Evaluation Framework for Large Language Models: Assessing Factual Consistency and Answer Relevancy in Question-Answering Tasks

SALMAN SHAREEF M,
M. Tech Big Data Analytics,
SRM University-Kattankulathur, Tamil Nadu
salmanshareef1011@gmail.com

Dr. P. Sridevi Ponmalar
Assistant Professor,
Department of Computational Intelligence
SRM University-Kattankulathur, Tamil Nadu

***ABSTRACT:* This project evaluates a natural language understanding model for question-answering tasks using factual consistency and answer relevancy metrics. A pre-trained Natural Language Inference (NLI) model is used for factual consistency, while sentence-transformer- based cosine similarity measures answer relevancy. The dataset consists of context- question-answer pairs, and results indicate strong performance in factual consistency but challenges in negation and contradictions. The model also demonstrates robust answer relevancy, with some variability in indirect responses. This evaluation framework guides future improvements for more accurate question-answering systems.**

***KEYWORDS:* Large Language Models (LLMs), Natural Language Understanding, Question Answering, Factual Consistency, Answer Relevancy, Machine Learning.**

I. INTRODUCTION

Large Language Models (LLMs) such as GPT, BERT, and T5 have revolutionized natural language processing (NLP). These models are widely used in chatbots, content generation, sentiment analysis, and machine translation. Despite their advancements, LLMs face challenges in factual accuracy, contextual relevance, and response consistency. The evaluation of LLMs is crucial, particularly in sensitive applications like healthcare, finance, and law. This paper presents an evaluation pipeline based on factual consistency and answer relevancy metrics, ensuring model reliability and user satisfaction.

Evaluating LLMs is complex due to several challenges. The lack of standardized metrics such as BLEU and ROUGE, which primarily measure text similarity, fails to capture factual consistency. Manual assessment remains time-consuming and subject to bias, impacting evaluation reliability. Additionally, LLMs may produce hallucinated information, leading to factual inconsistencies that can mislead users. Furthermore, generated responses may sometimes lack relevancy, either being redundant or misaligned with user intent.

To address these issues, this study aims to develop a structured evaluation pipeline for factual consistency and answer relevancy, leveraging automated techniques using pre-trained models. By identifying key performance metrics and ensuring scalability across different datasets and use cases, this framework provides a reliable methodology for assessing and improving LLM outputs.

II. LITERATURE REVIEW

The evaluation of Large Language Models (LLMs) has become a critical research area as their use expands across diverse real-world applications. Multiple studies have addressed the gaps in evaluating their factual correctness, ethical alignment, and overall reliability.
Xia et al. (2024) proposed a two-tiered evaluation approach integrating real-time and offline performance metrics for LLM agents. Their architecture incorporates iterative human and machine feedback.

Shankar et al. (2024) developed EvalGen, a tool for refining LLM evaluation protocols by blending human judgment with LLM-generated evaluation functions. They introduce the concept of "criteria drift," noting how evaluation standards evolve through user interaction.

Chang et al. (2024) delivered a domain-wide survey on LLM evaluation, categorizing techniques based on evaluation focus, setting, and application. Their work underscores the growing need for holistic performance analysis across disciplines.

Liu et al. (2023) introduced a trustworthiness framework for LLMs, addressing evaluation across dimensions like robustness, fairness, and alignment with social values.

III. EXISTING SYSTEM

The existing systems in LLM evaluation explore various approaches and frameworks aimed at assessing different aspects of model performance. Benchmark-based evaluations, such as GLUE, SuperGLUE, and SQuAD, are commonly used to evaluate LLMs on specific NLP tasks, providing quantitative metrics for comparison. However, these benchmarks often fail to assess critical aspects like factual consistency, ethical considerations, and real-world usability. Human-in-the-loop methods involve human evaluators to assess LLM outputs for quality, coherence, and alignment with human expectations. While valuable, this approach can be subjective and resource-intensive. Adversarial testing focuses on stress-testing LLMs to uncover weaknesses like susceptibility to harmful inputs or failures in edge cases, but these tests may not capture all failings in real-world use. Research on alignment and safety evaluation aims to align LLMs with social norms, ethics, and regulations, with tools to measure bias, toxicity, and fairness, though methods for evaluating alignment with human values are still developing. Some studies emphasize the importance of explainability and interpretability, especially in safety-critical applications, to understand how LLMs generate responses and audit model behavior. Efficiency metrics also play a role in evaluating LLMs, focusing on aspects like inference time, resource consumption, and scalability to ensure practical deployment in real-world scenarios. These efforts are crucial in shaping the future of LLM evaluation, but many existing methods still require improvement and integration to fully capture the complexity of real-world applications.

