LLM-as-a-Judge for Low-Resource Languages: Adapting Ragas and Comparative Ranking for Romanian

Claudiu Creanga[2,3], **Liviu P. Dinu**[1,3]
[1] Faculty of Mathematics and Computer Science,
[2] Interdisciplinary School of Doctoral Studies,
[3] HLT Research Center,
University of Bucharest, Romania
claudiu.creanga@fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Evaluating Retrieval-Augmented Generation (RAG) systems remains a challenge for Low-Resource Languages (LRLs), where standard reference-based metrics fall short. This paper investigates the viability of the "LLM-as-a-Judge" paradigm for Romanian by adapting the Ragas framework using next-generation models (Gemini 2.5 and Gemini 3). We introduce **AdminRo-Eval**, a curated dataset of Romanian administrative documents annotated by native speakers, to serve as a ground truth for benchmarking automated evaluators. We compare three evaluation methodologies—direct scoring, comparative ranking, and granular decomposition—across metrics for Faithfulness, Answer Relevance, and Context Relevance. Our findings reveal that evaluation strategies must be metric-specific: granular decomposition achieves the highest human alignment for *Faithfulness* (96% with Gemini 2.5 Pro), while comparative ranking outperforms in *Answer Relevance* (90%). Furthermore, we demonstrate that while lightweight models struggle with complex reasoning in LRLs, the Gemini 2.5 Pro architecture establishes a robust, transferable baseline for automated Romanian RAG evaluation.

1 Introduction

Building AI systems is inherently an iterative loop of design, evaluation, and refinement. However, in the context of **Low-Resource Languages (LRLs)**, this loop is often disrupted by a scarcity of reliable benchmarks (Joshi et al., 2020). While recent initiatives have started to bridge this gap for Romanian in tasks like lexical simplification (Anghel et al., 2025), resources for complex reasoning pipelines remain limited. In the era of LLMs, where outputs are non-deterministic, the ability to rapidly and accurately measure system fidelity is important. Evaluation in this context serves as the functional equivalent of unit tests in traditional software engineering, providing the essential signal required to distinguish improvement from regression, a signal that is frequently absent or noisy for underrepresented languages (Kreutzer et al., 2022).

To address the intrinsic limitations of frozen LLMs, specifically their tendency toward **hallucinations** and their inability to access information beyond their training data cutoffs, RAG systems have emerged as the dominant architectural paradigm. By dynamically anchoring the model's generative capabilities to verifiable, external knowledge sources, RAG reduces hallucinations. While the conceptual framework of RAG is well-established, evaluating these pipelines presents unique challenges for the low-resource community.

A significant contribution in this domain is the **Ragas framework** (Es et al., 2023), which introduced a suite of reference-free metrics (such as context precision, faithfulness, and answer relevance) for the automated assessment of RAG pipelines. This framework leverages an "LLM-as-a-Judge" approach, utilizing an advanced model to critique the performance of the RAG system. However, the original validation of Ragas was heavily English-centric and relied on the now-dated **gpt-3.5-turbo** architecture (Ouyang et al., 2022). This English focus leaves a gap regarding the framework's applicability to lower-resource languages like **Romanian**, which often suffer from performance degradation when using tools optimized for English. We contend that for an "LLM-as-a-Judge" to be effective in a low-resource setting, the judge itself must possess state-of-the-art multilingual reasoning capabilities to compensate for the lack of training data. Consequently, we focus on applying the Ragas framework to Romanian datasets by replacing legacy evaluators with next-generation engines: **Gemini 2.5** (Gemini Team, 2025) and **Gemini 3** (Gemini Team, Google DeepMind, 2025). By leveraging these newer, more powerful models, we aim to establish a robust baseline for automated RAG evaluation in Romanian, providing a transferable

strategy for other low-resource languages.

