Methods and Instruments for Determining the Reliability for LLM and RAG Systems in Renewable Energy Domain

Vjeran Strahonja, Dijana Oreški
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
vjeran.strahonja@foi.hr, dijana.oreski@foi.hr

This paper explores the evaluation of large language models (LLMs) and retrieval-augmented generation (RAG) systems within the renewable energy domain, emphasizing the need for reliable advisory tools in high-stakes technical applications. We review existing evaluation methods, identifying gaps in answer quality, source grounding, and system stability. To address these, we propose a three-dimensional evaluation framework integrating correctness, attribution accuracy, and robustness, tailored for renewable energy's technical complexity and regulatory demands. This framework supports users without deep AI expertise and fits into RAG chatbot development and operational cycles. Drawing from a literature review, we synthesize complementary evaluation approaches and highlight domain-specific adaptations, including handling specialized terminology, multi-source integration, and evolving standards. Our study underlines the importance of combining automated and human-centered evaluation to ensure trustworthy deployment of LLM/RAG systems, bridging current methodological gaps and fostering safer, more transparent AI applications in renewable energy.

DOI
https://doi.org/10.18690/um.fov.3.2026.59

ISBN
978-961-299-124-1

Keywords:
large language models, retrieval-augmented generation, renewable energy, artificial intelligence, RAG evaluation

University of Maribor Press

1 Introduction

The widespread adoption of artificial intelligence has led to the growing use of AI-based advisory systems in domains where precision, transparency, and regulatory compliance are critical. Such systems, commonly implemented as conversational interfaces (chatbots), combine global data sources with domain- and region-specific knowledge. One prominent high-stakes application area is renewable energy and energy efficiency, which is the focus of this paper. These systems typically support problem diagnosis and solution formulation, while final decisions remain with human users. In addition, Retrieval-Augmented Generation (RAG) systems assist professionals in tasks such as technical documentation analysis, regulatory compliance, system design, and troubleshooting, where domain-specific terminology, evolving standards, and multi-source information present significant challenges. Such systems need robust reliability assessment methods, particularly in applications where incorrect information may lead to serious operational, financial, or security consequences. Existing literature shows that current LLM and RAG evaluation approaches remain fragmented and lack comprehensive frameworks that jointly address answer quality, grounding and source attribution, domain-specific evaluation benchmarks, and system robustness, especially in renewable energy applications. Key challenges identified in the literature include trustworthiness and grounding issues (Song et.al, 2025), (Kenthapadi et.al., 2024), limitations of general LLM evaluation and hallucination detection (Chang et.al., 2023), (Niu et.al. 2024), challenges of RAG deployment in high-stakes domains (Tu et.al., 2024), (Prabhune and Berndt, 2024), RAG-specific evaluation gaps [24], unmet domain-specific evaluation needs (Tu et.al., 2024), and ethical considerations (Jiao et.al., 2025).

This paper addresses these gaps by synthesizing existing LLM and RAG evaluation methods into a unified three-dimensional reliability assessment framework tailored to renewable energy systems. The objectives are to: (1) systematically review and categorize LLM/RAG evaluation methods, (2) develop a three-dimensional evaluation approach integrating answer quality, grounding, and stability, and (3) propose domain-specific adaptations suitable for renewable energy applications. The proposed approach is designed to be easy to apply by non-AI experts, usable throughout the RAG-based chatbot development and deployment lifecycle. Existing work is organized into six evaluation categories: (1) human evaluation, (2) LLM reliability evaluation methods, (3) RAG-specific evaluation methods, (4) multi-

dimensional evaluation frameworks, (5) AI–human and AI–AI comparison methods, and (6) risk management and governance frameworks. To ensure practical applicability in high-stakes technical domains, these approaches are consolidated into a three-dimensional verification framework addressing (1) answer quality, (2) grounding and source attribution, and (3) stability and robustness.