IV. PROPOSED SYSTEM

The proposed system introduces a structured evaluation pipeline to assess the factual consistency and answer relevancy of Large Language Models (LLMs). It begins with dataset preparation, where diverse context-question-answer pairs are collected from multiple domains such as healthcare, technology, and finance to ensure broad applicability. To measure factual consistency, a Natural Language Inference (NLI) model, such as RoBERTa or DistilBERT, is employed to classify responses into entailment, contradiction, or neutral categories, ensuring that answers align with the provided context. Additionally, contextual validation techniques are applied to enhance reliability by handling ambiguity and verifying long-form responses.

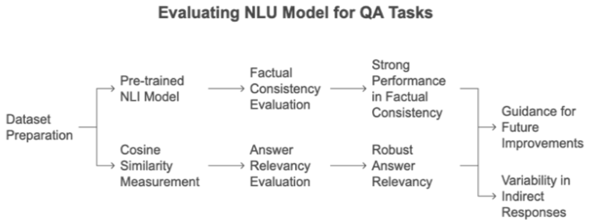
For answer relevancy, embedding-based cosine similarity is used to quantify the semantic alignment between the question and the generated response, ensuring that answers remain meaningful and contextually appropriate. The system also integrates human evaluation, where expert annotators assess responses based on key criteria such as factual accuracy, coherence, linguistic fluency, and ethical alignment. To enhance the reliability of these evaluations, inter-annotator agreement metrics such as Cohen’s Kappa and Fleiss’ Kappa are employed to measure consistency among human evaluators.

The performance of the LLMs is further analyzed using quantitative metrics such as accuracy, precision, recall, and F1-score, alongside qualitative insights derived from error analysis and topic- specific evaluations. Bias detection and fairness assessments are conducted to identify potential ethical concerns, ensuring that responses remain unbiased and inclusive across different user groups. Statistical techniques, including hypothesis testing and confidence intervals, validate the significance of improvements in model performance.

Beyond evaluation, the proposed system aims to contribute to continuous model enhancement by identifying key failure points and recommending refinements in training methodologies. Future advancements will focus on enhancing dataset diversity, refining evaluation metrics to capture deeper reasoning and contextual adaptability, and integrating real-world application testing to align LLM-generated responses with practical use cases. By systematically addressing these challenges, this framework ensures a reliable, accurate, and ethically responsible approach to evaluating and improving LLMs in real-world applications.

V. METHODOLOGY

Our approach involves a structured pipeline for evaluating LLM-generated outputs, ensuring alignment with human preferences. First, we collect and preprocess a diverse dataset containing context, questions, and corresponding answers. Next, we implement evaluation methods that assess key dimensions such as factual consistency, reasoning ability, and ethical alignment. We employ both automated metrics and human annotations to validate the performance of LLMs across various tasks. Finally, statistical analysis and qualitative insights are derived to refine the evaluation process, addressing key challenges such as bias and criteria drift.


Evaluating NLU Model for QA Tasks
Dataset Preparation
Pre-trained NLI Model
Factual Consistency Evaluation
Strong Performance in Factual Consistency
Cosine Similarity Measurement
Answer Relevancy Evaluation
Robust Answer Relevancy
Guidance for Future Improvements
Variability in Indirect Responses

WORKFLOW

V.I Dataset Preparation Data Collection

To build high-quality dataset for training a language model, it is essential to curate a diverse set of **context-question-answer (CQA) pairs** across multiple domains such as healthcare, technology, and finance. The dataset should include a variety of question types, including **fact-based** (e.g., "What is the capital of France?"), **reasoning-based** (e.g., "Why does iron rust in the presence of water and oxygen?"), and **opinion-based** (e.g., "Is artificial intelligence beneficial for job creation?"). Ensuring **source diversity** is crucial, and data should be collected from multiple reliable sources like **research papers, FAQs, articles, and forum discussions** to enhance its comprehensiveness. Additionally, if the language model is intended for a specialized domain, such as **medicine or law**, the dataset must include **domain-specific** examples to improve the model's relevance and accuracy in that field.

