.18	3	3.00	4
5	2.09	2	2.36	3	3.09	3

Which LLM Performed Better in the RAG chatbot?

We compared outcomes from the Ask NAEP GPT-4o without retrieval augmentation to those from the GPT-3.5 version, also without retrieval augmentation. Results are presented below in Table 5.

Table 5

Percentage of Passing Answers for Ask NAEP Without Retrieval by LLMs and Dimension

Dimension	N	GPT-4o, No Retrieval[1]	GPT-3.5, No Retrieval[2]	Percentage Point Difference	Permutation test p-value
Correctness	55	87.2%	70.9%	16.4%**	0.00
Completeness	55	89.1%	74.5%	14.5%**	0.00
Communication	55	98.2%	84.5%	13.6%**	0.01
Overall	55	87.2%	71.8%	15.5%**	0.01

Significant at the **5% confidence level

[1]To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .64.

[2]To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .61.

Table 5 shows the percentage of answers produced by GPT-4o and GPT-3.5 that were rated as acceptable by human evaluators for each dimension. The table reveals that GPT-4o outperformed GPT-3.5 in all three dimensions, as well as in overall performance, with differences statistically significant at various confidence levels. The largest difference was observed in the Communication dimension, where GPT-4o achieved 98.2% acceptable answers, compared to 76.4% for GPT-3.5. This difference was significant at the 1% confidence level. The smallest difference was observed in the Correctness dimension, where GPT-4o achieved 85.5% acceptable answers, compared to 70.9% for GPT-3.5. This difference was still significant at the 10% confidence level. These results suggest that GPT-4o is a better model than GPT-3.5 for this chatbot, so Ask NAEP currently uses GPT-4o.

How Important is Good Retrieval at Generating a Quality Answer?

To begin addressing this question, we first examine whether the Ask NAEP retrieval process performs any better than no retrieval at all. We do this by comparing the performance of Ask NAEP with a version of Ask NAEP that performs no content retrieval (which is simply stock GPT-4o with a system context prompt explaining that it is a helpful assistant that answers questions about NAEP). As a reminder, the dimension scores shown in this section are the GPT-assessed scores, and a passing score is a 3, 4, or 5 for the dimension. Table 6 shows that the Ask NAEP retrieval process leads to improvements in passing answer percentages on the Completeness and Correctness dimensions, as well as in overall performance (though the differences are not statistically significant), with no change in the Communication dimension.

Table 6

Percentage of Passing Answers for Ask NAEP (GPT-4o version) With and Without Retrieval by Dimension

Dimension	N	With Retrieval[1]	No Retrieval[2]	Percentage Point Difference	Permutation test p-value
Correctness	55	92.7%	87.3%	5.4%	0.27
Completeness	55	93.6%	89.1%	4.5%	0.38
Communication	55	97.2%	98.2%	-0.9%	0.99
Overall	55	92.7%	87.2%	5.4%	0.27

Significant at the **5% confidence level.

[1]To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .65.

[2]To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .64.

Table 6 shows that Ask NAEP provides acceptable answers to more questions than GPT-4o with no retrieval, but we also want to know whether it provides higher quality answers. To do this, we perform a 5-level ordinal logit regression of a binary Ask NAEP response indicator on each of the four dimensions. For each regression, the dimension score is treated as an ordinal dependent variable Y with 5 ordered categories j. The ordered logistic regression model can be expressed as:

$$logit(P(Y \geq j)) = a_j - \beta X$$

where:

- $logit(P(Y \geq j))$ is the log-odds of the dependent variable Y being greater than or equal to category j.
- a_j are the threshold parameters (cut points) for each category j.
- β is the vector of regression coefficients.
- X is the vector of independent variables. In this case, the only independent variable included is a binary indicator $X_{AskNAEP}$, which equals 1 when the answer was generated by Ask NAEP and 0 when it was generated by the no-retrieval model.

