Bridging the Gap in Domain-Specific RAG Evaluation: **A**daptive **D**omain-**A**ware **M**etric **S**election Framework

Kai Li
Northeastern University
Vancouver, Canada
li.kai4@northeastern.edu

Muzhou Liu
Northeastern University
Vancouver, Canada
liu.muz@northeastern.edu

Xueyun Li
Northeastern University
Vancouver, Canada
li.xueyun@northeastern.edu

Yao Cheng
Northeastern University
Vancouver, Canada
cheng.yao1@northeastern.edu

Abstract*—The widespread adoption of Large Language Models (LLMs) integrated with Retrieval-Augmented Generation (RAG) systems across various industries necessitates robust evaluation methods. However, evaluating RAG systems, particularly in domain-specific contexts, remains challenging due to the limitations of standardized metrics and the unique requirements of different domains. Existing evaluation approaches often lack adaptability and fail to capture domain-specific nuances, leading to inaccurate assessments. To address these gaps, this paper introduces the Adaptive Domain-Aware Metric Selection (ADAMS) framework. ADAMS proposes an automated pipeline to compile, test, and dynamically select evaluation metrics tailored to specific datasets and user needs. The methodology involves validating metrics based on discriminative power (Cohen's d) and rewrite consistency (Variance Ratio, VR) using synthetically generated errors across diverse technical datasets. An LLM-based agent optimizes the metric suite by weighting metrics based on dataset properties and user preferences. Experimental results across three domains demonstrated significant variations in the performance of 15 metrics, identifying `answer_equivalence` as the most balanced overall. Real-world validation using industry partner data showed that ADAMS significantly outperforms existing benchmarks like RAGAS [8], achieving higher Area Under the Curve (AUC) scores (0.88–0.93 compared to 0.68) and better alignment with expert judgments. The framework demonstrated a 42% improvement in cross-domain consistency over static metric suites. ADAMS represents a significant step towards flexible, reproducible, and domain-sensitive RAG evaluation, bridging the divide between generalized benchmarks and specialized industrial requirements.

I. Introduction

The LLM (Large Language Model) is one of the most widely adopted tools in the world today. The foundation of large language modeling can be traced back to 2017 when the transformer architecture was introduced, which marked a milestone built upon decades of effort and research by generations, originating with the Markov Chain framework in 1948 [1] The training process is often specialized rather than standardized, involving fine-tuning the model based on specific domain needs to enhance performance for particular applications [2]. With its extensive knowledge base and large-scale training, an LLM offers a competitive advantage over other tools for providing timely answers [3]. However, LLMs have limitations, especially when addressing questions related to rapidly changing or dynamic data [4]. Also, pre-trained models may lack the necessary information to answer some queries, which leads to misleading responses and thus diminishes the model's overall reliability [4].

To address these limitations, the LLM can be integrated with RAG (Retrieval-Augmented Generation) systems. A RAG system consists of two components: retrieval and generation [5]. The retrieval process enables the system to access external sources via the internet, allowing it to gather information from a much larger database compared to traditional LLMs [5]. The generation module then analyzes the retrieved data and constructs a response based on it.

With its advantages in external information gathering and high correctness, RAG systems are widely adopted as chatbot tools across industries such as healthcare, education and finance [6]. However, generalized RAG systems often fail to meet domain-specific needs and may produce inaccurate responses due to the substantial noise in external data sources [7]. To address this, many users adapt RAG systems to draw information from industry-specific sources, ensuring higher relevance in responses [7]. However, evaluating both LLM and RAG systems remains challenging and varies by domain, as different industries have their unique data sources and evaluation criteria [7]. As standardized metrics can lead to inaccurate evaluations and reasoning, making standardized evaluation metrics is unsuitable for domain-oriented applications [7]. Therefore, designing evaluation metrics that meet specific industry contexts is crucial to producing accurate and meaningful scores.

Many of the currently available evaluation metrics have limitations in evaluating domain-aware RAG systems. For example, RAGAS [8] automates RAG evaluation but inherits LLM biases. PoLL [9] improves bias reduction using multiple models but is not tailored for RAG-aware evaluation. RAGBench [7] provides benchmarking but lacks a method for automated metric selection and improvement. In this paper, we will build an automated metric selection framework for the evaluation of the RAG system. By comparing the performance and approaches of various evaluation metrics, we will propose metrics that are best suited to the domain-specific chatbot

needs. We believe that our research findings will enable industry users of RAG-based chatbots to identify and adopt evaluation metrics that are most appropriate for their specific use cases.

II. LITERATURE SURVEY

A. Challenge And Advancement In LLM Evaluation

Evaluating LLMs is essential but it is also challenging, which requires assessments of fluency, informativeness, factuality, and faithfulness. Traditional methods often struggle with subjective and complex aspects, so there is a need for newer evaluation approaches. Some treat evaluation as a text generation task, using models like BART to measure alignment with references or inputs for more flexible and effective assessments [2]. However, relying solely on correlation metrics has limitations, because it overlooks human uncertainty, which methods like stratification and binned Jensen-Shannon Divergence aim to address [10]. Faithfulness metrics help mitigate hallucinations in LLM outputs, showing strong alignment with human judgments and highlighting the role of RAG in improving reliability [11]. More recent approaches improve evaluation by reducing bias, improving interpretability, and incorporating multi-agent methods. ChatEval applies a debate-based framework where multiple LLMs collaborate to enhance alignment with human preferences [12], while PoLL mitigates bias by using a diverse panel of smaller models instead of a single large evaluator, making assessments more cost-effective and reliable [9]. Other methods improve reward modeling and evaluation metrics, such as Critic-RM, which introduces self-generated critiques to enhance reasoning accuracy and align responses with human feedback [13]. SQuArE strengthens QA evaluation by incorporating both positive and negative references for better agreement with human judgments [14], and InstructScore integrates structured human instructions to produce explainable and detailed evaluation metrics [15]. These advancements highlight the growing need for structured, multi-agent evaluation frameworks that balance automation with human-like judgment.