Our **contributions** are:

- **First Application of Ragas to Romanian**: We conduct the first rigorous evaluation of the RAGAs framework in a non-English context. By validating these metrics on Romanian data, we demonstrate their cross-lingual portability and establish a precedent for reference-free evaluation in LRLs.
- **Benchmarking Next-Generation Evaluators**: Deviating from the framework's original reliance on GPT-3.5, we integrate and benchmark a suite of modern architectures, including **Gemini 2.5 Pro**, **Gemini 3 Pro**, and **Gemini 3 Flash**. Our comparative analysis reveals that specific Gemini variants significantly outperform legacy baselines in aligning automated scores with human judgment, challenging the "one-size-fits-all" approach to model selection.
- **Introduction of AdminRo-Eval**: We release a novel, domain-specific evaluation dataset curated from official Romanian public administration documents. Named **AdminRo-Eval**, this resource provides a challenging, real-world benchmark for RAG systems in a formal, knowledge-intensive domain. It is specifically designed to aid in the development of public sector AI agents, offering a direct metric for how accurately systems can retrieve and synthesize official documentation for citizen queries.

2 Related Work

Evaluating RAG systems is uniquely challenging in low-resource settings where "gold standard" reference texts are virtually non-existent. Historically, metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have dominated NLP. However, their reliance on lexical overlap with human-curated references makes them inadequate for RAG, where correct answers often diverge in vocabulary. For Low-Resource Languages (LRLs), where creating high-quality reference datasets is prohibitively expensive, these metrics are particularly ill-suited. Furthermore, while detecting the *presence* of machine-generated text is feasible (Marchitan et al., 2024), assessing its *accuracy* without gold references is far more complex.

While embedding-based metrics like BERTScore (Zhang et al., 2020) improve upon lexical rigidity, they often typically retain the dependency on reference text. To bypass this data bottleneck, the field has shifted toward the "LLM-as-a-Judge" paradigm (Zheng et al., 2023). This approach employs a strong LLM as an adjudicator, offering a scalable alternative to human evaluation.

External verification methods like FActScore (Min et al., 2023), which cross-reference claims against knowledge bases (e.g., Wikipedia), face scalability hurdles in low-resource domains where such bases are often sparse or outdated. Consequently, the **Ragas framework** (Es et al., 2023) has emerged as a vital tool. By atomizing evaluation into reference-free metrics such as **Faithfulness** and **Context Relevance**, Ragas assesses reasoning directly. Our work adopts this methodology to determine if such reference-free frameworks can be robustly transferred to the Romanian linguistic context.

3 Methodology

3.1 Data Source

Official documents from Romanian public administration bodies provided the raw material for our corpus. The pipeline for gathering and preparing this data involved the following steps:

- **Data Sources**: The initial collection comprised 84 PDF documents from institutions such as the Romanian Health Ministry, the Senate, and official publishing houses. These sources guarantee that the dataset contains the formal language and complex subject matter typical of real-world retrieval tasks.
- **Filtering**: We restricted the dataset to documents published within the two months preceding collection. This temporal filter prevents data leakage, ensuring the LLM did not encounter the text during its training. We further refined the selection by removing pages containing mostly figures, signatures, or blank space, leaving 27 usable multi-page PDFs.
- **Chunking**: Since RAG systems process smaller segments rather than full documents, we divided the filtered PDFs into contiguous 3-page blocks. This process yielded a final set of 100 specific PDF chunks to serve as the knowledge context.

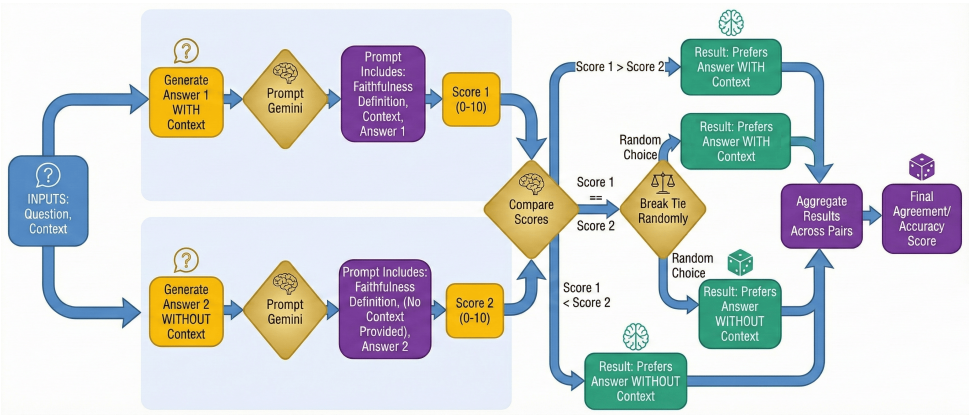

Figure 1: Operational flow of the Baseline (LLM-as-Judge) evaluation. This method isolates the assessment process by prompting the LLM to assign an independent scalar score (0-10) to the output.