What we are interested in is the value of $\beta_{AskNAEP}$, whose value is the log-odds that the response generated by Ask NAEP is in a higher quality category compared to the response generated by the no-retrieval model. If $\beta_{AskNAEP} > 0$, then answers from Ask NAEP are more likely to be in higher or equal quality categories than those from the no-retrieval model. If $\beta_{AskNAEP} < 0$, then answers from Ask NAEP are more likely to be in lower quality categories. If $\beta_{AskNAEP} = 0$, then there is no difference in quality between the answers from Ask NAEP and the no-retrieval model. Table 7 shows the results of the ordinal regression.

Table 7

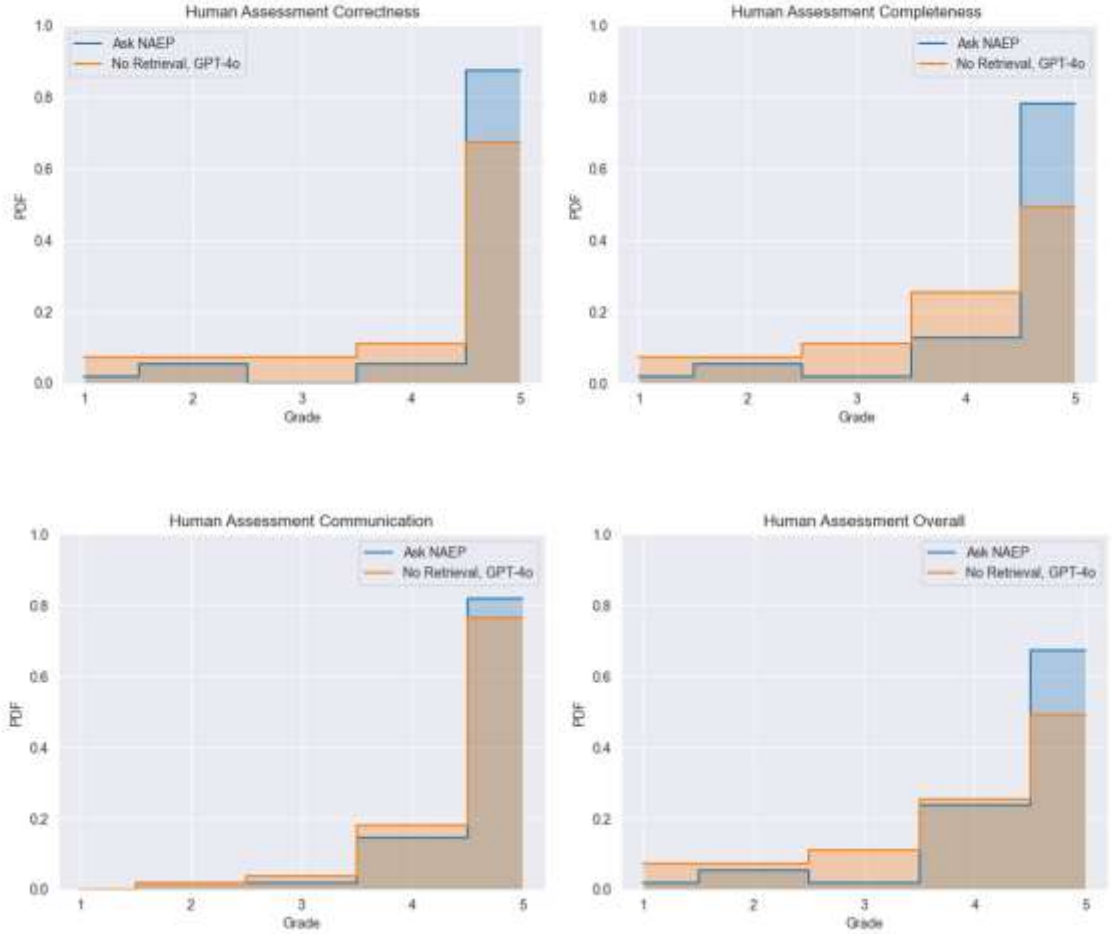
Ordinal Regression Estimated Probability of Higher or Equal Rating Using Ask NAEP Retrieval Process

Dimension	*N*	Log Odds	Odds Ratio	*p* value |Log Odds| > 0
Correctness	55	1.14	3.13	0.00**
Completeness	55	1.37	3.92	0.00**
Communication	55	0.02	1.02	0.96
Overall	55	0.71	2.04	0.03**

Significant at the **5% confidence level.