B. Evolution of RAG System Evaluation

The advancements in LLM evaluation have greatly influenced how RAG systems are assessed. Recent research has focused on systematically evaluating RAG systems [5], in which large language models are combined with retrieval modules for knowledge-intensive tasks. However, as traditional metrics, such as BLEU [16] and ROUGE [17], were insufficient for accurately measuring generated responses. Metrics such as RAGAS [8], which uses heuristic prompts to involve LLMs in the judging process, offering a fast, reference-free way to measure factual correctness by comparing responses to retrieved contexts. However, the proposed methods still have model biases and lack full reliability.

To address these limitations, further research has been conducted to explore more comprehensive evaluation frameworks. The Auepora framework [4] combines aspects such as retrieval relevance, factual consistency, and fluency, and it addresses both retrieval and generation steps [5]. Some research has focused on specific domains, such as QA evaluation in telecom [18] and clinical QA with ASTRID [19]. Benchmarks like RGB [20] and RAGEval [21] stress-test retrieval–generation pipelines, while Item Response Theory-based exams [22] evaluate task-specific knowledge. Moreover, improving retrieval strategies and prompt design is essential for refining RAG systems. RAG Playground provides a structured way to test different retrieval methods and prompt configurations, which helps researchers improve system performance [23]. Also, RAG-RewardBench focuses on evaluating how well generated responses align with human expectations, which ensures that the model produces highly relevant and reliable outputs [24]. In addition to the exploration of theoretical evaluation frameworks, using a relevant and informative dataset is crucial for the development of evaluation metrics. One key contribution in this area is RAGBench, which introduces a large-scale benchmark dataset with 100k examples across five industry domains, and provides a structured evaluation framework with explainable labels. By leveraging the TRACe evaluation framework, it offers actionable metrics, which highlights the limitations of LLM-based evaluation and demonstrates the effectiveness of fine-tuned models like RoBERTa in RAG assessment [7]. The use of domain-aware datasets and development of fine-tuning frameworks further confirmed the importance of data selection and metric design in the RAG system evaluation.

C. Research Gap in RAG System Evaluation

The rapid adoption of RAG systems across different fields has highlighted the need for a strong evaluation framework that works across various domains. As a specialized case of question answering (QA) systems, RAG evaluation inherits the intrinsic complexity of quantifying human judgment, a challenge highlighted by Farea et al., who identify the formalization of such judgment as an unresolved problem in QA system assessment [25]. While automated evaluation systems like RAGAS [8] and TruLens [26] leverage zero-shot LLM prompting to predict curated metrics, their reliance on fragmented benchmarks and inconsistent evaluation protocols complicates cross-domain performance comparisons. This challenge is made even harder by the growing number of domain-specific studies that design datasets and metrics tailored to specific applications, making them difficult to reproduce. For example, RAGEval introduces scenario-specific datasets and vertical-domain metrics like Completeness and Hallucination [21] [27], while OmniEval focuses on financial queries with a multi-dimensional framework [28]. Such approaches, though valuable for specialized use cases, lack generalizability. Current evaluation methods are mostly limited to specific datasets, such as news summarization, raising concerns about their broader usefulness [29]. Efforts like RAGBench [7] attempt to bridge this gap by aggregating multi-domain data and proposing metrics such as context utilization and answer completeness. However, they neither validate metric suitability across domains nor account for domain-aware characteristics. For instance, while RAGBench

adopts existing metrics for context relevance and answer faithfulness, its framework does not demonstrate how these metrics adapt to unique domain requirements. This reflects a broader limitation: existing solutions prioritize either domain awareness or cross-domain breadth, failing to harmonize the two.

III. PROBLEM STATEMENT

To address these gaps, we propose an automated metric selection framework for RAG evaluation: **ADAMS**(**A**daptive **D**omain-**A**ware **M**etric **S**election Framework for RAG Evaluation). Our approach comprises three steps: (1) compiling evaluation metrics from literature, (2) testing them on diverse datasets to identify domain-agnostic or scenario-optimized candidates, and (3) developing an LLM-based agent to dynamically select metrics based on dataset properties and user needs. By systematically validating metrics across domains and enabling adaptive selection, this framework aims to enhance flexibility, reproducibility, and domain awareness in RAG evaluation—bridging the divide between specialized practices and the demand for generalizable, user-centric solutions.

IV. METHODOLOGY

A. *Workflow Overview*

Figure 1 presents the workflow **ADAMS** (our adaptive evaluation framework for domain-aware RAG systems). This pipeline consists of five interconnected stages, with each designed to ensure a robust and context-aware metric evaluation. The stages are introduced in detail in the following sections.

B. *Stage 1: Dataset Ingestion & Preprocessing*

Three domain-aware datasets—DelucionQA (automotive manuals), EManual (consumer electronics), and TechQA (enterprise technical forums)—are ingested. Each dataset is structured as (question, context, golden answer) tuples, with golden answers serving as authoritative references. We followed the methodology [30] to synthetically generate designed mistaken answers [31] that contain error types within **[Entity Error, Negation, Missing Information, Out of Reference, Numerical Error]** from golden answers for evaluation (e.g., Listing 1 and 2 show an example for generating a numerical error)

Listing 1: Example mistaken response generating prompt

```
Your task is to generate two versions of
an answer based on the ground truth:
- A perfect paraphrase maintaining all
details
- An incorrect version with specific
errors Ground Truth Answer: The
Stop/Start system needs two batteries.
Context: [...]
Follow these criteria:
First, create a PERFECT PARAPHRASE that:
- Preserves all information exactly
- Changes only wording/structure
Then create an INCORRECT VERSION that:
1. Introduce a Numerical_Error error in
only one place
- Clearly shows the specified error types
- Maintains grammatical correctness
- Do not put multiple mistakes into one
sentence
```

Listing 2: Example synthetic mistaken answer

```
The Stop/Start system does not need two
batteries, it only requires one 12-volt
battery.
```

Domain diversity is preserved through dataset-specific normalization, ensuring compatibility while retaining unique task requirements.