2 Literature Review

Research on LLM/RAG evaluation methods is a very dynamic field, but relatively new, so many research papers are in the pre-print stage. However, there are also comprehensive surveys (Chang et.al., 2024) and evaluation framework taxonomies (Yu et.al., 2024). Based on preliminary research, LLM/RAG evaluation methods are organized into six complementary categories, each addressing specific aspects of system reliability. This is elaborated further on in this chapter.

2.1 Human Evaluation

Human evaluation remains the gold standard for assessing LLM outputs, requiring careful calibration and inter-rater agreement protocols to ensure reliability. The basis of this approach are inter-rater validation methods, agreement metrics and consistency measures. Research emphasize the importance of human annotation quality in evaluation datasets (Niu et. al., 2024) and human-annotated databases for model validation, for example 5,000-sample validation in Human-Calibrated Automated Testing (HCAT) is a two-stage calibration approach aligning machine evaluations with human (Sudjianto et.al., 2024). Probability calibration and conformal prediction techniques are applied. This method embeds metrics for explainable assessment and stratified sampling for automated test generation.

2.2 LLM Reliability Evaluation Methods

Systematic reliability testing across stability, consistency, hallucination and robustness is essential for ensuring widespread LLM use, and especially trustworthy deployment in high-stakes applications. It is conducted against adversarial, out-of-distribution, and varied inputs (Tu et.al., 2024). It confirms output consistency and stability across similar queries and identifies marginal, entropy-based and bivariate weaknesses (Joshi, 2025). ***Hallucination detection*** (self-consistency checks) is well

described in the literature, as is the phenomenon of LLM hallucinations. There are comprehensive surveys on their detection, explanation, and mitigation approaches. Methods like SelfCheckGPT use sampling-based consistency across multiple stochastic generations to detect hallucinations at sentence and passage levels without external knowledge sources (Manakul et.al., 2023). ***Stability testing*** is the examination of the performance of some tasks, with the aim of determining whether there is an absence of response stability across repeated runs, systematic yes-response biases, changes in conclusions, changes in numerical values, logical reversals or other inconsistent answers with identical inputs.***Calibration assessment*** (knowing when it doesn't know) is the evaluation of how well an LLM's stated confidence (explicit or implicit) aligns with its actual correctness. It is carried out by using probability calibration techniques (Sudjianto et.al., 2024), comparing predicted confidence scores or probabilities against empirical accuracy using metrics such as Expected Calibration Error (ECE). Calibration metrics are part of the HELM framework (Bommasani, 2023).

2.3 RAG-Specific Evaluation Methods

RAG systems imply the reliability of the components they rely on, but require specialized evaluation approaches that assess retrieval quality and generation fidelity.

Retriever evaluation implies the application of typical metrics: (i) Retrieval relevance and quality metrics (Es et.al.,2024) - Did you miss relevant documents? (ii) Context precision / relevance metrics – How much noise was retrieved? (iii) Redundancy and coverage assessment (Birur et.al., 2024) (iv) Multi-step retrieval pipeline effectiveness (Krishna et.al., 2025) ***Generator evaluation*** contains evaluation of: (i) Faithfulness / roundedness' to retrieved context (Es et.al.,2024) - Are claims supported by retrieved text? (ii) Answer precision and accuracy (Birur et.al., 2024) (iii) Answer completeness and hallucination metrics (Zhu et. al.,2025). (iv) Answer relevance detection (Zhu et. al.,2025) - Does it actually answer the question? If RAG (and this also applies to LLM) provides citations, ***citation-level evaluation*** comprises (Song et.al., 2024): (i) Citation precision – does the quoted passage truly support the claim? (ii) Citation recall – are all major claims backed by citations? (iii) Attribution granularity – paragraph-level vs whole-document citation. ***Trustworthiness and attribution*** evaluation refer to: *(i)* Trust-Score for RAG

trustworthiness (Song et.al., 2024). (ii) Grounded attributions and learning to refuse. (iii) Citation accuracy and source verification.
The literature also describes end-to-end RAG Assessment using comprehensive ranking scores (Birur et.al., 2024) and factuality, retrieval, and reasoning measurement (Krishna et.al., 2025) and unified evaluation frameworks (FRAMES, VERA) (Krishna et.al., 2025), (Birur et.al., 2024).