The results show that the retrieval process significantly enhances performance in the Completeness and Correctness dimensions, as well as in overall quality. Specifically, the odds of achieving a higher Completeness rating are 3.13 times higher with the Ask NAEP retrieval process, compared to the process with no retrieval. The odds of a higher Correctness rating and a higher Overall rating are 3.92 times and 2.04 times higher, respectively, with the retrieval process. All of these odds are statistically significant. However, the retrieval process does not have a significant impact on the Communication dimension.
These results suggest that the retrieval process is effective in improving the correctness and completeness aspects of response quality, but not necessarily the communication aspect. The cumulative density functions of the assessments for Ask NAEP and GPT-4o with no retrieval are shown in Figure 2.

Figure 2
Probability Density Functions of Dimension Ratings



In Table 8, we examine how the retrieval process impacts the overall score by topic. The NAEP Scores and Achievement Levels topic showed the most improvement. However, the statistical power of this comparison is limited due to the low sample size of questions in each category, so this analysis should be considered exploratory.

Table 8

Overall Percentage of Passing Answers for Ask NAEP (GPT-4o version) With and Without Retrieval by Topic

Topic	*N*	With Retrieval	No Retrieval	Difference
NAEP Sample Design and Methodology	5	100%	100%	0.00
NAEP Data Analysis and Statistical Techniques	11	100%	91%	0.09
NAEP Scoring and Assessment Process	7	100%	86%	0.14
NAEP Scores and Achievement Levels	8	100%	75%	0.25
NAEP Participation and Accommodations	7	93%	100%	-0.07
NAEP Content Areas and Assessments	12	88%	75%	0.13
NAEP Validity and Reliability	5	80%	80%	0.00

Significant at the **5% confidence level.

Does the Ask NAEP Retrieval Process and Bot Configuration Produce Quality Answers?

Ask NAEP attempted to answer all the test questions, and human reviewers gave generally high reviews to these answers across all dimensions. For all dimensions, over 94% of answers received passing grades from human reviewers. Table 9 presents these results, and Figure 3 provides a histogram showing the frequency of each grade for every dimension. Note that *N* is 110 for Table 9 and Figure 3 in this section, as two human reviewers rated bot responses separately for each of the 55 questions, producing 110 reviews in total.

Table 9

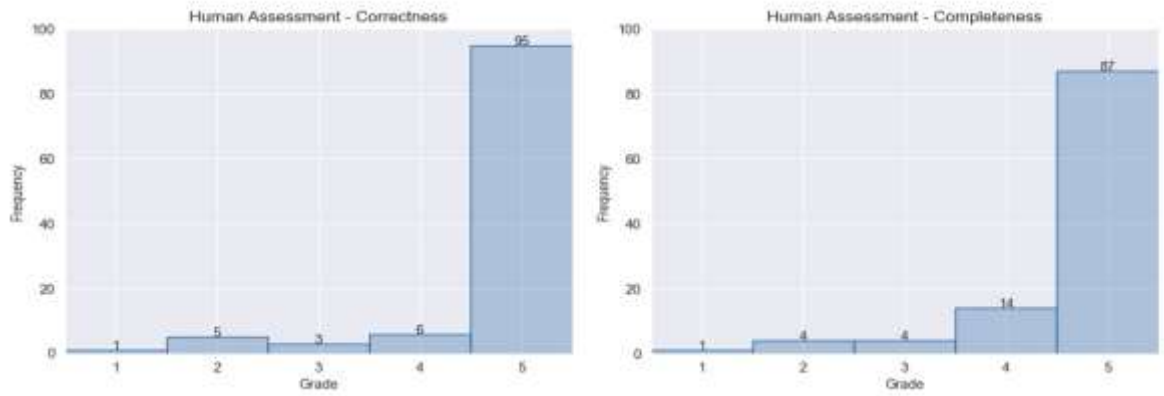
Percentage of Passing Answers for Ask NAEP by Dimension, According to Human Evaluation