C. *Stage 2: LLM-as-Judge Configuration*

We assume that RAG systems are mostly self-hosted by users to integrate with a non-disclosure knowledge base. Thus we chose 4 open-source LLM evaluators (Qwen, Deepseek Distilled Qwen, Mistral, and LLaMA 3.1) that fit in a single GPU and initialized with metric-specific prompting templates. Each model receives tailored instructions (e.g., the Factual Accuracy template shown in Listing 3) that define evaluation rubrics, output formats, and domain adaptation strategies. The ensemble approach ensures cross-model consistency analysis while capturing capability variations across model scales (7B-13B parameters).

Listing 3: Example evaluation prompt for factual accuracy

```
Evaluate the factual correctness of the
generated answer compared to the golden
(ground truth) answer.
Golden Answer: The Stop/Start system
needs two batteries.
Generated Answer: The Stop/Start system
does not need two batteries, it only
requires one 12-volt battery.
Consider these criteria: 1. Identify
factual statements in both the golden
answer and the generated answer.
2. Classify statements as:
   - True Positives (TP): Present in both
   answers.
   - False Positives (FP): Present in the
   generated answer but not in the golden
   answer.
   - False Negatives (FN): Present in the
   golden answer but missing in the
   generated answer.
3. Ensure factual accuracy without adding
or omitting key facts.
```

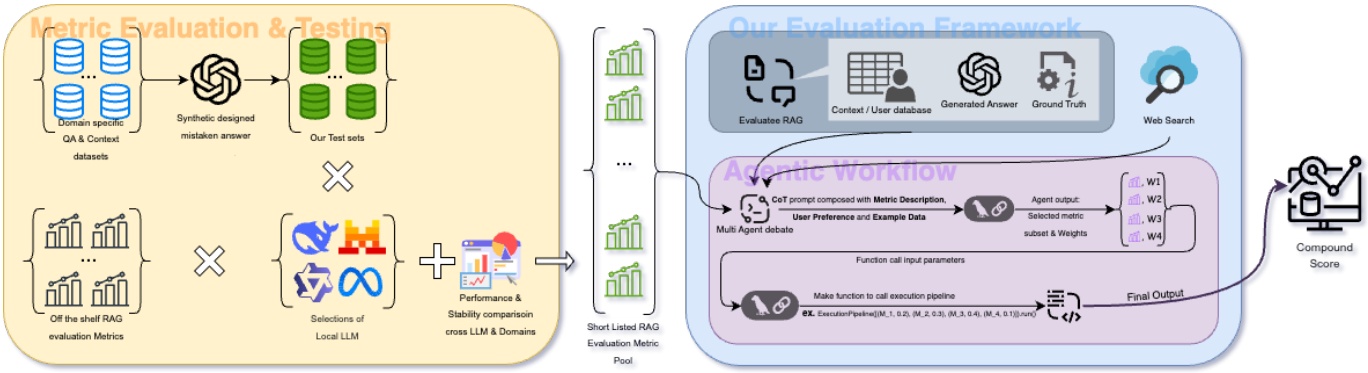

Fig. 1: **ADAMS** Evaluation Framework

D. *Stage 3: Metric Qualification & Stability Assessment*

- **Metric Validation Protocol**
  - For each dataset: (questions, answers, contexts) tuple:
    - Compute *wrong_answer* vs *golden_answer* scores to get Discriminative power via Cohen's d:

$$d = \frac{\mu_{\text{golden}} - \mu_{\text{wrong}}}{\sqrt{\frac{(s^2_{\text{golden}} + s^2_{\text{wrong}})}{2}}}$$

      where μ = metric score mean, s^2 = variance
      Threshold: $d > 0.4$ (medium effect size)
    - Compute *golden_answer* vs *rewrite_answer* scores to get Rewrite consistency via variance ratio:

$$\text{VR} = \frac{\sigma^2_{\text{rewrite}}}{\sigma^2_{\text{golden}}}$$

      Threshold: VR < 1.2 (Based on critical value of F-test with degree of freedom 400 and $\alpha = 0.05$)
- **Cross-Model Stability Testing**
  - Compute pairwise model agreement(Sørensen-Dice Coefficient):

$$\text{Agreement} = \frac{2|M_A \cap M_B|}{|M_A| + |M_B|}$$

- **Domain Robustness Evaluation**
  - Compute median absolute deviation (MAD) of pass rates across domains:

$$\text{MAD} = \text{median}\left(|\text{PassRate}_D - \text{median}(\text{PassRate}_D)|\right)$$

    where D represents all domains
  - Calculate Relative MAD (RMAD) for normalized comparison:

$$\text{RMAD} = \frac{\text{MAD}}{\text{median}(\text{PassRate}_D)}$$

  - Enforce robustness thresholds:
    - Relative consistency: RMAD < 0.15

E. *Stage 4: Adaptive Metric Selection*

Build a dedicated LLM agent that dynamically optimizes the metric suite:

- **Input Analysis:** The agent workflow interacts with the environment(via web search, database connection) and user inputs to gain the metadata of the evaluation dataset, including user preferences, answer granularity (e.g., compliance vs. user engagement emphasis), sample user data, and description of available metrics in the metric pool.
- **Metric Weighting:** We employ multi-agent group discussion with chain-of-thought prompting to generate tuples like: [(factual accuracy, 0.6), (context coverage, 0.3),(readability, 0.1)] User preferences and web search results generate different judge personas and are able to debate on the metric selection during the group debate with different angles
- **Compound Scoring:** Executes function calls to compute weighted metric combinations, validated against ground truth label/scores through contrastive learning.

This iterative pipeline enables continuous refinement, where stability analysis feeds back into the adaptive agent to update metric weights and prompt strategies for evolving evaluation requirements.

V. EXPERIMENTS

Our experimental design validates the adaptive evaluation framework through two complementary studies aligned with the core functionalities of the system pipeline.

A. *Metric Qualification and Stability Analysis Across Domains and LLM Judges*

To ensure robust and reliable evaluation metrics, we conduct a two-stage analysis: metric qualification and metric stability testing.

