Evaluation of Models and RAG Improvement using LLM As A Judge

Gregory Labbe
Case Studies in Machine Learning
Master of Science in AI
The University of Texas at Austin
labbegc@my.utexas.edu

This paper explores using large language models (LLMs) as evaluators to assess the performance of other LLMs. We demonstrate the effectiveness of this approach in discriminating between models with varying architectures and sizes, using both binary classification and long-form question-answering tasks. Our results show that OpenAI's o1-preview model outperforms the Llama models in both tasks, while the smaller Llama models exhibit comparable performance in binary classification. Notably, the application of Retrieval Augmented Generation to the smallest Llama model significantly enhances its performance in long-form answers, making it competitive with larger models. Our study highlights the value of using LLMs as evaluators to gain insight into model strengths and weaknesses. It demonstrates the potential of retrieval mechanisms to improve the performance of smaller models. These findings have important implications for developing and fine-tuning LLMs and underscore the importance of using LLMs as evaluators to inform model improvement.

Introduction

The rapid advancement of artificial intelligence (AI) has led to the development of sophisticated language models capable of performing complex natural language processing (NLP) tasks. Evaluation systems play a crucial role in assessing the performance and capabilities of these models. Traditional evaluation methods, however, cannot properly measure the subtleties of similarities in human-readable questions and answers. Large language models (LLMs) have recently been explored as subjects of evaluation and evaluators themselves, offering scalable and human-like assessment capabilities (Liang et al., 2022; Meta, 2024a; OpenAI, 2023b).

This paper investigates the use of LLMs as judges in evaluating the foundation model's performance. It highlights the evolution of evaluation systems, the challenges faced by traditional methods, and the potential of LLM-based evaluation to provide more comprehensive and human-like results. We explore how these techniques differ from older systems and discuss their implications for the future of AI evaluation (Srivastava et al., 2023).

Why older evaluation systems struggle

Traditional evaluation systems often rely on high-level metrics such as accuracy, precision, and recall, which require a binary truth for calculations. While these metrics provide quantitative measures, they fail to capture the deeper understanding required for more advanced model output, which often mimics human language.

Older measurement systems struggle with the following:

- **Semantics vs. Syntax**: Focusing on correctness rather than means the evaluations do not reflect understanding (Liang et al., 2022).
- **Reasoning Abilities**: Unable to evaluate model outputs' logical reasoning and inference meaning (Mondorf & Plank, 2024).
- **Language Nuances**: Inability to appreciate cultural references and context-dependent meanings (van Dijk et al., 2023).
- **Relevance and Coherence**: It is difficult to evaluate whether the generated text relates to the question and makes sense (Mackie et al., 2023).
- **Hallucinations**: Is the model simply providing an answer even if that answer is not correct (Khatun & Brown, 2024).

These limitations show a need for improved evaluation techniques that can understand complex questions and answers. This is quite often the job of a human.

How are these techniques different?

LLM as a Judge is a new development representing a shift in evaluation techniques. Unlike traditional metrics, LLM as a Judge leverages contextual understanding and reasoning of more recent advanced language models. Key differences include:

- **Open-ended Evaluation**: allows for more detailed evaluations beyond fixed metrics.
- **Unified Framework**: LLMs can serve as generic evaluators across tasks and produce homogenous output.
- **Contextual Understanding**: Evolving capabilities to comprehend context and semantics leading to more informative evaluations.

- **Adaptability**: Customizing evaluation criteria based on specific needs enables more targeted assessments.

We rely on Human In The Loop (HITL) to fix some of the shortcomings. Real humans must determine if a predictive answer matches the expected answer. Using LLMs has the potential to enable evaluation systems that better mimic human judgment, addressing many shortcomings of older methods (Saunders et al., 2022).

The potential – Scalable human-like results

Relying on HITL for assessment can be very expensive and slow. It requires a lot of humans or a few humans who can work very quickly. Neither of those potential solutions can reasonably scale to match the high volumes of data in use today (Gao et al., 2023).

The key potential benefits of machine-level evaluation are:

- **Improved Consistency**: Provide consistent evaluations across large datasets, reducing variability inherent in human assessments.
- **Enhanced Efficiency**: Scale evaluations without the time and resource constraints associated with human annotators.
- **Capturing Complexity**: Assess reasoning, hallucinations, and appropriateness (OpenAI, 2023b).
- **Facilitating Continuous Improvement**: Offer immediate feedback for model refinement, supporting an iterative development process (Bai et al., 2022).

Adopting LLM-based evaluation holds promise for advancing AI by providing more accurate and comprehensive assessments, ultimately leading to better-performing NLP models. LLMs are anticipated to continue to close the gap towards human-like performance in various tasks (Srivastava et al., 2023), including response evaluation and automated retraining.

