Northeastern University
Khoury College of Computer Sciences

TECHNICAL SAFETY BC

Bridging the Gap in Domain-Specific RAG Evaluation: Adaptive Domain-Aware Metric Selection Framework

Kai Li, Muzhou Liu, Yao Cheng, Xueyun Li
Khoury College of Computer Science, Northeastern University

Instructor: Dr. Aanchan Mohan

1. Background

- **Rise of RAG:** Retrieval-Augmented Generation (RAG) combines an external retrieval module with a generative model to address knowledge-intensive tasks.
- **Evaluating RAG systems remains a challenge:** Standard metrics fail to capture domain-specific requirements, leading to inaccurate assessments in specialized fields such as regulation, healthcare, or finance.

2. Motivations

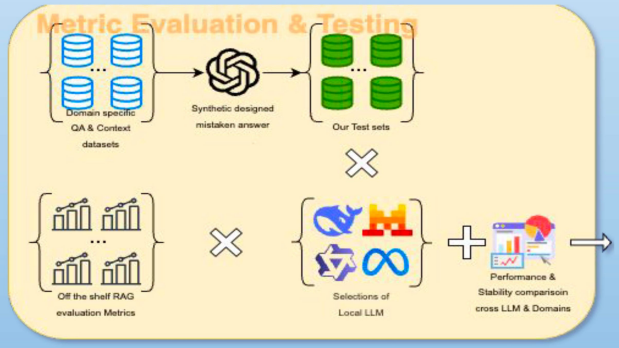
- Domain-oriented RAG chatbots need tailored evaluation criteria (not just generic faithfulness or fluency).
- Existing automated metrics (e.g., RAGAS) struggle with domain nuance and might inherit LLM biases.
- Lack of a flexible, automated approach that selects the "right" metrics depending on domain characteristics.

3. Objectives

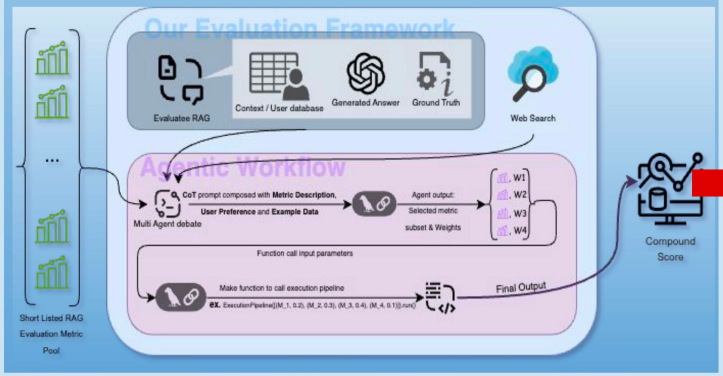
- Propose an Adaptive Domain-Aware Metric Selection Framework (ADAMS) that dynamically chooses optimal metrics for domain-specific RAG assessment.

4. Methodology

- **Dataset & Annotation:** We used 3 datasets and generated synthetic mistakes to test metrics' ability to distinguish errors.
- **Metric qualification:** We use 4 LLMs as judge LLM to score metrics consistently across domains to verify their performance.

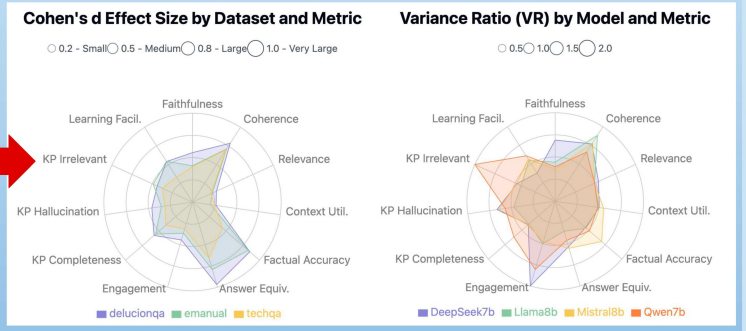


- **Adaptive Metric Selection:** Multi-agent debate for metric and weights based on user inputs to generate a compound score.

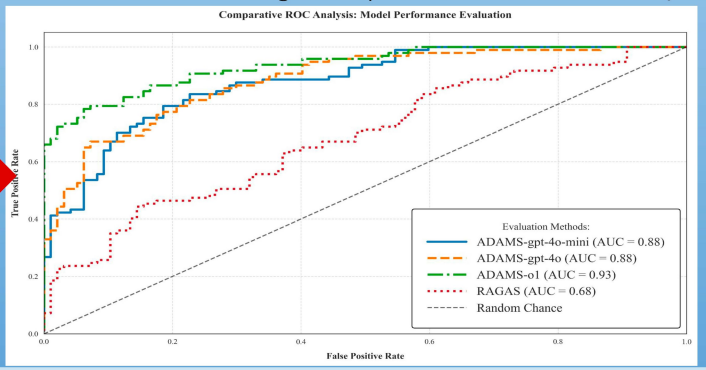


5. Experimental Results

- **Experiment on the metrics:**
  - *Discriminative Power*: A good metric should show a clear score gap(Cohen's d) between correct and incorrect answers.
  - *Rewrite Consistency*: The metric should remain stable (Variance Ratio) when evaluating semantically equivalent rewrites.



- **Performance comparison with RAGAS:** ADAMS outperforms RAGAS in discriminating correct/incorrect answers in TSBC QA.



6. Conclusion & Future Work

- ADAMS outperforms one-size-fits-all approaches like RAGAS, achieved 42% improvement in cross-domain consistency.
- ADAMS enabled domain-aware, user-aligned evaluation with customizable metric weights.
- Next steps include full-scale deployment with human-aligned agent training and expansion to more domains.