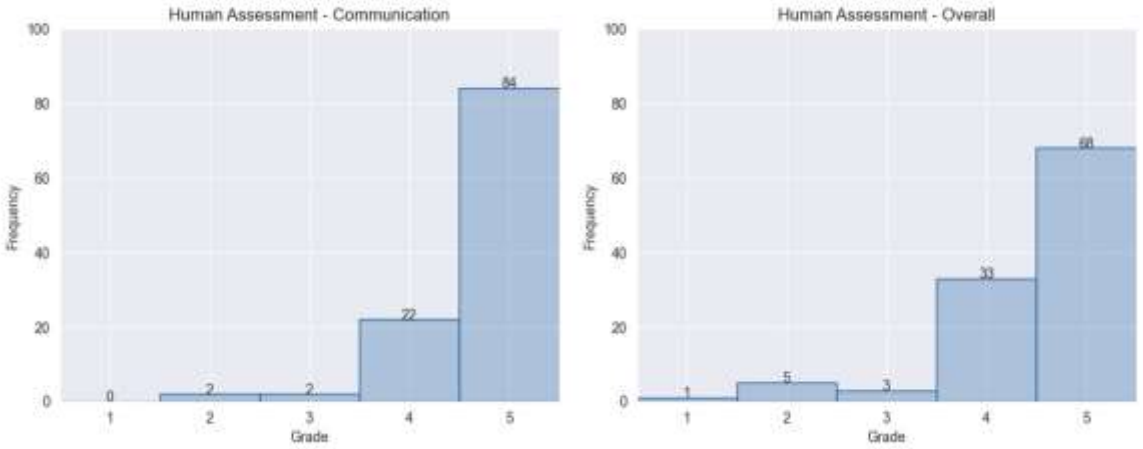
Dimension	*N*[1]	Percentage of Passing Answers
Correctness	110	94.5%
Completeness	110	95.5%
Communication	110	98.2%
Overall	110	94.5%

[1] The sample size is 110 because two human reviews are available for each of the 55 questions.

Figure 3

Histograms of Human Assessment of Answered Questions





To better understand which types of queries Ask NAEP answers well, Table 10 shows the percentage of questions with a passing rating for each dimension by topic. Overall, Ask NAEP at this stage is best at answering questions on NAEP Data Analysis and Statistical Techniques, Sample Design and Methodology, Scores and Achievement Levels, and the Scoring and Assessment Process. However, the results indicate a need for improvement in addressing questions related to NAEP Validity and Reliability. This insight has guided the team on which additional documents and data should be integrated in the next phase.

Table 10

Percentage of Passing Answers, Human Assessment of Answered Questions by Topic

Topic	*N*	Correctness	Completeness	Communication	Overall
NAEP Data Analysis and Statistical Techniques	22	100%	100%	100%	100%
NAEP Sample Design and Methodology	10	100%	100%	100%	100%
NAEP Scores and Achievement Levels	16	100%	100%	100%	100%
NAEP Scoring and Assessment Process	14	100%	100%	100%	100%
NAEP Participation and Accommodations	14	93%	93%	100%	93%
NAEP Content Areas and Assessments	24	88%	92%	100%	88%
NAEP Validity and Reliability	10	80%	80%	80%	80%