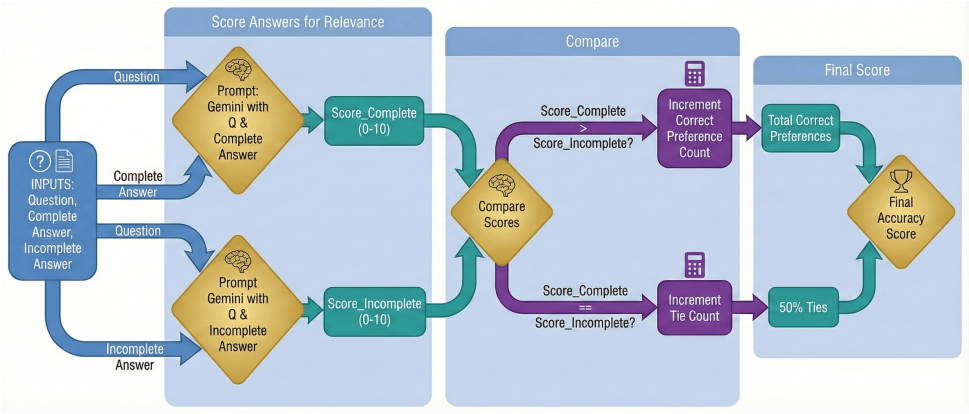

Figure 2: **Answer Relevance Baseline.** Gemini independently scores a complete and an incomplete answer. The scores are compared to determine correct preferences.

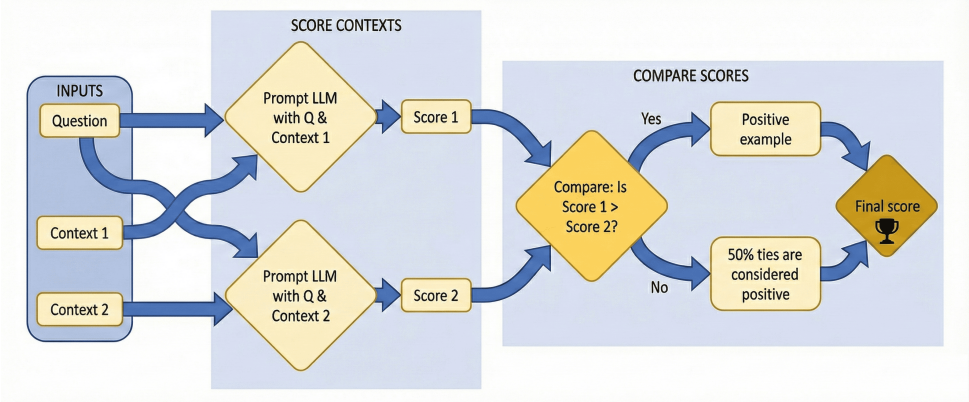

Figure 3: **Context Relevance Baseline.** The LLM independently assigns scalar scores to a relevant and an irrelevant context relative to the query.

3.2 Data Generation in Romanian

Creating clear test cases required generating contrastive pairs for each metric: one correct example and one intentionally flawed counterpart. This binary approach simplifies the decision for both human judges and the LLM. We conducted all prompting and generation using the Gemini model in Romanian to maintain linguistic accuracy.