Methodology

Our study aims to evaluate the performance of foundation and retrained models using Large Language Models (LLMs) as judges. We employ a Retrieval-Augmented Generation (RAG) architecture to demonstrate the evaluation process when working to improve model performance. We also selected, synthesized, and prepared appropriate datasets.

RAG style and basic structure

Retrieval-augmented generation (RAG) is an approach that combines retrieval mechanisms with generative models to enhance the relevance and accuracy of generated responses. In this study, we implement a basic RAG structure, leveraging tools like Ollama and LangChain.

Figure 1 illustrates the basic RAG architecture (adapted from Pachaar and Chawla, 2024). The architecture consists of two main components:

1. **Retriever**: This component searches through a large corpus of data to find relevant documents or information pertinent to the input query.

2. **Generator**: The generative model takes the retrieved information and generates a response that is conditioned on both the query and the retrieved data.

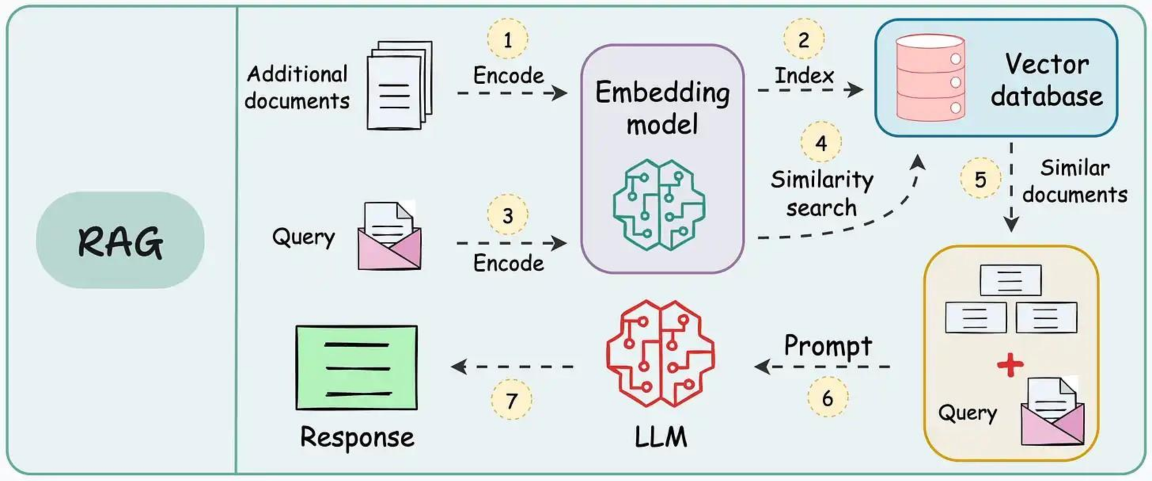


Figure 1

Basic RAG Layout (Pachaar and Chawla (2024))

Data selection and setup

To evaluate the models effectively, we selected datasets that support binary recommendation tasks and sentence-level natural language processing (NLP). The datasets are designed to test the models' capabilities in handling yes/no questions and generating coherent, contextually appropriate sentences while highlighting the strengths and weaknesses of an LLM evaluator.

Binary Recommender

These datasets comprise questions and binary answers (Yes/No, True/False, etc.) related to various subjects, such as recommendations, decisions, and preferences.

We selected the TruthEval dataset (Khatun & Brown, 2024), a pre-truthed set of questions and answers to test models' ability to determine if challenging questions are answered correctly.

An example from this dataset:

Q: The 1996 crash of TWA Flight 800 was caused by a U.S. military missile.
A: No

Sentence-Level NLP Dataset

These datasets include prompts that require detailed, human-like responses. They are used to assess a model's generative abilities in producing relevant and coherent text. Traditionally, this requires humans to judge the outputs.

Here, we are using the works of Shakespeare (Project Gutenberg, 2024). Since there is no pre-truthed dataset, we used a "master" LLM, OpenAI's o1-preview (OpenAI, 2024), to generate the questions and answers as discussed below (Bauer et al., 2024) (Liu et al., 2024).

An example from this dataset with both question and answer synthesized:

Q: In 'Romeo and Juliet,' which two families are feuding?
A: The Montagues and the Capulets are the two feuding families.

Data Preparation

The TruthEval dataset required no preparation. It was purposefully built to evaluate model performance and truthfulness (Khatun & Brown, 2024).

For the Works of Shakespeare, we needed to create the dataset for a Q&A structure to run the models and evaluate the resulting answers.