2.4 Multi-Dimensional Evaluation Frameworks

Most of the commonly used benchmarks are incomplete, narrow and non-transparent. For example, SuperGLUE focuses on natural language understanding tasks, MMLU (Massive Multitask Language Understanding) tests understanding across academic subjects, and HumanEval's ability to generate correct Python code functions. Therefore, comprehensive multi-dimensional evaluation and benchmarking is necessary, by integrating quantitative and domain-specific metrics across multiple scenarios (Joshi, 2025). Such a multi-dimensional evaluation framework is HELM (Holistic Evaluation of Language Models), developed by the Stanford Center for Research on Foundation Models (CRFM). Some of the advantages of HELM are: (i) deeply evaluates each model across seven distinct dimensions (accuracy, calibration, robustness, fairness, bias, toxicity, efficiency) and provides a multifaceted profile, instead of a shallow one-dimensional answer as to whether the answer is "correct" (Bommasani, 2023). (ii) the evaluation spans 42 real-world use-case scenarios, such as question answering, multilingual dialogue, summarization, information extraction and classification, etc. which gives it diversity. (iii) enables trade-off analysis between different performance dimensions. (iv) is designed as an open, extensible ecosystem that encourages contributions from the research and industry. (v) strives for transparency and reproducibility through public release of prompts, datasets, and model access methods. There are other multi-dimensional benchmark frameworks with the focus on LLMs, e.g. COMPL-AI (Guldimann, 2024), intended for technical interpretation of the EU AI Act.

2.5 AI-Human and AI-AI Comparison Methods

LLMs as smart assessment judges bring a revolutionary change compared to solo human evaluators and their scalability limits, time and budget constraints. LLM-as-a-judge validation is an evaluation approach where an LLM is used to assess, score,

or compare the outputs of another model (or multiple models) against predefined criteria (Joshi, 2025). These methods are based on automated, scalable comparison, with human oversight and maintaining alignment with human preferences. LLM-as-judge architectures exhibit excellent capability to judge multiple quality dimensions in parallel, such as semantic and contextual correctness, conversational quality and fulfillment of user intent. Accuracy and reliability are increased through feedback loops. Today, the best practice is for prompt design in evaluation, used for validation against human judgments (e.g., 91% alignment in RAG-Check) (Mortaheb, 2025).

2.6 Risk Management and Governance Frameworks

Beyond the narrow evaluation of LLM/RAG reliability, comprehensive risk management aligned with regulatory requirements and effective governance are essential. Applying the NIST AI Risk Management Framework to LLM/RAG systems enables a structured assessment of risks across the AI lifecycle, including data quality, model behavior, security, and governance, by mapping system-specific hazards to the framework's functions for identifying, measuring, managing, and governing AI risks (Kenthapadi et.al, 2024). Similarly, applying the COMPL-AI technical interpretation of the EU AI Act allows high-level regulatory obligations to be translated into concrete, and measurable technical requirements for systematic risk evaluation and system validation (Guldimann, 2024). LLM/RAG evaluation poses distinct ethical challenges (Jiao et.al.,2025) as harms may arise from the interaction between opaque model reasoning and dynamically retrieved external content. These include bias amplification, misinformation propagation, and overreliance on unverified outputs. Such risks can be mitigated through curated and auditable retrieval sources, human-in-the-loop review for high-risk scenarios, explicit citation signals, and continuous monitoring with feedback loops. Although not part of evaluation frameworks in the strict sense, governance models for LLM/RAG deployment and use (Prabhune and Berndt, 2024) must clearly define accountability, decision rights, risk ownership, and escalation mechanisms, while ensuring regulatory compliance and secure technical architectures. In practice, these challenges can be addressed by establishing centralized AI governance structures, embedding human review into workflows, and operationalizing governance through logging, audits, training, and continuous compliance monitoring.