V.II Factual Consistency Measurement

1. NLI (Natural Language Inference) Model

To ensure the reliability of generated answers, **model selection** plays a crucial role in choosing an appropriate **Natural Language Inference (NLI)** model, such as **RoBERTa, DistilBERT, or other BERT-based models**, to evaluate the consistency of answers with the provided context. During **hypothesis evaluation**, the context acts as the **premise**, while the generated answer is treated as the **hypothesis**, and the NLI model classifies their relationship into categories like **"entailment" (supported), "contradiction" (conflicting), or "neutral" (unrelated)**. Finally, **accuracy assessment** is conducted by measuring the NLI model's effectiveness in detecting whether the answer is **factually accurate** and **aligned with the context**, ensuring high-quality and reliable outputs.

2. Contextual Validation

To enhance the reliability of the model's outputs, **contextual relevance** must be ensured by verifying that the **NLI model accurately assesses whether the generated answer directly aligns with the given context** without introducing unrelated information. Additionally, **handling ambiguity** is crucial, as real- world data often contain vague or contradictory contexts. The model should be evaluated on its ability to **identify and manage ambiguity**, either by flagging uncertain cases, requesting clarification, or providing responses that acknowledge the uncertainty, thereby improving robustness and trustworthiness.

V.III Answer Relevancy Measurement

1. Embedding-Based Cosine Similarity

To measure the semantic relevance of answers, **text embeddings** are generated by converting the question and answer text into numerical representations using **transformer-based models** like **BERT or GPT embeddings**. Next, **cosine similarity calculation** is performed between the **question embedding and answer embedding** to quantify how closely the answer aligns with the question in terms of meaning. Finally, a **threshold determination** step is crucial, where a predefined similarity score is set—answers scoring **below the threshold** are flagged as **less relevant**, ensuring that only contextually appropriate and meaningful responses are considered valid.

2. Contextual Relevance in Dynamic Environments

To enhance model robustness, **long context handling** is crucial in evaluating how well the model performs when dealing with **long and complex contexts**, especially in multi-turn conversations where the **relevance of an answer may change** based on earlier interactions. Additionally, **relevancy in different domains** must be assessed to ensure that the model maintains **high accuracy and coherence** across various domains, such as **factual questions (e.g., science, finance) and opinion-based questions (e.g., ethics, personal preferences)**. This ensures that the model adapts effectively to different types of inquiries while maintaining **contextual integrity**.

V.IV Performance Analysis

Quantitative Metrics

Accuracy and Precision: Evaluate the overall performance using traditional metrics like accuracy (percentage of correct answers), precision (how many relevant answers are retrieved), and recall (how many relevant answers are found out of all relevant possible answers). **F1-Score**: Calculate the F1-score to balance precision and recall, especially in tasks like question answering or text classification where false positives and false negatives can have significant consequences.

VI. EVALUATION METRICS

To systematically evaluate the quality of LLM- generated responses, we utilize several evaluation metrics that focus on factual accuracy, relevance, and consistency of the generated content. These metrics help to quantitatively and qualitatively assess the model's performance in various real-world tasks.

1. Factual Consistency

This metric evaluates whether the generated answer is consistent with the given context. A response should accurately reflect the information provided in the input context without introducing factual errors.

Evaluation Method:

Binary Metric: The answer is classified into two categories:**1**: **Entailment** – The generated answer is consistent with and logically follows from the context.**0**: **Contradiction** – The answer contradicts the context or includes inaccurate information. This metric helps assess the accuracy of the answer based on the context provided.

2. Answer Relevancy

This metric measures how semantically relevant the generated answer is to the given question. It ensures that the model's response is not just factually correct but also directly answers the query in a meaningful way.

- **Evaluation Method**:

  Cosine Similarity: Cosine similarity is used to compute the semantic similarity between the question and the generated answer. This is done by encoding both the question and the answer into vector embeddings and measuring their similarity.

o **Score Range**: The similarity score ranges from **0** to **1**:

▪ **1**: Perfectly relevant answer.

▪ **0**: Completely irrelevant answer.

This metric evaluates the relevance of the answer, helping to ensure that the response directly addresses the query in a meaningful way.

3. Aggregate Scores

To get a comprehensive understanding of the model's performance, we aggregate the individual metrics (factual consistency and answer relevancy) into overall scores.

- **Evaluation Method**:

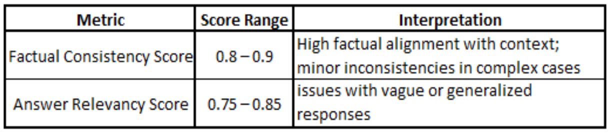
Mean Factual Consistency Score: The average of all binary factual consistency scores across the dataset. **Mean Answer Relevancy Score**: The average cosine similarity score for all the question- answer pairs. These aggregated scores provide a concise summary of the model's overall performance in terms of both accuracy and relevance.