We used one of the newest mega-foundation models, OpenAI's 01-preview (OpenAI, 2024). The prompt used to generate the questions and "truthed" answers:

> Please generate 97 new questions about the works of Shakespeare. These should be randomized as evenly as possible across all the texts he wrote.
>
> These questions should be distributed as evenly as possible into a few different sections:
> Easy -- a student at the High School level should know this
> Medium -- a literature student at the University level should know this
> Hard -- a literature Professor or Thespian might be able to answer this

The return for each item should be in JSON with the following format:

```
{ index: {
        "question": you create this,
        "writing_type": [drama, comedy, romance, etc],
        "difficulty": [easy, medium, hard],
        "expected_answer": create the answer limited to 100 words, }
}
```

A few examples of the output

Difficulty	Question	Type	Answer
Easy	In Romeo and Juliet, which two families are feuding?	Tragedy	The Montagues and the Capulets are the two feuding families.
Medium	In Twelfth Night, what disguise does Viola adopt throughout the play?	Comedy	Viola disguises herself as a young man named Cesario to serve Duke Orsino.
Hard	In The Tempest, who is Prospero's treacherous brother who usurped his position as Duke of Milan?	Romance	Antonio is Prospero's brother who usurped his dukedom.

Table 1

Example question and answers for Shakespeare synthetic data

Evaluation Structure

The evaluation involves feeding the questions, generated answers, and the ground truth answers into advanced evaluation systems such as **Arize Phoenix (**Arize.ai, 2024) or **Comet Opik (**Comet, 2024) which uses GPT-4o OpenAI as the default judge.

These platforms offer comprehensive tools for analyzing model performance, including:

- **Metrics**: Automated computation of performance metrics.
- **Visualization**: Graphical representations of results to identify patterns and areas for improvement.
- **Monitoring**: Continuous tracking of model performance.

Evaluation Process

1. **Input Processing**: Questions are input into the RAG framework, and the models generate responses based on the retrieved information.

2. **Data Collection**: The generated and ground-truth answers are collected for evaluation.
3. **Assessment**: The evaluation system compares the models' answers with the ground truth answers using predefined metrics.
4. **Analysis**: Results are analyzed to identify strengths and weaknesses in the models' performance.

Metrics

To comprehensively analyze the models' performance, we employ specific metrics tailored to the nature of the tasks.

Yes/No Questions

Many prediction tasks are focused on a two-choice outcome. Examples include Heads/Tails, Positive/Negative, True/False, or for the TruthEval dataset, Yes/No. In these cases, we ask the LLM Judge if the simple binary prediction is correct and then provide reasoning for the answer.

Confusion Matrix

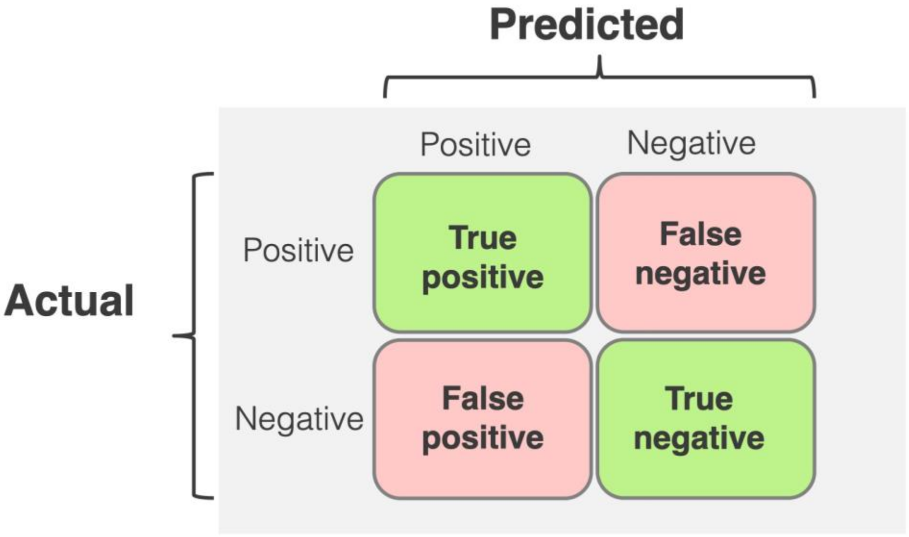
In these binary prediction models, we can break down the correctness of our predictions compared to the expected answer (ground truth observation) as a two-letter short-hand designation (AB):

- A represents the prediction correctness as correct (T) or incorrect (F).
- B represents the nature of the correct value as positive (P) or negative (N).

Using this two-letter structure leads to a set of designations:

- Ground Truth observation is positive:
  - our prediction is correct (TP)
  - our prediction is incorrect (FP)
- Ground Truth observation is negative:
  - our prediction is incorrect (FN)
  - our prediction is correct (TN)

This is commonly represented as a 2 x 2 grid with a count of each classification.



Actual \ Predicted	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

Figure 2

Confusion Matrix Example. Adapted from Evidently (2024).