3 Three-Dimensional Evaluation Approach

The evaluation methods reviewed provide solid coverage of individual aspects; however, comprehensive surveys have identified notable limitations (Chang et.al, 2024), (Yu et.al., 2024), underscored domain-specific evaluation requirements (Tu et.al., 2024), and highlighted challenges related to practical deployment (Song et.al, 2024). These findings suggest significant opportunities for advancement, particularly in the development and application of integrated evaluation frameworks tailored to domain-specific contexts, alongside the provision of practical implementation guidance. In the following, a three-dimensional evaluation approach is proposed that simultaneously addresses quality, grounding/source attribution, and stability/robustness to ensure reliable LLM/RAG system deployment.

The main objectives of the proposed three-dimensional approach are: simplicity and comprehensibility and interdependencies between dimensions, comprehensive coverage of reliability concerns, alignment with existing evaluation methods and multi-method strategy, open to accepting new methods.

3.1 Definition of dimensions

The Answer Quality dimension is primarily focused on the evaluation of correctness, relevance, completeness, and coherence of generated responses. Human evaluation provides the gold standard for quality assessment (Tu et.al, 2024), (Joshi, 2025), relying on relevant metrics, such as RAGEval completeness and irrelevance metrics (Zhu et.al., 2025). Evaluation components are presented and explained in Table 1.

Dimension 2 (Grounding and Source Attribution) evaluates if generated content is properly grounded in authoritative sources with accurate attribution. Evaluation Components are explained in Table 2.

Dimension 3: Stability and Robustness is checked by evaluating system consistency, and resistance to adversarial or out-of-distribution inputs (Table 3).

Table 2: Grounding and Source Attribution dimension definition

Dimension	Metrics	Methods	Key Literature
Retrieval Quality (Effectiveness of source identification and ranking)	Retrieval precision, recall, relevance scores	Retrieval component evaluation (Es et.al., 2024), multi-step pipeline assessment (Krishna et.al., 2025)	RAGAS retrieval metrics (Es et.al., 2024); FRAMES retrieval evaluation (Krishna et.al., 2025)
Attribution Accuracy (Correctness of source citations and references)	Citation accuracy, attribution precision	Citation-level evaluation (Song et.al., 2024), grounded attribution assessment	Trust-Score framework for grounded attributions (Song et.al., 2024); RAGTruth corpus (Niu et.al., 2024)
Faithfulness (Alignment of generated content with retrieved sources)	Faithfulness scores, entailment measures	Automated faithfulness checking (Es et.al., 2024), human verification	RAGAS faithfulness metrics (Es et.al., 2024); VERA context adherence (Birur et.al., 2024)
Source Transparency (Clarity and accessibility of source information)	Source disclosure rates, transparency scores	Source tracking and verification (Birur et.al., 2024)	VERA validation framework for retrieval systems (Birur et.al., 2024)

Table 3: Stability and Robustness dimension definition

Dimension	Metrics	Methods	Key Literature
Consistency (Stability of outputs across similar queries)	Output variance, consistency scores	Repeated query testing, marginal analysis (Sudjianto et.al., 2024)	HCAT robustness testing (Sudjianto et.al., 2024); entropy-based stability (Joshi, 2025)
Calibration (Alignment of confidence estimates with actual accuracy)	Calibration error, confidence–accuracy correlation	Probability calibration (Sudjianto et.al., 2024), conformal prediction (Sudjianto et.al., 2024)	HELM calibration and robustness [2]; HCAT calibration approach (Sudjianto et.al., 2024)
Adversarial Resistance (Performance under challenging or adversarial conditions)	Adversarial robustness scores, failure rates	Adversarial testing (Sudjianto et.al., 2024), out-of-distribution evaluation	HCAT robustness testing (Sudjianto et.al., 2024); S-Eval safety evaluation (Yuan et.al., 2024)
Error Detection (Ability to identify and refuse inappropriate queries)	Refusal rates, error detection accuracy	Learning to refuse (Song et.al., 2024), uncertainty quantification	Trust-Align method (Song et.al., 2024); S-Eval safety evaluation frameworks (Yuan et.al., 2024)

3.2 Method Mapping Matrix

The three-Dimensional Approach integrates methods from all six evaluation categories to provide comprehensive reliability assessment. The mapping of evaluation methods to three dimensions is described in Table 4.