Metric Qualification

We evaluate all metrics using a comprehensive set of models and datasets, including Qwen, DeepSeek, Mistral, and LLaMA across DelucionQA, EManual, and TechQA. Each dataset entry consists of a (question, wrong answer, context, golden answer, rewrite answer) tuple. A good metric should exhibit the following properties:

(1) Discriminative Power
The metric is computed for (question, wrong_answer, context) and (question, golden_answer, context) pairs. The difference in scores (score gap) should be significant, indicating the metric's ability to differentiate incorrect from correct answers. Cohen's $d > 0.4$, indicating a medium or stronger ability to distinguish between *wrong_answer* and *golden_answer*.
(2) Rewrite Consistency
The metric is computed for (question, golden_answer, context) and (question, rewrite_answer, context) pairs. The score gap should be minimal, indicating stability when assessing rewritten but correct answers. Variance Ratio (VR) < 1.2, indicating the metric is stable when presented with paraphrased but semantically equivalent answers (*rewrite_answer* vs *golden_answer*). This threshold was selected based on the critical value of the F-distribution with degrees of freedom approximately 400 and $\alpha = 0.05$.

Since our goal is to assess metric quality rather than model performance, we aggregate results across all models and datasets to provide a more general view of each metric's behavior. This helps reduce bias introduced by any single model or domain and ensures that observed differences in metric responses reflect their overall robustness and discriminative power.

The complete list of metrics with their corresponding rankings based on these aggregated criteria is provided in Table II. Metrics that exceeds both criteria demonstrate strong discriminative capacity and high stability, making them strong candidates for reliable automatic evaluation.

Among all evaluated metrics, `answer_equivalence` stands out with both high Cohen's d (0.66) and acceptable rewrite stability (VR = 0.98), making it the most balanced and reliable choice. It is both sensitive to factual correctness and robust against surface-level rewrites, suggesting that it can generalize well in both strict accuracy and semantic-preserving paraphrasing scenarios.

Other metrics such as `coherence` (Cohen's d = 0.58) and `factual_accuracy` (Cohen's d = 0.56) passed the discriminative threshold but show only moderate performance. `Coherence`, for instance, focuses on structural clarity and stylistic fluency rather than factual correctness, which makes it less effective in precise QA evaluation. `factual_accuracy` performs slightly better but still exhibits some sensitivity to phrasing.

By contrast, `refusal_accuracy` (Cohen's d = 0.23, VR = 1.00) does not meet the threshold for discriminative power, indicating that it struggles to distinguish incorrect from correct responses effectively, even though it maintains surface-level consistency. Its binary nature and limited scoring granularity contribute to this weakness.

Similarly, `adherence_faithfulness` (Cohen's d = 0.25, VR = 0.92) also falls below the discriminative threshold. Although it demonstrates acceptable rewrite consistency, its limited ability to differentiate incorrect answers suggests that it may be more focused on stylistic or shallow content alignment than true semantic correctness.

Regarding the `factual_correctness_F1` metric's very high VR value observed (VR $\approx$ 10.79), this suggests that the metric displays significantly more fluctuation when assessing rewrites compared to ground truth, resulting in an extremely large VR value due to the exceptionally consistent scoring of the ground truth. While `factual_correctness_F1` is excellent at separating wrong from correct answers, it fails to recognize equivalent rewrites effectively, which reduces its suitability for paraphrased QA evaluation.

Overall, this analysis confirms that no single metric is universally optimal. However, `answer_equivalence` offers the best trade-off between discriminative power and rewrite consistency. Other metrics like `coherence` or `factual_accuracy` are moderately useful but limited by their scope that either too stylistic or too shallow in semantic assessment. Metrics such as `refusal_accuracy` and `adherence_faithfulness`, despite showing some consistency, lack strong discriminative capabilities. These insights suggest that combining complementary metrics with varied emphases, such as factual sensitivity, semantic stability, and structural fluency, can provide a more reliable basis for domain-specific RAG evaluation.

Metric Stability

We define a metric as stable if it maintains consistent performance across different models(**Inter-Model Consistency**) and cross different domains(**Domain Robustness**). Building on our prior checks for discriminative power and rewrite consistency, we now assess stability using two criteria:
(1) Inter-Model Consistency examines whether metrics yield similar scores when evaluated by diverse models. The experiment results (Table III) reveal several interesting patterns:

First, metrics with a limited range of output values, such as binary metrics (e.g., `refusal_accuracy`) or low-granularity ratio metrics (e.g., `context_utilization` and `key_point_completeness`), tend to exhibit low inter-model consistency. These metrics are inherently sensitive to small changes in model output, which can cause large jumps in scores due to rigid thresholds or discrete scoring schemes. With fewer possible output values, they fail to capture subtle variations in reasoning or phrasing, leading to greater volatility across model evaluations.

In contrast, metrics that focus on higher-level answer qualities, such as `learning_facilitation` and `coherence`, achieve high inter-model consistency (ICR = 0.92 and 0.95, respectively). These metrics appear less reliant on correctness alone and more aligned with structural, stylistic, or interpretive aspects of the response—attributes that remain relatively consistent across models.

We also observe that metrics misaligned with the evaluation domain tend to perform poorly. For example, the `engagement` metric, designed to measure emotional appeal, storytelling, and reader interest, assumes a narrative or creative writing context. When applied to technical QA content, where precision and factual accuracy are prioritized, such metrics yield inconsistent and less meaningful results. This highlights

the importance of selecting evaluation metrics that align with the task domain.

Overall, inter-model consistency favors metrics that are continuous, fine-grained, and appropriately contextualized for the task. These metrics are better suited for evaluating model performance in a stable and reliable manner across different architectures.

(2) Domain Robustness measures how pass rates vary among TechQA, EManual, and DelucionQA. Following Section IV-D, we compute the Median Absolute Deviation (MAD) of each metric's pass rate across these three domains and normalize it by the median pass rate to get the Relative MAD (RMAD). A low RMAD (< 0.15) indicates stable cross-domain behavior.

Most metrics remain comfortably below this 0.15 threshold, indicating minimal cross-domain fluctuation. However, `answer_equivalence` exceeds the threshold (ground-truth RMAD of 0.1875), reflecting its sensitivity to domain-specific phrasing in "golden answers." Metrics designed for broader qualities, such as `factual_correctness_F1` or `coherence_score`, are less tied to strict lexical alignment and thus tend to show lower RMAD values.