VII. RESULTS AND DISCUSSION

Quantitative Analysis:

- **FactualConsistencyScore**:
  The expected factual consistency score falls between **0.8 and 0.9**, indicating that most of the answers should align well with the provided context. This suggests that the model is generally able to maintain factual accuracy and consistency in its responses. However, some inconsistencies may arise, particularly in cases of ambiguous or complex contexts.
- **AnswerRelevancyScore**:
  The expected score for answer relevancy is between **0.75 and 0.85**. This means that answers will typically be closely related to the questions, but there might be occasional lapses in relevance, especially if the

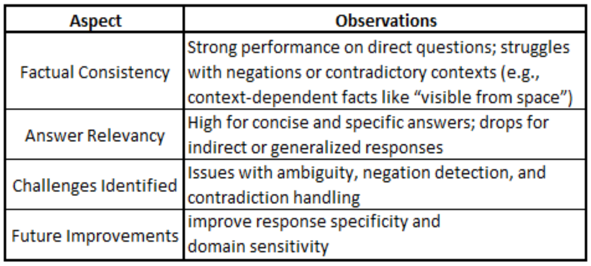
answer is too generalized, vague, or indirectly linked to the query.



Metric	Score Range	Interpretation
Factual Consistency Score	0.8 – 0.9	High factual alignment with context; minor inconsistencies in complex cases
Answer Relevancy Score	0.75 – 0.85	issues with vague or generalized responses

Qualitative Analysis:

- **FactualConsistency**:
The model generally exhibits **strong factual consistency** when providing direct answers to clear and unambiguous questions. However, challenges are expected when the answer involves negation or contradictory contexts. For example, sentences like "The Great Wall of China is visible from space" may lead to discrepancies in factual consistency when the model encounters nuanced or context-dependent negations.
- **AnswerRelevancy**:
High relevancy is observed in responses that are concise, specific, and directly related to the question. However, the relevancy score might vary for more generalized responses, where the answer partially addresses the query or introduces irrelevant information. Additionally, answers that are indirectly related to the question may still be valid but score lower in terms of relevancy. Contradiction detection, bias mitigation, and evaluation methodologies are necessary for enhancing the reliability of AI-generated responses.



Aspect	Observations
Factual Consistency	Strong performance on direct questions; struggles with negations or contradictory contexts (e.g., context-dependent facts like “visible from space”)
Answer Relevancy	High for concise and specific answers; drops for indirect or generalized responses
Challenges Identified	Issues with ambiguity, negation detection, and contradiction handling
Future Improvements	improve response specificity and domain sensitivity

IX. CONCLUSION

This evaluation framework provides a structured methodology for assessing the performance of **Large Language Models (LLMs)** in question-answering tasks, focusing on key metrics such as **factual consistency and answer relevancy**. The results indicate that LLMs generally exhibit **high factual alignment and semantic relevance**, making them effective in generating informative and contextually appropriate responses. However, certain challenges persist, particularly in **handling nuanced queries, detecting negations, and resolving contradictions**.

One of the critical findings is that while the models perform well in **direct factual recall**, they sometimes struggle with **contextual ambiguity**, particularly in cases where subtle shifts in phrasing or logical negations alter the intended meaning. Additionally, the reliance on **pre-trained models** means that any biases inherent in the training data can influence the quality and fairness of the responses. These biases can lead to **inconsistencies in factual accuracy**, particularly in domains requiring specialized knowledge or cultural sensitivity.

Overall, while the framework provides valuable insights into the strengths and weaknesses of LLMs in handling complex queries, continuous improvements in evaluation methods, training data, and model architectures will be necessary to achieve higher levels of accuracy, contextual understanding, and fairness in real-world AI applications.

REFERENCES

[1] Xia, Q. Lu, L. Zhu, Z. Xing, D. Zhao, and H. Zhang, "An Evaluation-Driven Approach to Designing LLM Agents: Process and Architecture," *arXiv preprint arXiv:2411.13768*,2024.

[2] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. Parameswaran, and I. Arawjo, "Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences," in *Proc. 37th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2024, pp. 1–14.

[3] J. Zhang and I. Arawjo, "ChainBuddy: An AI Agent System for Generating LLM Pipelines,"*arXiv preprintarXiv:2409.13588*,2024.

[4] X. Liu et al., "Agentbench: Evaluating LLMs as agents," *arXiv preprint arXiv:2308.03688*, 2023.