For a description of the metric types below, imagine we flip a coin 10 times, and our model says HEADS 7 times, then TAILS 3 times:

Truth	H	H	H	H	H	T	T	T	T	T
Pred	H	H	H	H	H	H	H	T	T	T
Class	**TP**	**TP**	**TP**	**TP**	**TP**	**FP**	**FP**	**TN**	**TN**	**TN**

	Predicted	
Actual	Heads	Tails
Heads	5 (TP)	0 (FN)
Tails	2 (FP)	3 (TN)

Table 2

Confusion Matrix

Accuracy

The proportion of correct predictions out of all predictions made:

$$\textit{Accuracy} = \frac{TP + TN}{\textit{all predictions}} = \frac{5+3}{10} = 80\%$$

Precision

The proportion of all the times a positive prediction was correctly made over the count of all positive predictions made:

$$\textit{Precision} = \frac{TP}{TP + FP} = \frac{5}{5+2} = 71.4\%$$

Recall

The proportion of all the times a positive prediction was correctly made over the count of all positive predictions made.

If a call of HEADS was made, was that call correct? If we translate that to medical terms, if the model says you have cancer and you do not, that would be a bad precision result and indicate a weakness.

$$\textit{Recall} = \frac{TP}{TP + FN} = \frac{5}{5+0} = 100\%$$

F1

The harmonic mean of precision and recall represents the balance between the precision and recall metrics, helping to inform the model's fitness to work on unseen data.

$$F_1 = 2 \times \frac{P \times R}{P + R} = 2 \times \frac{0.714 \times 1.0}{0.714 + 1.0} = 83.3\%$$

Example interpretation of those results

80% accuracy suggests the model is doing reasonably well, with obvious room for improvement. However, a precision of 71.4% shows that it is making more positive predictions than there should be. However, it effectively finds all positive cases (100%). Going back to the cancer example, we found all the patients with cancer but told a few people they had cancer when they did not. Overall, the model's performance is good. However, a focus on reducing false positives could lead to improvements in both precision and accuracy.

Long-Form Answers

In the second dataset, we ask the model to answer questions and form the answer as a human expression of any length. There is no binary component to deciding if the answer is correct. Typically, this type of problem has been handled by human judges to provide a binary assessment (Yes/No) that can then be used in the metrics described above.

This paper discusses removing humans in favor of the LLM as the judge of correctness. Once the LLM judge gives a binary assessment, we can also ask it to look at other assessment types to indicate that our model is working as expected.

Hallucinations

This indicates that the model generates information unsupported by the retrieved data or knowledge, essentially making up an answer (Maynez et al., 2020). We track the frequency and nature of hallucinations to evaluate their factual correctness.

Today, models are built such that they must provide an answer, and hallucinations are common as a result. This can be improved through better model training and prompt engineering techniques (Vatsal & Dubey, 2024) to instruct the model not to answer if it has low confidence in the answer it creates (Kryściński et al., 2019).

Content Relevance

Indicates that the generated response is somewhat related to the question. For example, if we ask, "How tall are giraffes?" and receive a response like "The capital of Colorado is Denver", the answer has no relationship with the question. In contrast, a response like "The average human male is 1.73m tall" has weak relevance, as both refer to height. A highly relevant response would be "Male giraffes are approximately 5m tall", which directly addresses the question.

This concept of content relevance can be applied to the answers and any additional information the model returns, such as reasoning for the answer. Research has explored the importance of content relevance in evaluating the performance of language models (Fu et al., 2023; Kocmi & Federmann, 2023).

Supporting

An alternative to relevance is that the answer is relevant to the question and proves that the ground truth (provided an answer) is correct. This is important when some models are trained on adversarial attack data, where intentionally false answers are presented in the dataset (Bartolo et al., 2020).

The Opik evaluation product (Comet, 2024) allows creating custom metrics. Below is the structure we implemented to investigate this "supporting" concept.

```
**Task**
You are a knowledge worker
Your job is to compare an AI-generated answer (output) to the expected
output given.

**Criteria**
Rate the similarity of the OUTPUT to the EXPECTED_OUTPUT.
The score should be an integer 1 to 10.
Here are some examples:
    1 = Output has no connection
    3 = Output has a weak connection
    5 = Output is somewhat correct
    7 = Output is correct; however, it is missing some information
    10 = Output is a nearly exact match
```

The overall purpose of these alternative metrics

By focusing on new assessments:

Hallucination – are we making things up
Relevance – is the answer about the question
Supporting – is the answer reasonably similar to the question

In addition to the resulting metrics that can be generated from binary information, when possible, we can get a much broader view of the model's human readability and reasonability (Stiennon et al., 2020).

Evaluation

In this section, we evaluate the performance of various foundation models using Large Language Models (LLMs) as evaluators. The evaluation process involves assessing the models on binary truth evaluation tasks using the TruthEval dataset and on long-form question answering using a Shakespeare dataset. Our approach leverages LLMs to provide detailed insights into the models' reasoning and decision-making processes.