Effective implementation of the Three-Dimensional Approach and related evaluation methods requires a practical, scalable methodological framework that includes an evaluation workflow design, a description and methodological instructions for each activity, and expected inputs and deliverables. It is important to balance rigor and comprehensiveness with applicability and resource constraints.

Evaluation workflow design also includes defining criteria for different application scenarios, for example: continuous vs. periodic evaluation, sequential vs. parallel evaluation, prioritization based on application requirements.

3.3 Domain-Specific Adaptation for Renewable Energy

Renewable energy applications require specialized adaptations of the three-Dimensional Approach to address unique domain characteristics, such as:

- Technical complexity and specialized terminology
- Regulatory compliance requirements (safety, environment, grid standards)
- Multi-source information integration (e.g. research, standards,..)
- Evolving knowledge base (new technologies, updated regulations)
- High-stakes decision contexts (safety, financial, operational).

Preliminary requirements for adaptation to renewable energy are in Table 4.

Table 4: The mapping of evaluation methods to three dimensions

Evaluation Method	Answer Quality	Grounding/Sources	Stability/Robustness
Human evaluation	Human judgment as ground truth for correctness and relevance (Sudjianto et.al., 2024), (Joshi, 2025)	Human verification of source attribution accuracy (Sudjianto et.al., 2024)	Human assessment of consistency across test cases (Sudjianto et.al., 2024)
LLM reliability reliability evaluation	Hallucination detection directly impacts correctness (Joshi, 2025), (Niu et.al., 2024)	Consistency, calibration, and adversarial resistance (Niu et.al., 2024)	Consistency, calibration, and adversarial resistance (Sudjianto et.al., 2024), (Joshi, 2025), (Yuan et.al., 2024)
RAG-specific evaluation	Completeness, correctness, and relevance of generated responses (Krishna et.al., 2025), (Es et.al., 2024), (Zhu et.al., 2025)	Retrieval quality, attribution accuracy, and citation verification (Song et.al., 2024), (Krishna et.al., 2025), (Birur et.al., 2024), (Es et.al., 2024)	Consistency across retrieval variations and adversarial contexts [L1]
Multi-Dimensional Approaches	Accuracy, completeness, and relevance metrics [2]	Accuracy, completeness, and relevance metrics [2]	Calibration, consistency, and adversarial resistance testing [2]
AI-AI comparison	Automated assessment of correctness, completeness, and relevance (Mortaheb et.al., 2025), (Joshi, 2025)	Citation verification and attribution checking (Mortaheb et.al., 2025)	Consistency testing across multiple evaluators (Joshi, 2025)
Risk management	Accuracy requirements for regulatory compliance (Guldimann, 2024), (Yuan et.al., 2024)	Verifiable accountability and source transparency (Jiao et.al, 2025)	Reliability requirements for high-stakes applications (Kenthapadi et.al, 2024), (Guldimann, 2024), (Yuan et.al., 2024)

4 Conclusion

This study underscores the critical importance of rigorous evaluation for the responsible deployment of LLMs RAG systems, particularly within high-stakes technical domains such as renewable energy. We propose a Three-Dimensional Framework for assessing LLM/RAG reliability, synthesizing insights from existing evaluation methodologies. Our literature review highlights the necessity of

comprehensive, multi-faceted evaluation approaches that integrate domain expert input. Current research trends emphasize the development of evaluation frameworks that balance automation and human oversight to ensure scalability and robustness.

We categorize LLM/RAG evaluation techniques into six methods, organized along three key dimensions: answer quality, grounding, and stability assessment. Additionally, we introduce domain-specific modifications tailored to renewable energy applications to address unique contextual requirements.