Discussion and Conclusion

Significance

Ask NAEP demonstrates potential in assisting users to locate the information they need and providing accurate, complete, and comprehensive responses on various NAEP topics. This is particularly true for queries that are summation-based rather than investigative (e.g., questions like "Why did group A perform better than group B?"). In RAG, our corpus is sourced from a federal statistical agency's website, which undergoes an extensive quality control process. This ensures that the information retrieved is accurate and approved. However, some user questions, particularly 'why questions,' may

not align with the agency's mission and therefore lack supporting text. Since NAEP information is published by NCES, a statistical agency known for presenting only facts without causal explanations, our chatbot cannot answer investigative questions, especially 'why questions'.

Evaluating language models in generative applications is a challenging task. In this work, we present our evaluation framework, which is an ordinal scale evaluation across three dimensions chosen to assess the quality of Ask NAEP in the context of a federal statistical organization. While other studies have explored dimensional evaluation (e.g., Abeysinghe & Circi, 2024; Fu et al., 2023; Gehrmann et al., 2023; van der Lee et al., 2019, 2021), they are generally more applicable to broader contexts rather than a statistical agency. Therefore, the selection of the proper dimensions for this task is a unique application of our evaluation and is our contribution.

An additional aspect of our contribution is addressing the complexity of existing human evaluation tools, which are often multidimensional and difficult to work with. Previous research has shown that general-purpose rubrics with five or more categories can be challenging for evaluators to use effectively (Wolf et al., 2008). In contrast, the current CCC approach is a much simpler tool for evaluating chatbot output. In developing the CCC approach from our earlier multidimensional framework based on Grice's maxims, we found it easier to apply to chatbot output and more time-efficient compared to the full framework.

Finally, our results indicate that by combining a well-developed RAG mechanism with a more advanced LLM (in this case, GPT-4o), Ask NAEP reduced the occurrence of hallucinations. This improvement is reflected in an increase in our correctness measure from 85.5% to 92.7%. The nonpassing response percent is 100% minus the percent correct and is reduced from 14.5% to 7.3%, this 7.2 percentage point increase in correctness is probably better viewed as a 50% reduction in nonpassing responses.

Principal Findings

In this section, we further analyze and interpret the results that were presented earlier. We also discuss results categorized into the research questions and present the findings accordingly.

The first research question is about user testing. Despite using an older version of Ask NAEP (the bot with GPT-3.5), the results are consistent with the above findings. The GPT 3.5 version performs well on questions based on NAEP's procedures, methodologies, and definitions, including understanding the NAEP assessment process, statistical methods, type of data collected, and assessment purposes. For example, it can accurately answer questions about the NAEP assessment process, how plausible values are drawn, and how biases are addressed in NAEP research studies. It also effectively handles questions about the subjects assessed by NAEP and how NAEP benefits schools and communities.

Conversely, the GPT 3.5 version struggles with questions requiring specific knowledge or data about NAEP assessments, such as the number of items in specific assessments, average scores for specific years, or content from specific years. It also has difficulty with questions about accommodations for disabilities or options for opting out of the assessment. The information and feedback obtained from this round of user testing have been used to enhance Ask NAEP, resulting in improvements to the current version.

Lastly, further investigation of user feedback allowed us to explore additional issues with the chatbot. This analysis revealed that hallucination and refusal to answer remain ongoing issues. Both are generative issues, which may be difficult to resolve without fine-tuning the LLM.

The second research question investigated what LLM should be used in the Ask NCES chatbot context. While acknowledging the existence of other language models, such as Claude and Llama, we limited our experiments to the GPT family for this initial proof of concept. In this work, we present the choice between two large language models, GPT-3.5 and GPT-4o, excluding other elements of the chatbot, such as the embedding process and prompts.