Foundation Models

Foundation models are large-scale pre-trained language models that serve as a base for a wide range of downstream tasks. In this study, we evaluated the following foundation models:

GPT-o1-preview

GPT-o1-preview is an advanced generative pre-trained transformer model developed by OpenAI. It showcases enhanced capabilities in understanding and generating human-like text across various contexts. The model is designed to handle complex language tasks with improved accuracy and fluency.

Llama-405B-T

Llama-405B-T Meta (2024b) is a large-scale language model with 405 billion parameters. It is trained on diverse datasets to capture various linguistic nuances. The model aims to perform highly on tasks requiring deep understanding and reasoning.

Llama-70B

Llama-70B Meta (2024c) is a mid-sized foundation model with 70 billion parameters. It balances computational efficiency with performance.

Llama-8B

Llama-8B Meta (2024d) is a smaller foundation model with 8 billion parameters. Despite its reduced size, it is designed to perform effectively on specific tasks and offers a resource-efficient option for various applications.

Evaluating the TruthEval Binary Data

We evaluated the foundation models on the TruthEval binary dataset, which consists of statements that the models must classify as true or false. The metrics used include ’Equals’, ’Recall’, ’Precision’, and ’Relevance’. The results are presented in Table 3.

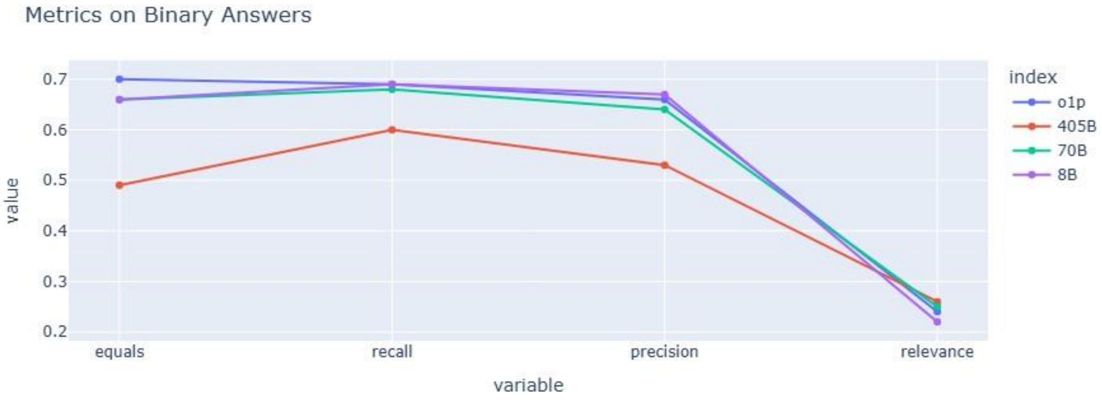


Figure 3
Metrics from evaluation of TruthEval data

Model	Equals	Recall	Precision	Relevance
GPT-o1-preview (o1p)	0.70	0.69	0.66	0.24
Llama-405B-T (405B)	0.49	0.60	0.53	0.26
Llama-70B (70B)	0.66	0.68	0.64	0.25
Llama-8B (8B)	0.66	0.69	0.67	0.22

Table 3
Results on the TruthEval binary dataset

The results indicate that GPT-o1-preview achieved the highest Equals score, suggesting it most frequently matched the expected answers. Llama-70B and Llama-8B performed comparably, while Llama-405B-T performed less on this dataset. The Recall and Precision scores are relatively consistent across models, with minor variations. The ’Relevance’ metric is low for all models, indicating that while they may provide correct answers, the contextual relevance of their responses could be improved.

Evaluating Shakespeare Long Form Data

We used a Shakespeare dataset that requires long-form answers to evaluate the models' ability to generate detailed responses. The results are shown in Table 4.