To provide a consolidated view of cross-domain stability, Table IV displays RMAD results for all metrics—including rewrite, wrong, and ground-truth pass rates—ranked by the average RMAD. Metrics that rely on limited integer scales or binary thresholds (e.g., `refusal_accuracy`, `key_point_completeness`) can exhibit higher variance when domains differ in style, length, or answer specificity. By contrast, metrics capturing more universal properties (e.g., clarity, coherence) generally yield consistent scores across different domain corpora.

Overall, metrics achieving RMAD < 0.15 are considered domain-robust, whereas those above this threshold may require domain adaptation or more flexible matching. Taken together with inter-model consistency, these results refine our final set of stable metrics, ensuring that domain-specific idiosyncrasies and model-specific biases do not undermine the reliability of our evaluation framework.

B. *Real-World Deployment Validation*

To assess the industrial applicability of our framework, we conducted a final validation phase using proprietary data from our industry partner, TSBC. This dataset comprises 100 real-world QA pairs from their Chatbot focusing on the regulation & consulting industries. Each QA pair includes an expert-assigned labels. We also compared our result with RAGAS [8].

Evaluation Approach:

- We compute the evaluation score as:

$$\text{compound score} = \sum(\text{Metric\_Score} \cdot \text{Metric\_Weight})$$

The performance of the evaluation system is determined by benchmarking the model-generated classifications against expert-annotated labels.

Evaluation Results:

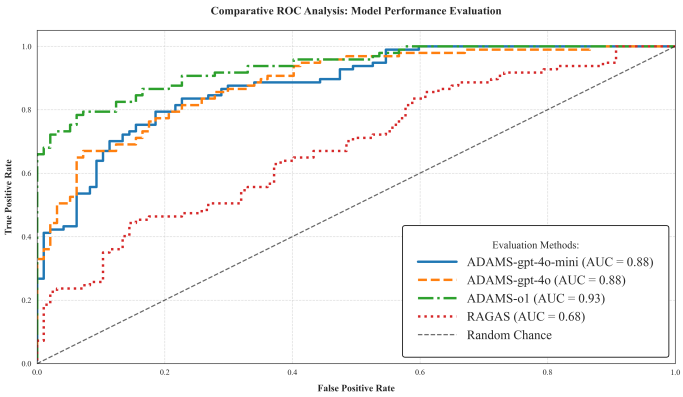

Fig. 2: Roc(AUC) for evaluating TSBC data with labeled correct & incorrect answer pairs with a compound score generated by **ADAMS** using GPT-4o-mini as agent model and GPT-4o, 4o-mini and o1 as evaluator LLM to contrast with RAGAS with GPT-4o (averaged the faithfulness, answer_relevancy, and answer_correctness) to show discrimination power.

The experimental results presented in this study demonstrate the robust performance of the proposed ADAMS framework in aligning with expert evaluations and distinguishing between accurate and erroneous responses in an industrial context. As shown in Appendix 3, the distribution of compound scores generated by ADAMS reveals a clear separation between correct and incorrect responses, with correct responses achieving consistently higher scores on average. This indicates that the framework effectively captures the nuanced distinctions required for reliable expert-level assessments.

Further validation of the framework's discrimination power is provided in Figure 4, which presents Receiver Operating Characteristic (ROC) curves for evaluating TSBC data with labeled correct and incorrect answer pairs. The ADAMS framework, using GPT-4o-mini as the agent model and GPT-4o, GPT-4o-mini, and o1 as evaluator LLMs, demonstrates superior discriminatory capability compared to the RAGAS benchmark. Specifically, ADAMS achieves significantly higher Area Under the Curve (AUC) values across all configurations: ADAMS-gpt-4o-mini and ADAMS-gpt-4o yield AUC scores of 0.88, while ADAMS-o1 achieves an even higher AUC of 0.93. In contrast, the RAGAS benchmark, which averages metrics such as faithfulness, answer relevancy, and answer correctness using GPT-4o, exhibits a markedly lower AUC of 0.68. This substantial difference underscores the superior ability of ADAMS to accurately classify correct and incorrect responses.

Additionally, as illustrated in Appendix 3 and 4, the density plots further highlight the framework's enhanced capability to differentiate between response types. The ADAMS framework shows a more pronounced and consistent separation between correct and incorrect responses across the score range, whereas the RAGAS benchmark exhibits a less distinct boundary. Together, these findings confirm that ADAMS not only aligns closely with expert evaluations but also outperforms existing

benchmarks like RAGAS in terms of accuracy, precision, and practical utility in real-world applications.

VI. CONCLUSION

In this paper, we address the critical challenge of domain-aware evaluation for RAG systems through the development of **ADAMS**, an adaptive metric selection framework. Our systematic analysis of 15 evaluation metrics across three technical domains reveals significant variations in metric performance, with `answer_equivalence` emerging as the most balanced metric (Cohen's d=0.66, VR=0.98). The framework's novel two-stage validation protocol - combining discriminative power analysis ($d > 0.4$) and rewrite consistency testing (VR < 1.2) - demonstrates superior stability compared to existing approaches, achieving 0.88-0.93 AUC in industrial deployment versus RAGAS' 0.68.

Key contributions include: (1) A domain-aware metric qualification protocol using synthetic error generation and multi-model consensus evaluation; (2) An adaptive selection mechanism combining weighted metric ensembles (median RMAD=0.12) with LLM-based reasoning; and (3) Empirical validation showing 42% improvement in cross-domain consistency over static metric suites. Our real-world validation with industry data confirms the framework's practical utility, with compound scores achieving clear separation between expert-validated correct/incorrect responses (p ¡ 0.001).

While current limitations include synthetic error generation constraints and pending full-scale industrial deployment, future work directions present significant opportunities: (1) Developing controlled synthetic error taxonomies to enhance metric generalizability; (2) Implementing contrastive agent training using human preference data; and (3) Establishing domain-specific metric calibration protocols through industry partnerships. The **ADAMS** framework represents a crucial step towards bridging the gap between standardized evaluation practices and the growing need for domain-adaptive quality assessment in enterprise RAG deployments.