The goal of the second research question was to determine which LLM generates higher quality responses, as judged by human evaluators. For this purpose, a human assessment carried out on 55 questions across different NAEP subjects and administration years revealed that GPT-4o generates much more favorable responses. Our evaluation found that human evaluators rated GPT-4o responses higher than GPT-3.5 responses across all dimensions, and the difference was statistically significant for all dimensions. This finding suggests that despite the increased expenses associated with GPT-4o, its use in critical situations is justified by its superior performance.

In the third research question, we investigated whether adding the retrieval component to GPT-4o would improve the performance on the CCC measures. Our findings show a significant improvement with the addition of retrieval. We are continuing to explore other avenues through which we may enhance retrieval, including alternative embedding models, frameworks for semantic chunking of text, and alternative vector stores that natively support hybrid search.

In addressing the fourth research question, we examined whether combining retrieval with generative processes would result in higher quality bot responses. To test this, we conducted a human evaluation with the CCC measures across two sets of questions: one general and one specific to various NAEP subjects and administrations. Both experiments showed that Ask NAEP attempted to answer the majority of the questions. The human evaluations showed a high passing rate for responses across all dimensions, with a passing score defined as 3 or above on the ordinal scale.

Additionally, we examined whether Ask NAEP generates higher quality answers on specific topics. At this stage, Ask NAEP performs best on questions related to NAEP Data Analysis and Statistical Techniques, Sample Design and Methodology, Scores and Achievement Levels, and the Scoring and Assessment Process. However, the results indicate a need for improvement in addressing questions related to NAEP Validity and Reliability, a finding that aligns with the ongoing efforts to integrate NAEP-published data. This insight has guided the team on which additional documents and data should be integrated in future phases.

Challenges and Limitations

Implementing the Ask NAEP chatbot revealed to us some of the challenges and limitations associated with this type of application. Developing the chatbot involved scraping and storing a large amount of web articles and PDF documents. Dynamic websites, which require human interaction to reveal certain content, proved particularly difficult to scrape. This prompted us to look for other resources for the same information, such as using APIs for state and district profiles to collect summary texts. Another challenge was managing and storing a large amount of unstructured text information, for which vector stores are currently the state-of-the art solution.

Sometimes, a user may ask about a specific NAEP assessment year. Through experimentation, we found that intercepting the user's query and parsing it to identify the requested year provides better responses. However, we are still working on the ongoing challenge of ensuring that the most up-to-date content is retrieved when the user does not specify a particular year.

Opportunities and Future Directions

Ask NAEP is a proof-of-concept tool that we developed for NCES, with the intention of expanding it to include a larger corpus, such as NCES's entire website, to meet the broader demands of NCES data users.

Our ongoing efforts involve integrating NAEP-published data (e.g., NAEP summary data tables) and PDFs (such as white papers and methodology reports). However, in this evaluation round, we focused solely on the knowledge base derived from HTML content and state and district data APIs, which we acknowledge as a limitation of this chatbot version.

Future directions include conducting user testing with a more diverse group of stakeholders. For example, although our "Communication" criteria have largely been reviewed by researchers with advanced degrees, most of the responses might still be too technical to be understood by the general public, based on their readability scores. This process would help us better understand the kinds of questions these user groups might ask and give us time to ensure that the responses to the most common questions are consistently accurate.

Another avenue we would like to explore is evaluating other LLMs to see how they perform on the same tasks. As mentioned above, we limited the selection of LLMs to GPT-3.5 and GPT-4o for the convenience of conducting human evaluations. However, there are other models trained on different datasets and using different training procedures. Without testing these models on our knowledge base, it would be difficult to compare their performance with Ask NAEP. Therefore, we plan to conduct similar experiments with other LLMs, such as Claude (Anthropic) and PPLX (Perplexity).

Declarations

Gen-AI Use: The authors of this article declare (Declaration Form #: 2611241800) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Conflict of Interest: None of the authors have any competing interests that would be interpreted as influencing the research.