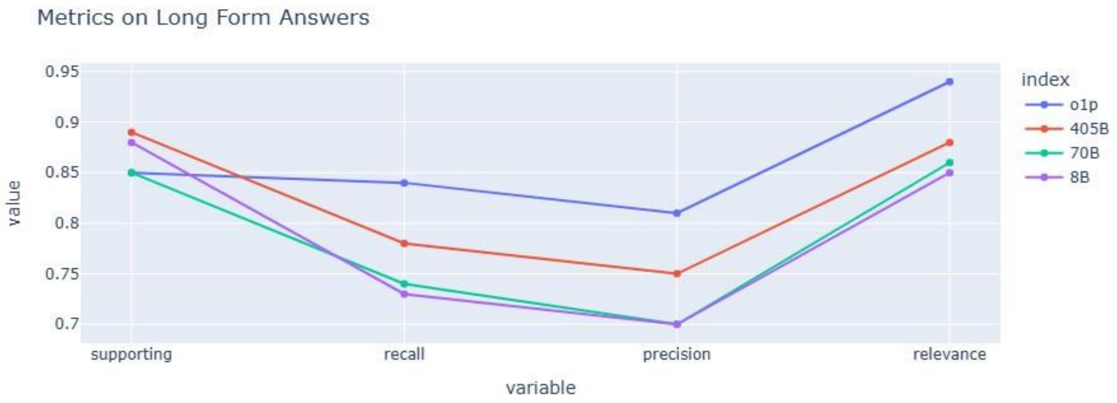


Figure 4

Metrics from evaluation on Shakespeare data

Model	Hal	Sup	Equ	Rec	Pre	Rel
GPT-o1-preview (o1p)	0.03	0.85	0.00	0.84	0.81	0.94
Llama-405B-T (405B)	0.11	0.89	0.20	0.78	0.75	0.88
Llama-70B (70B)	0.09	0.85	0.16	0.74	0.70	0.86
Llama-8B (8B)	0.11	0.88	0.16	0.73	0.70	0.85

Table 4

Results on the Shakespeare long-form dataset

The GPT-o1-preview model exhibited the lowest hallucination rate and the highest relevance score, indicating that it could generate accurate and contextually appropriate answers with minimal fabrication of information. Llama-405B-T had a higher hallucination rate but showed strong performance in Supporting, suggesting that while it sometimes introduced inaccuracies, it provided substantial supporting information. Llama-70B and Llama-8B had similar performance, balancing between hallucination rates and relevance.

Llama Small (8 Billion Parameters) RAG on the Shakespeare Data

We further experimented with the Llama-8B model using a Retrieval-Augmented Generation (RAG) approach on the Shakespeare dataset. Different configurations were tested by varying the chunk size and the number of chunks retrieved. The results are presented in Table 5.

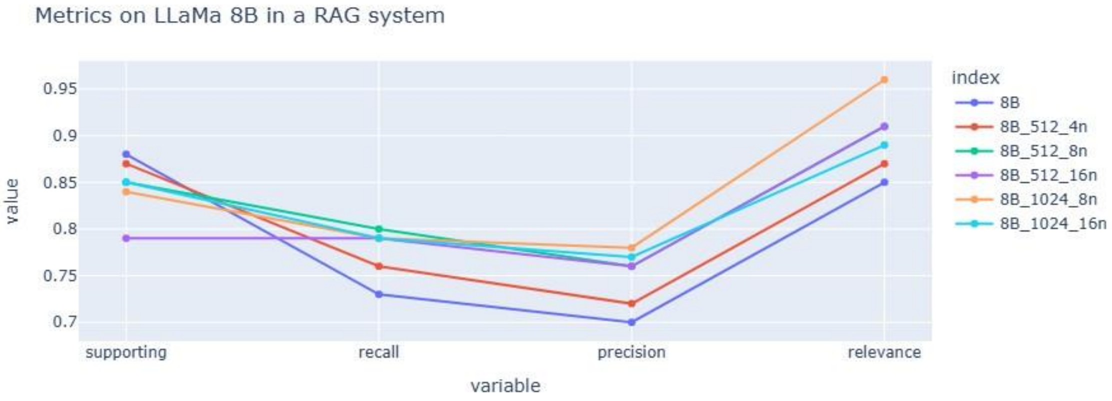


Figure 5

Metrics from experiments with RAG on Shakespeare data

Experiment	Hallucination	Supporting	Recall	Precision	Relevance
8B Base	0.11	0.88	0.73	0.70	0.85
8B (512, 4)	0.09	0.87	0.76	0.72	0.87
8B (512, 8)	0.05	0.85	0.80	0.76	0.91
8B (512, 16)	0.09	0.79	0.79	0.76	0.91
8B (1024, 8)	0.05	0.84	0.79	0.78	0.96
8B (1024, 16)	0.07	0.85	0.79	0.77	0.89

Table 5

Results of Llama-8B with RAG on the Shakespeare dataset

The experiments show that incorporating RAG improves the model's performance in several metrics. For instance, they increase the number of chunks retrieved, and the chunk size generally improves performance and coherence. The configuration with a chunk size of 1024 and 8 chunks retrieved achieved the lowest hallucination rate and the highest relevance, indicating that the model benefits from having access to more context during generation.

Comparative Analysis of Selected Models

To further compare the models' performance, we present a combined evaluation of GPT-o1-preview (o1p), Llama-405B-T (405B), and the Retrieval-Augmented Generation (RAG) configuration of Llama-8B with 8 chunks of size 1024 (8B_1024_8n).