For every one of the 100 PDF chunks, the process included:

- **Question Extraction**: Using prompts in Romanian, Gemini analyzed each PDF chunk to formulate a single, text-supported question. This resulted in 100 base questions.
- **Generating Paired Data for Evaluation**:
  - **For Faithfulness**: We aimed to distinguish between grounded answers and hallucinations. First, the model generated "Answer 1" using the provided context. We then requested a response to the same question without supplying the context, forcing the model to rely on internal memory and increasing the risk of error ("Answer 2").
  - **For Answer Relevance**: We contrasted helpful responses with unhelpful ones. The model first produced a direct, complete response ("Answer 1"). A subsequent prompt explicitly requested an evasive or partial reply to the same question, creating "Answer 2."
  - **For Context Relevance:** The original 3-page chunk served as the precise "Context 1." To test noise handling, we added an unrelated page from a different document to the original chunk, resulting in a less relevant "Context 2."

This approach produced a dataset containing 100 contrastive pairs per metric, enabling a controlled comparison in Romanian.

3.3 Human Annotation

To define a gold standard for the Romanian language, we first used Gemini to generate 100 questions (one per PDF chunk) and contrastive synthetic pairs for each metric: grounded vs. ungrounded answers for Faithfulness, complete vs. evasive answers for Answer Relevance, and original vs. noisy chunks for Context Relevance.

Two native speakers then acted as annotators, evaluating these 100 generated questions and their corresponding synthetic pairs through a comparative process.

The evaluation relied on a forced-choice format. For metrics like Faithfulness, annotators reviewed the question, the context, and two distinct answers, selecting the one that adhered better to the text (see Figure 10). Similar comparative tasks were used for Context Relevance (Figure 11) and Answer Relevance (Figure 12). We provided detailed guidelines in Romanian to ensure consistent scoring. These human judgments function as the ground truth for testing our automated evaluation.

4 Framework for Romanian RAG Evaluation

To rigorously assess the output quality of our Retrieval-Augmented Generation (RAG) system, we implemented an evaluation protocol strictly aligned with the metrics established in the Ragas research.

4.1 Ragas Metrics

We adapted three core metrics from the Ragas framework to specifically measure the fidelity and utility of generated Romanian content:

- Faithfulness: This metric evaluates the extent to which the generated response, $a_s(q)$, is factually grounded in the retrieved context, c(q). A response is deemed faithful only if every assertion it contains can be logically inferred from the source material.
- Answer Relevance: This indicator assesses the alignment between the generated response and the user's initial inquiry.
- Context Relevance: This metric audits the retrieval component itself. A context c(q) is classified as relevant if it provides exactly the information required to address the query, minimizing the signal-to-noise ratio by excluding superfluous content.

5 Automated Assessment Strategies

We conducted a comparative study of three distinct automated evaluation methodologies, utilizing the Gemini models as the adjudicators. The objective was to identify which scoring technique demonstrated the highest correlation with human-annotated ground truth.

5.1 Method 1: Direct Scoring (LLM-as-Judge)

This baseline approach utilizes the LLM as a scalar scorer. For every answer pair, we issue distinct prompts to the model, requesting a numerical rating between 0 and 10 based on the specific metric definitions.

As illustrated in Figure 1 and Table 1, this technique secured an 86% alignment with human judgments regarding Faithfulness when using the Gemini 2.5 Pro model and 82% when using Gemini 3 Pro. We similarly applied this independent scoring methodology to Answer Relevance (Figure 2) and Context Relevance (Figure 3).

5.2 Method 2: Comparative Ranking (LLM-as-Ranker)

By shifting the paradigm from absolute scoring to relative ranking, this method simplifies the evaluation task. We present Gemini with the query alongside two competing answers in a single prompt, requesting a binary preference for the superior response.

This pairwise comparison yielded a 90% agreement rate with human judgments for Answer Relevance using Gemini 2.5 Pro (see Figure 4 and Table 1). The corresponding ranking architectures for Faithfulness and Context Relevance are detailed in Figures 5 and 6, respectively.

5.3 Method 3: Granular Decomposition (Ragas-style)

Representing the central methodology of the Ragas framework, this approach dissects the evaluation process into atomic verification steps.

- **Faithfulness Evaluation (Ragas-style):** As shown in Figure 7, Gemini parses the Romanian response into distinct statements (s_i) and verifies if each is supported by the context. The final score is the ratio of verified statements to total statements:

$$F = \frac{|V(a)|}{|S(a)|}$$

  This granular approach achieved a **96%** agreement with human annotations when using Gemini 2.5 Pro.