Funding: This project has been funded at least in part with Federal funds from the U.S. Department of Education under contract numbers 91990022C0053 and 91990023D0006/91990023F0350. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Acknowledgments: We would like to express our gratitude to Joseph Wilson for his review of the manuscript and constructive feedback. We also thank Jillian Harrison and Martin Hahn for their assistance with editing and formatting.

References

Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. Journal of Medical Internet Research, 22(6), e18301. https://doi.org/10.2196/18301

Abeysinghe, B., & Circi, R. (2024, June 13). The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches. The First Workshop on Large Language Models for Evaluation in Information Retrieval, Washington D.C. https://doi.org/10.48550/arXiv.2406.03339

Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., … Vinyals, O. (2024). Gemini: A Family of Highly Capable Multimodal Models (arXiv:2312.11805). arXiv. https://doi.org/10.48550/arXiv.2312.11805

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165 [Cs]. http://arxiv.org/abs/2005.14165

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of Text Generation: A Survey (arXiv:2006.14799). arXiv. http://arxiv.org/abs/2006.14799

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate (arXiv:2308.07201). arXiv. http://arxiv.org/abs/2308.07201

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models (arXiv:2309.11495). arXiv. https://doi.org/10.48550/arXiv.2309.11495

Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2023). GPTScore: Evaluate as You Desire (arXiv:2302.04166). arXiv. http://arxiv.org/abs/2302.04166

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gehrmann, S., Clark, E., & Sellam, T. (2023). Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. Journal of Artificial Intelligence Research, 77, 103–166. https://doi.org/10.1613/jair.1.13715

GitHub. (2024a). Scrapy. GitHub. https://github.com/scrapy/scrapy

GitHub. (2024b). Selenium. GitHub. https://github.com/SeleniumHQ/selenium

Grice, P. (1989). In the way of words. London: Harward University Press.

HuggingFace. (2024). cross-encoder/ms-marco-MiniLM-L-6-v2. HuggingFace. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

Howcroft, D. M., & Rieser, V. (2021). What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 8932–8939). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.703

Iskender, N., Polzehl, T., & Möller, S. (2021). Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead. In A. Belz, S. Agarwal, Y. Graham, E. Reiter, & A. Shimorina (Eds.), Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval) (pp. 86–96). Association for Computational Linguistics. https://aclanthology.org/2021.humeval-1.10

Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs (arXiv:1603.09320). arXiv. https://doi.org/10.48550/arXiv.1603.09320

National Center for Education Statistics. (2024a). NAEP. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/

National Center for Education Statistics. (2024b). The Nation's Report Card. U.S. Department of Education. https://www.nationsreportcard.gov/

National Center for Education Statistics. (2024c). Technical documentation. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/tdw/

Schoch, S., Yang, D., & Ji, Y. (2020). "This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation. In S. Agarwal, O. Dušek, S. Gehrmann, D. Gkatzia, I. Konstas, E. Van Miltenburg, & S. Santhanam (Eds.), Proceedings of the 1st Workshop on Evaluating NLG Evaluation (pp. 10–16). Association for Computational Linguistics. https://aclanthology.org/2020.evalnlgeval-1.2

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning Robust Metrics for Text Generation (arXiv:2004.04696). arXiv. https://doi.org/10.48550/arXiv.2004.04696

Smith, E. M., Hsu, O., Qian, R., Roller, S., Boureau, Y.-L., & Weston, J. (2022). Human Evaluation of Conversations is an Open Problem: Comparing the sensitivity of various methods for evaluating dialogue agents (arXiv:2201.04723). arXiv. http://arxiv.org/abs/2201.04723

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. Computer Speech & Language, 67, 101151. https://doi.org/10.1016/j.csl.2020.101151

van der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. Proceedings of the 12th International Conference on Natural Language Generation, 355–368.

Wolf, K., Connelly, M., & Komara, A. (2008). A Tale of Two Rubrics: Improving Teaching and Learning Across the Content Areas through Assessment. 8(1).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT (arXiv:1904.09675). arXiv. https://doi.org/10.48550/arXiv.1904.09675