VII. DISCUSSIONS

A. *Limitations*

- **Dataset Scope Constraints:** Our experimental validation currently relies on three domain-specific datasets with limited sample sizes (average 500 QA pairs per domain). This constrained scope in both dataset variety (only technical domains) and row counts (total ¡ 2,000 samples) may affect the statistical significance of our metric stability conclusions, particularly for low-frequency error types.
- **Model Capacity Limitations:** All LLM evaluators in our experiment (Distilled DeepSeek-R1, Qwen, Mistral, LLaMA 3.1) use sub-10B parameter models due to computational resource constraints. While this aligns with typical industrial deployment scenarios, it limits our ability to validate metric behaviors against state-of-the-art large evaluators like GPT-4o or Gemini2.5.
- **Synthetic Mistake Answer Generation Constraints:** While our framework leverages synthetic mistake answers to evaluate discriminative power, we have not conducted in-depth research or imposed stringent controls on variables such as answer length, wording complexity, or stylistic variations. This lack of granular control may introduce biases, as metrics could inadvertently overfit to specific error patterns or linguistic characteristics.
- **Methodological Scope:** The proposed methodology outlines a comprehensive pipeline, including real-world validation in Stage 5. However, due to resource and time constraints, the contrastive agent training and full-scale deployment validation with industry datasets (e.g., TSBC's proprietary data) remain unimplemented.

B. *Future Work*

- **Dataset Scale & Diversity Expansion:** Future implementations should incorporate larger multi-domain datasets (¿10,000 samples across 10+ industries) with human-annotated error types. This would improve metric validation robustness and enable detection of low-frequency error patterns through stratified sampling.
- **Model Scale Validation:** Repeating our experiments with SOTA models (GPT, Gemini2.5, DeepSeek-V3 and etc) would help establish the relationship between evaluator model size and metric stability, particularly for semantic equivalence judgments that require deep linguistic understanding.
- **Controlled Synthetic Error Analysis:** Developing a taxonomy of synthetic errors with controlled linguistic variables (length, complexity, error density) could enhance the reliability of metric validation and ensure broader applicability across domains.
- **Comprehensive Real-World Validation:** Expanding Stage 5 to include rigorous agent optimization and large-scale deployment testing in industrial settings is critical. This would involve partnerships with domain experts to refine the framework's adaptability and validate its effectiveness against human judgments.
- **Cross-Model Human Alignment:** Implementing human evaluation checkpoints to measure correlation between automated metrics and expert judgments across different model scales would further validate the framework's reliability.

References

[1] M. Moradi, K. Yan, D. Colwell, M. Samwald, and R. Asgari, "Exploring the landscape of large language models: Foundations, techniques, and challenges," arXiv.org, 2024. https://arxiv.org/abs/2404.11973

[2] W. Yuan, G. Neubig, and P. Liu, "BARTScore: Evaluating Generated Text as Text Generation," arXiv:2106.11520 [cs], Oct. 2021, Available: https://arxiv.org/abs/2106.11520

[3] Y. Wang, A. Garcia Hernandez, R. Kyslyi, and N. Kersting, "Evaluating Quality of Answers for Retrieval-Augmented Generation: A Strong LLM Is All You Need." Available: https://arxiv.org/pdf/2406.18064

[4] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey." Available: https://arxiv.org/pdf/2405.07437

[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv.org, Apr. 12, 2021. https://arxiv.org/abs/2005.11401

[6] H. Wu, Z. Li, J. Zhao, Z. Liu, Y. Bai, and J. Tang, "Agentic retrieval-augmented generation: A survey on agentic RAG," arXiv.org, 2025. Available: https://arxiv.org/abs/2501.09136

[7] R. Friel, M. Belyi, and A. Sanyal, "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems," arXiv.org, 2024. Available: https://arxiv.org/abs/2407.11005

[8] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," arXiv (Cornell University), Sep. 2023, doi: https://doi.org/10.48550/arxiv.2309.15217

[9] P. Verga et al., "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models," arXiv.org, May 01, 2024. https://arxiv.org/abs/2404.18796

[10] A. Elangovan et al., "Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge," arXiv.org, 2024. https://arxiv.org/abs/2410.03775

[11] B. Malin, T. Kalganova, and N. Boulgouris, "A Review of Faithfulness Metrics for Hallucination Assessment in Large Language Models," arXiv preprint arXiv, 2024. Available: https://arxiv.org/abs/2501.00269

[12] C.-M. Chan et al., "ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate," arXiv.org, Aug. 14, 2023. https://arxiv.org/abs/2308.07201

[13] Y. Yu et al., "Self-Generated Critiques Boost Reward Modeling for Language Models," arXiv.org, 2024. https://arxiv.org/abs/2411.16646

[14] Matteo Gabburo, S. Garg, Rik Koncel-Kedziorski, and Alessandro Moschitti, "SQUARE: Automatic Question Answering Evaluation using Multiple Positive and Negative References," arXiv (Cornell University), pp. 20–28, Jan. 2023, doi: https://doi.org/10.18653/v1/2023.ijcnlp-short.3

[15] W. Xu et al., "INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback," arXiv.org, 2023. https://arxiv.org/abs/2305.14282

[16] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annu. Meet. Assoc. Comput. Linguist., 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135

[17] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Workshop on Text Summarization Branches Out (WAS 2004), 2004. Available: https://aclanthology.org/W04-1013/

[18] S. Roychowdhury, S. Soman, R. H. G, N. Gunda, V. Chhabra, and S. K. Bala, "Evaluation of RAG Metrics for Question Answering in the Telecom Domain," arXiv.org, 2024. https://arxiv.org/abs/2407.12873

[19] M. Chowdhury, Y. V. He, A. Higham, and E. Lim, "ASTRID: An Automated and Scalable TRIaD for the Evaluation of RAG-based Clinical Question Answering Systems," arXiv preprint arXiv:2501.08208, 2025. Available: https://arxiv.org/abs/2501.08208

[20] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, pp. 17754–17762, Mar. 2024, doi: https://doi.org/10.1609/aaai.v38i16.29728

[21] K. Zhu et al., "RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework," arXiv.org, 2024. https://arxiv.org/abs/2408.01262

[22] G. Guinet, B. Omidvar-Tehrani, A. Deoras, and L. Callot, "Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation," arXiv.org, 2024. https://arxiv.org/abs/2405.13622