- **Answer Relevance Evaluation (Ragas-style):** We prompt Gemini to synthesize n potential inquiries for the generated answer (Figure 8). We then compute the cosine similarity between the original query q and these synthetic questions q'_i:

$$AR = \frac{1}{n} \sum_{i=1}^{n} \text{sim}(q, q'_i)$$

  This technique resulted in an **85%** agreement for Gemini 2.5 Pro.

- **Context Relevance (Ragas-style):** Illustrated in Figure 9, this method tasks the model with extracting only the "essential sentences" required to answer the query. The score is calculated as the ratio of essential sentences to the total sentence count in the chunk. Here Gemini 3 pro obtained the best score of 93%.

6 Results

Table 1 presents a comparison of the alignment between automated metrics and human judgments across three dimensions: Faithfulness, Answer Relevancy, and Context Relevancy. The evaluation contrasts the performance of three distinct judge models (Gemini 3 Pro, Gemini 3 Flash, and Gemini 2.5 Pro) across both Romanian (Low-Resource) and English (High-Resource) datasets.

6.1 Performance on Romanian Data

In the context of the Romanian dataset, the **Gemini 2.5 Pro** model demonstrated superior performance in assessing **Faithfulness**, achieving a 96% alignment with human annotators using the Ragas-style decomposition method. This outperformed Gemini 3 Pro (84%) and Gemini 3 Flash (87%) using the same methodology. For **Answer Relevancy**, the Comparative Ranking (LLM-as-Ranker) approach proved most effective. Gemini 2.5 Pro again led the benchmarks with a 90% agreement rate, whereas Gemini 3 Pro and Flash achieved lower correlations of 80% and 77%, respectively.

In **Context Relevancy**, **Gemini 3 Pro** utilizing the Ragas framework achieved the highest score in the entire Romanian evaluation set (93%), demonstrating a good capability in filtering noise. In contrast, the lightweight Gemini 3 Flash model failed to produce meaningful signal for this metric, scoring between 0% and 5% across all methodologies.

6.2 Performance on English Data

On the English control dataset, performance was universally higher. **Gemini 3 Pro** achieved near-perfect alignment (100%) for Context Relevancy

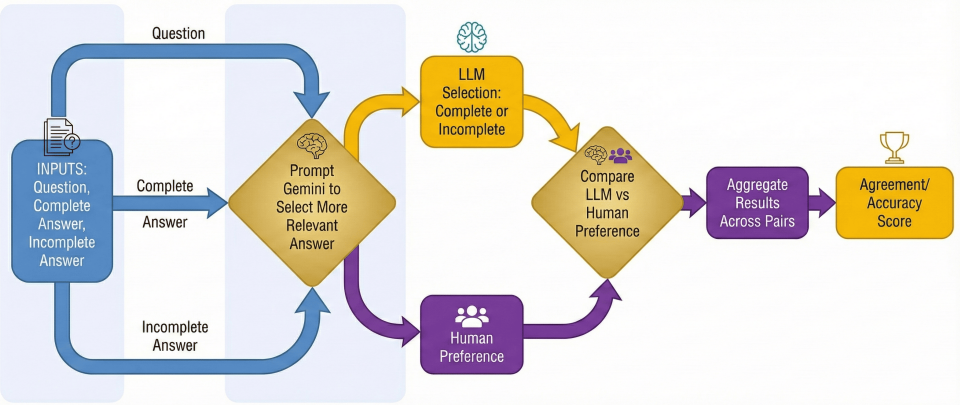

Figure 4: **Answer Relevance Ranking**: Pairwise comparison method where Gemini is prompted to identify the more relevant response between a complete and an incomplete answer.