This work reflects the present capabilities and constraints of evaluation frameworks, acknowledging that the field remains rapidly evolving. Our scope is limited to text-based LLM/RAG systems, reflecting current boundaries in integrating numerical data and performing quantitative calculations. In the future research it is necessary to advance evaluation methods that accommodate these limitations and support broader, more integrated system assessments.

Acknowledgment

This research is supported by Erasmus + KA220-HED-000166765 project: AI2SEP: Developing Talents in Artificial Intelligence to Solve Disruptive Environmental Problems.

References

Birur, N. A., Baswa, T., Kumar, D., Loya, J., Agarwal, S., & Harshangi, P. (2024). Vera: Validation and enhancement for retrieval augmented systems. arXiv. https://arxiv.org/abs/2409.15364

Bommasani, R., et al. (2023). Holistic evaluation of language models. Annals of the New York Academy of Sciences, 1525(1), 140–146. https://doi.org/10.1111/nyas.15069

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., … Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45. https://doi.org/10.1145/3637928

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval-augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150–158). https://aclanthology.org/2024.eacl-demo.16

Guldimann, P., et al. (2024). Compl-AI framework: A technical interpretation and LLM benchmarking suite for the EU Artificial Intelligence Act. arXiv. https://arxiv.org/abs/2410.07959

Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2025). Navigating LLM ethics: Advancements, challenges, and future directions. AI Ethics, 5, 5795–5819. https://doi.org/10.1007/s43681-024-00513-6

Joshi, S. (2025). Evaluation of large language models: Review of metrics, applications, and methodologies. Preprints. https://www.preprints.org/manuscript/202504.0369

Joshi, H. (2025). Joint evaluation (Jo. E): A collaborative framework for rigorous safety and alignment evaluation of AI systems integrating human expertise, LLMs, and AI agents. Preprints. https://doi.org/10.20944/preprints202509.0042.v1

Kenthapadi, K., Sameki, M., & Taly, A. (2024). Grounding and evaluation for large language models: Practical challenges and lessons learned. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6523–6533). https://doi.org/10.1145/3637528.3671897

Krishna, S., Krishna, K., Mohananey, A., Schwarcz, S., Stambler, A., Upadhyay, S., & Faruqui, M. (2025). Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 4745–4759). https://aclanthology.org/2025.naacl-long.370

Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9004–9017). https://aclanthology.org/2023.emnlp-main.561

Mortaheb, M., Khojastepour, M. A. A., Chakradhar, S. T., & Ulukus, S. (2025). RAG-Check: Evaluating multimodal retrieval-augmented generation performance. arXiv. https://doi.org/10.48550/arXiv.2501.03995

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., … Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 10862–10878). https://aclanthology.org/2024.acl-long.591

Prabhune, S., & Berndt, D. J. (2024). Deploying large language models with retrieval-augmented generation. arXiv. https://doi.org/10.48550/arXiv.2411.11895

Song, M., Sim, S. H., Bhardwaj, R., Chieu, H. L., Majumder, N., & Poria, S. (2024). Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. arXiv. https://arxiv.org/abs/2409.11242

Sudjianto, A., Zhang, A., Neppalli, S., Joshi, T., & Malohlava, M. (2024). Human-calibrated automated testing and validation of generative language models. arXiv. https://arxiv.org/abs/2411.16391

Tu, S., Wang, Y., Yu, J., Xie, Y., Shi, Y., Wang, X., … Li, J. (2024). R-Eval: A unified toolkit for evaluating domain knowledge of retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 5813–5824). https://doi.org/10.1145/3637528.3671664

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). Evaluation of retrieval-augmented generation: A survey. In CCF Conference on Big Data (pp. 102–120). Springer. https://doi.org/10.1007/978-981-99-9063-1_9

Yuan, X., et al. (2024). S-Eval: Towards automated and comprehensive safety evaluation for large language models. arXiv. https://arxiv.org/abs/2405.08974

Zhu, K., Luo, Y., Xu, D., Yan, Y., Liu, Z., Yu, S., … Sun, M. (2025). Rageval: Scenario-specific RAG evaluation dataset generation framework. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8520–8544).