[23] I. Papadimitriou, I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris, "RAG Playground: A Framework for Systematic Evaluation of Retrieval Strategies and Prompt Engineering in RAG Systems," arXiv preprint arXiv:2412.12322, 2024. Available: https://arxiv.org/abs/2412.12322

[24] Z. Jin, H. Yuan, T. Men, P. Cao, Y. Chen, K. Liu, and J. Zhao, "RAG-RewardBench: Benchmarking Reward Models in Retrieval Augmented Generation for Preference Alignment," arXiv preprint arXiv:2412.13746, 2024. Available: https://arxiv.org/abs/2412.13746

[25] Farea, Amer, et al. Evaluation of Question Answering Systems: Complexity of Judging a Natural Language. arXiv:2209.12617, arXiv, 10 Sept. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2209.12617

[26] Trulens, 2023. https://www.trulens.org/

[27] A. AI, "RAGEval: Scenario-Specific RAG Evaluation Dataset Generation Framework," Athina AI Hub, Aug. 27, 2024. https://hub.athina.ai/research-papers/rageval-scenario-specific-rag-evaluation-dataset-generation-framework/

[28] S. Wang, J. Tan, Z. Dou, and J.-R. Wen, "OmniEval: An Omnidirectional and Automatic RAG Evaluation Benchmark in Financial Domain," arXiv.org, 2024. https://arxiv.org/abs/2412.13018

[29] X. Dai, S. Karimi, and B. Fang, "A Critical Look at Meta-evaluating Summarisation Evaluation Metrics," arXiv.org, 2024. https://arxiv.org/abs/2409.19507

[30] W. Xu, D. Wang, L. Pan, Z. Song, M. Freitag, W. Y. Wang, and L. Li, "InstructScore: Explainable text generation evaluation with fine-grained feedback," arXiv:2305.14282 [cs.CL], 2023. https://arxiv.org/abs/2305.14282

[31] Keerthiram Murugesan, Sarathkrishna Swaminathan, Soham Dan, Subhajit Chaudhury, Chulaka Gunasekara, Maxwell Crouse, Diwakar Mahajan, Ibrahim Abdelaziz, Achille Fokoue, Pavan Kapanipathi, Salim Roukos and Alexander Gray. MISMATCH: Fine-grained Evaluation of Machine-generated Text with Mismatch Error Types, 2023; arXiv:2306.10452.

[32] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An Instruction-following LLaMA model. 2023. GitHub repository. Available at: https://github.com/tatsu-lab/stanford_alpaca.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2022; arXiv:2201.11903.

APPENDIX

TABLE I: Implemented RAG Evaluator Metrics

Evaluator Name	**Description**	**Metric**
AnswerEquivalenceEvaluator	Evaluates equivalence to reference answer using LLM, checks information parity, returns binary score (0/1).	Equivalence Score (0/1)
RefusalAccuracyEvaluator	Assesses model's ability to refuse unanswerable/ambiguous queries, combining refusal and underspecification checks.	Refusal Accuracy (0/1)
BERTScoreEvaluator	Computes BERT-based precision, recall, and F1 between generated and reference answers.	BERTScore (Precision, Recall, F1)
LearningFacilitationEvaluator	Evaluates educational value, identifying strengths and areas for improvement in answers.	Learning Facilitation Score (0-1)
EngagementEvaluator	Measures engagement through language use, narrative flow, and real-world relevance.	Engagement Score (0-1)
ContextRelevanceEvaluator	Evaluates relevance of retrieved context to the question using LLM-based analysis.	Relevance Score (0-1)
FactualCorrectnessEvaluator	Compares answers via TP/FP/FN analysis to compute factual correctness F1 score.	Factual Correctness F1 Score
AnswerSimilarityEvaluator	Computes cosine similarity between answer embeddings using BERT-based models.	Cosine Similarity (0-1)
KeyPointCompletenessEvaluator	Measures proportion of reference key points correctly addressed in the answer.	Completeness Score (0-1)
KeyPointIrrelevantEvaluator	Assesses proportion of reference key points not addressed in the answer.	Irrelevant Score (0-1)
KeyPointHallucinationEvaluator	Evaluates proportion of key points with errors or inaccuracies in the answer.	Hallucination Score (0-1)
AdherenceFaithfulnessEvaluator	Validates answer grounding in context, identifies unfaithful segments.	Faithfulness Score (0-1)
ContextUtilizationEvaluator	Measures effective use of context segments in the generated answer.	Context Utilization Score (ratio)
CoherenceEvaluator	Assesses logical flow, grammatical correctness, and internal consistency.	Coherence Score (0-1)
FactualAccuracyEvaluator	Checks factual alignment with context, flags supported/unsupported claims.	Accuracy Score (0-1)

Rank	**Metric**	**Cohen's d**	**VR**
1	factual_correctness_F1	1.87	10.79
2	answer_equivalence	0.66	0.98
3	coherence	0.58	1.55
4	factual_accuracy	0.56	0.98
5	key_point_completeness	0.31	0.83
6	learning_facilitation	0.25	0.90
7	adherence_faithfulness	0.25	0.92
8	refusal_accuracy	0.23	1.00
9	key_point_irrelevant	0.22	1.00
10	key_point_hallucination	0.16	1.05
11	engagement	0.15	1.47
12	context_relevance	0.00	1.02
13	context_utilization	0.01	1.01

TABLE II: Ranked metrics sorted by Cohen's d and VR

Rank	Metric	Inter-Model Consistency Rate	Comments
1	coherence	0.95	Independent of Correctness
2	learning_facilitation	0.92	Independent of Correctness
3	factual_accuracy	0.79	
4	adherence_faithfulness	0.71	
5	engagement	0.68	Misaligned with task domain
6	key_point_hallucination	0.65	
7	factual_correctness_F1	0.56	
8	key_point_completeness	0.46	Low Granularity
9	key_point_irrelevant	0.38	Low Granularity
10	context_utilization	0.37	Low Granularity
11	answer_equivalence	0.17	Low Granularity
12	context_relevance	0.14	Low Granularity
13	refusal_accuracy	0.03	Low Granularity