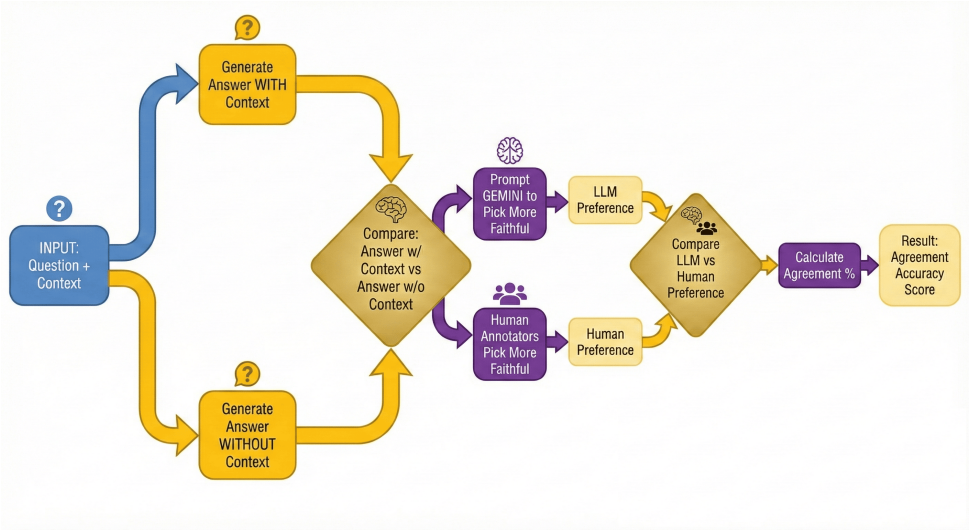

Figure 5: **Faithfulness Ranking**: The model is presented with a simultaneous view of competing answers and must discriminatively select the candidate with higher factual grounding.

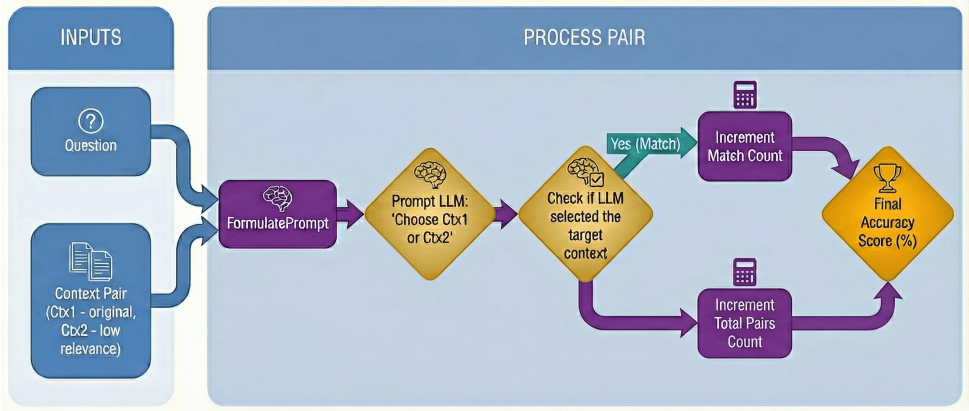

Figure 6: **Context Relevance Ranking**: Instead of independent scoring, the LLM evaluates a pair of contexts simultaneously to select the most relevant one.

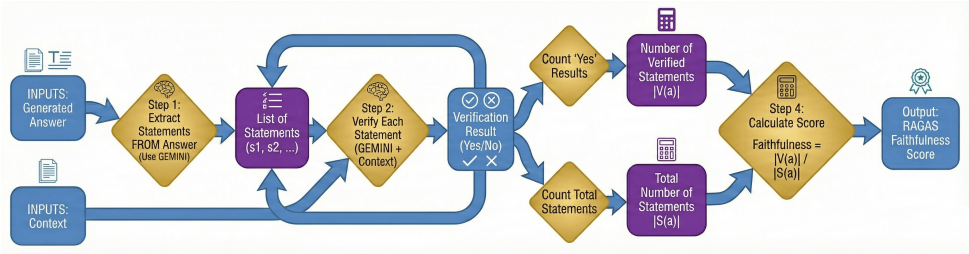

Figure 7: **Faithfulness Decomposition**: The answer is deconstructed into atomic statements and validated against the source context.