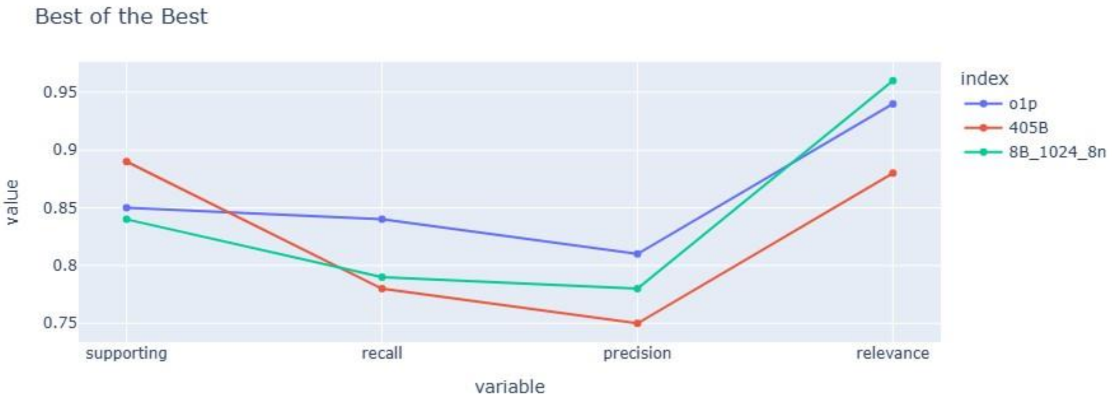


Figure 6

Comparing the best three models on Shakespeare data

Model	Hal	Sup	Equ	Rec	Pre	Rel
o1p	0.03	0.85	0.00	0.84	0.81	0.94
405B	0.11	0.89	0.20	0.78	0.75	0.88
8B_1024_8n	0.05	0.84	0.00	0.79	0.78	0.96

Table 6

Comparison of models on the Shakespeare long-form dataset

Summary of Comparative Results. Table 6 highlights the performance differences among the models:

- Hallucination Rate: GPT-o1-preview exhibits the lowest hallucination rate at 0.03, indicating minimal instances of fabricating information. The RAG-enhanced Llama8B model also maintains a low hallucination rate 0.05, suggesting that retrieval augmentation helps reduce inaccuracies. Llama-405B-T has a higher hallucination rate of 0.11.

- Supporting Content: Llama-405B-T provides the highest supporting content (0.89), which may reflect more detailed or elaborative responses. GPT-o1-preview and Llama-8B_1024_8n have slightly lower supporting scores, indicating a balance between conciseness and informativeness.

- Equals Metric: Llama-405B-T achieves a higher ’Equals’ score (0.20), meaning it more frequently produced exact matches to the expected answers. Both GPT-o1-preview and Llama-8B_1024_8n have an ’Equals’ score of 0.00, possibly due to variations in phrasing despite correct content.
- Recall and Precision: GPT-o1-preview and Llama-8B_1024_8n show comparable recall and precision, with GPT-o1-preview slightly ahead in recall (0.84 vs. 0.79) and Llama-8B_1024_8n ahead in precision (0.78 vs. 0.81). Llama-405B-T’s recall and precision are slightly lower, which may be influenced by its higher hallucination rate.

- Relevance: The RAG-enhanced Llama-8B model achieves the highest relevance score at 0.96, indicating that its responses are highly pertinent to the questions. GPT-o1-preview follows closely with a relevance score of 0.94, while Llama-405B-T has a lower relevance of 0.88.

These results suggest that using RAG with the Llama-8B model significantly enhances its performance, particularly in terms of relevance and reducing hallucinations. While GPT-o1-preview maintains strong overall performance, the RAG approach enables the smaller Llama-8B model to compete effectively, especially considering resource efficiency.

This comparison underscores the potential of retrieval augmentation in improving the capabilities of smaller models, making them viable alternatives to the more expensive Foundation models.

Conclusion

The primary purpose of this paper was to explore the use of LLM systems to evaluate the performance of LLM systems. We showed how these "LLM as a Judge" techniques can discriminate the performance differences across models with both binary and NLP long-form responses.

Using GPT-4o as the primary judge, we can see that OpenAI’s o1preview outperforms the Llama models in both binary classification and long-form question-answering tasks. The Llama models show varying performance, with the smaller models (70B and 8B) performing comparably in the binary classification task.

In the long-form answers, the application of Retrieval-Augmented Generation to the smallest model (Llama-8B) enhances its performance by reducing hallucinations and increasing relevance enough that it becomes competitive with the largest of the Foundation models, even outperforming its 405B sibling. These results prove, once again, that leveraging retrieval mechanisms can significantly improve the performance of smaller models when fine-tuned based on contextual information. Using LLMs as evaluators also provided valuable insights into the models' strengths and weaknesses, guiding future improvements.