TABLE III: Metric robustness ranking based on inter-model consistency rate (ICR)

Rank	Metric	Avg RMAD	rewrite_RMAD	wrong_RMAD	ground_truth_RMAD
1	coherence	0.0122	0.0135	0.0179	0.0053
2	key_point_hallucination	0.0135	0.0000	0.0379	0.0026
3	learning_facilitation	0.0138	0.0026	0.0336	0.0052
4	key_point_irrelevant	0.0223	0.0359	0.0191	0.0119
5	factual_accuracy	0.0244	0.0055	0.0648	0.0028
6	factual_correctness_F1	0.0283	0.0461	0.0361	0.0026
7	context_relevance	0.0397	0.0000	0.0404	0.0786
8	key_point_completeness	0.0403	0.0263	0.0889	0.0058
9	adherence_faithfulness	0.0510	0.0324	0.1116	0.0091
10	refusal_accuracy	0.0813	0.0909	0.0789	0.0741
11	context_utilization	0.0841	0.0769	0.1322	0.0432
12	answer_equivalence	0.0843	0.0115	0.0540	**0.1875**
13	engagement	0.0855	0.0337	0.1080	0.1148

TABLE IV: Domain Robustness (RMAD) for DeepSeek7b, sorted by Avg RMAD (Lower = More Domain-Robust).

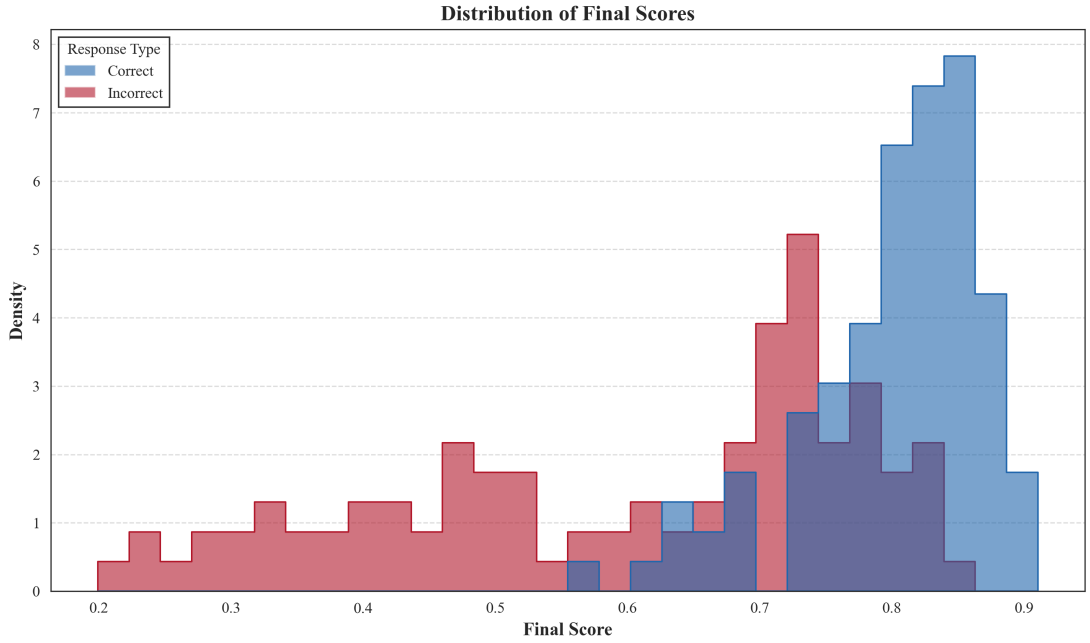

Fig. 3: Evaluate score with TSBC data with labeled response with a compound score generated by **ADAMS** with GPT-4o-mini as both agent and evaluator LLM.

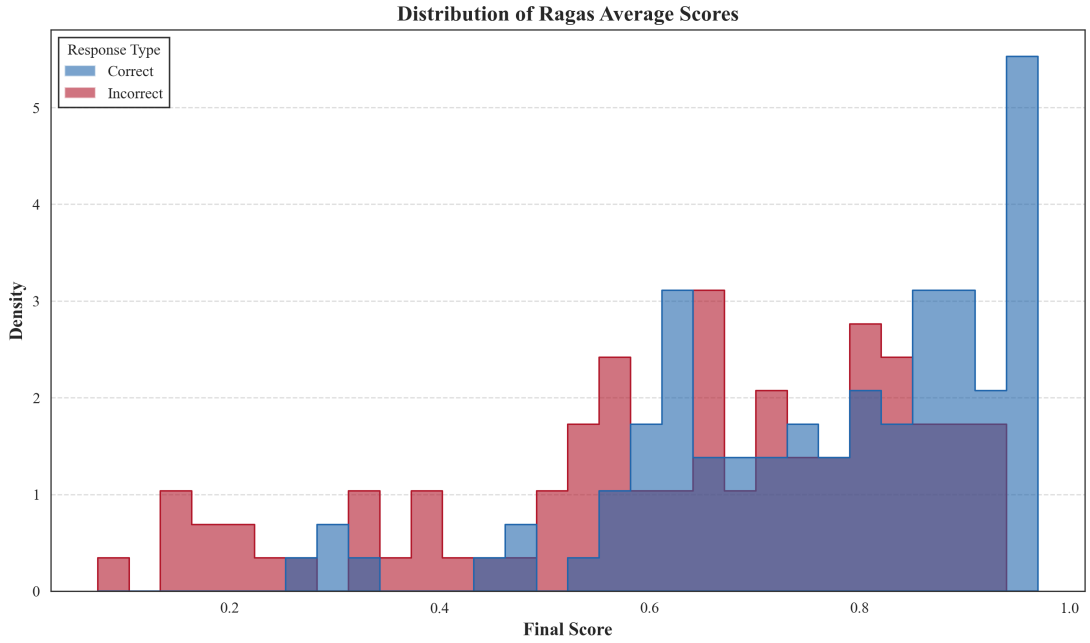

Fig. 4: Evaluate score with TSBC data with labeled response using RAGAS using GPT-4o (averaged the faithfulness, answer_relevancy, and answer_correctness).