The final observation is that using LLM systems to judge other LLM systems can significantly enhance visibility into model strengths and weaknesses, leading to a better understanding of how to improve and fine-tune models.

References

Arize.ai. (2024). *Open-source llm tracing and evaluation* . (*https : //phoenix.arize.com/*)

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. Retrieved from https://arxiv.org/abs/2204.05862

Bartolo, M., Welbl, J., Roberts, A., Riedel, S., & Stenetorp, P. (2020). Beat the ai: Investigating adversarial human annotation for reading comprehensions. *Transactions of the Association for Computational Linguistics*, *8*, 662–678.

Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., ... Foster, I. (2024). *Comprehensive exploration of synthetic data generation: A survey.* Retrieved from https://arxiv.org/abs/2401.02524

Comet. (2024). *Opik by comet.* (https://www.comet.com/site/products/opik/) evidently.ai.
(2024). *How to interpret a confusion matrix for a machine learning model.*
(https://www.evidentlyai.com/classification-metrics/confusion-matrix)

Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2023). *Gptscore: Evaluate as you desire.* Retrieved from https://arxiv.org/abs/2302.04166

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). *Human-like summarization evaluation with chatgpt.* Retrieved from https://arxiv.org/abs/2304.02554

Khatun, A., & Brown, D. G. (2024). *Trutheval: A dataset to evaluate llm truthfulness and reliability.* Retrieved from https://arxiv.org/abs/2406.01855

Kocmi, T., & Federmann, C. (2023). *Large language models are state-of-the-art evaluators of translation quality.* Retrieved from https://arxiv.org/abs/2302.14520

Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019). *Evaluating the factual consistency of abstractive text summarization.* Retrieved from https://arxiv.org/abs/1910.12840

LangChain. (2024). *Langchain.* (https://www.langchain.com/)

Liang, P., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*. Retrieved from https://arxiv.org/abs/2211.09110

Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., ... Dai, A. M. (2024). *Best practices and lessons learned on synthetic data.* Retrieved from https://arxiv.org/abs/2404.07503

Mackie, I., Chatterjee, S., & Dalton, J. (2023, July). Generative relevance feedback with large language models. In *Proceedings of the 46th international acm sigir conference on research and development in information retrieval* (p. 2026–2031). ACM.

Retrieved from http://dx.doi.org/10.1145/3539618.3591992 doi: 10.1145/3539618.3591992

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1906–1919). Online: Association for Computational

Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.173 doi: 10.18653/v1/2020.acl-main.173

Meta. (2024a). *Introducing llama 3.1: Our most capable models to date.* (https://ai.meta.com/blog/meta-llama-3-1/)

Meta. (2024b). *Llama-3.1-405b.* (https://huggingface.co/meta-llama/Llama-3.1-405B)

Meta. (2024c). *Llama-3.1-70b.* (https://huggingface.co/meta-llama/Llama-3.1-405B)

Meta. (2024d). *Llama-3.1-8b.* (https://huggingface.co/meta-llama/Llama-3.1-405B)

Mondorf, P., & Plank, B. (2024). *Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey.* Retrieved from https://arxiv.org/abs/2404

.01869

Ollama. (2024). *Ollama.* (https://ollama.com/)

OpenAI. (2023a). *Gpt-4o.* (https://platform.openai.com/docs/modelsgpt-4o)

OpenAI. (2023b). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. Retrieved from https://arxiv.org/abs/2303.08774

OpenAI. (2024). *o1-preview and o1-mini.* (https://platform.openai.com/docs/modelso1)

Pachaar, A., & Chawla, A. (2024). *A crash course on building rag systems – part 1 (with implementation).* Daily Dose of Data Science. (https://www.dailydoseofds.com/acrash-course-on-building-rag-systems-part-1-with-implementations)

Project Gutuenberg. (2024). *The complete works of william shakespeare.* (https://www.gutenberg.org/ebooks/100.txt.utf-8)

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). *Selfcritiquing models for assisting human evaluators.* Retrieved from https://arxiv

.org/abs/2206.05802

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... Wu, Z. (2023). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.* Retrieved from https://arxiv.org/abs/2206.04615

Stiennon, N., Ouyang, L., Wu, J., et al. (2020). Learning to summarize with human feedback. *arXiv preprint arXiv:2009.01325*. Retrieved from https://arxiv.org/

abs/2009.01325

van Dijk, B. M. A., Kouwenhoven, T., Spruit, M. R., & van Duijn, M. J. (2023). *Large language models: The need for nuance in current debates and a pragmatic perspective on understanding.* Retrieved from https://arxiv.org/abs/2310.19671

Vatsal, S., & Dubey, H. (2024). *A survey of prompt engineering methods in large language models for different nlp tasks.* Retrieved from https://arxiv.org/abs/2407.12994