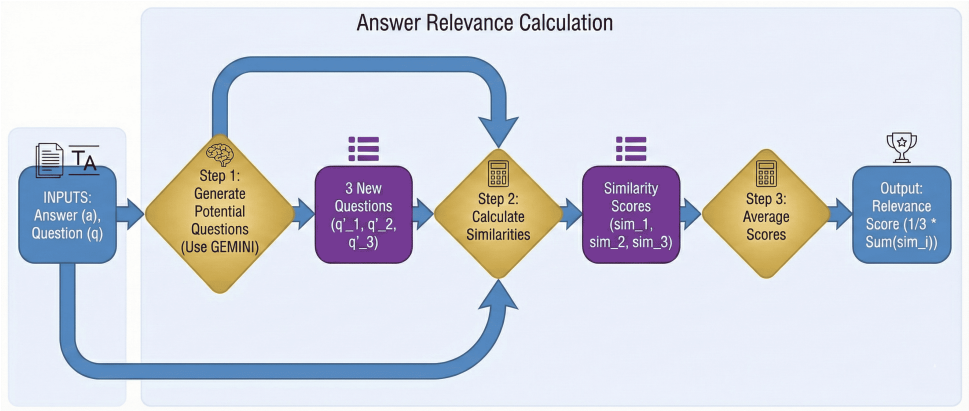

Figure 8: **Answer Relevance Ragas**: The LLM generates hypothetical questions (q_i') based on the answer to check semantic alignment with the original query.

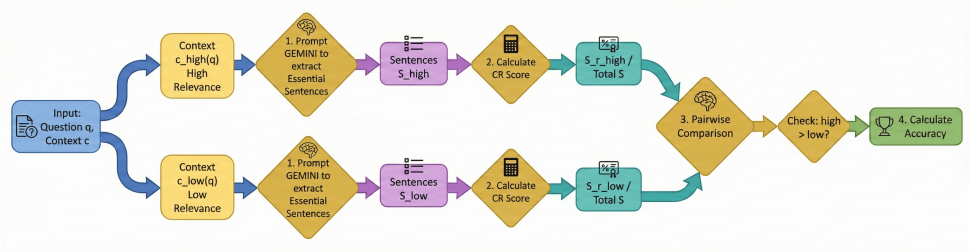

Figure 9: **Context Relevance Ragas**: The LLM extracts "essential sentences" to determine the signal-to-noise ratio of the retrieved context.

Table 1: Comparative Evaluation of LLM Judges on Romanian and English Datasets

Model	Faithfulness	Answer Relevancy	Context Relevancy
Romanian Dataset Evaluation			
Gemini 3 Pro	**Ragas(84%)** Baseline(82%) Ranking(81%)	Ragas(76%) **Baseline(83%)** Ranking(80%)	**Ragas(93%)** Baseline(0%) Ranking(8%)
Gemini 3 Flash	**Ragas(87%)** Baseline(76%) Ranking(76%)	Ragas(76%) Baseline(76%) **Ranking(77%)**	Ragas(0%) Baseline(2%) **Ranking(5%)**
Gemini 2.5 Pro	**Ragas(96%)** Baseline(86%) Ranking(85%)	Ragas(85%) Baseline(83%) **Ranking(90%)**	**Ragas(79%)** Baseline(70%) Ranking(65%)
English Dataset Evaluation			
Gemini 3 Pro	**Ragas(94%)** Baseline(90%) Ranking(92%)	Ragas(84%) Baseline(89%) **Ranking(92%)**	**Ragas(100%)** Baseline(90%) Ranking(90%)
Gemini 3 Flash	**Ragas(98%)** Baseline(86%) Ranking(85%)	**Ragas(84%)** Baseline(70%) Ranking(82%)	Ragas(8%) Baseline(20%) **Ranking(32%)**
Gemini 2.5 Pro	**Ragas(98%)** Baseline(85%) Ranking(88%)	Ragas(68%) Baseline(69%) **Ranking(73%)**	**Ragas(90%)** Baseline(85%) Ranking(78%)

using the Ragas method. Similarly, Faithfulness scores were elevated, with Gemini 3 Flash reaching 98% accuracy via the Ragas method, surpassing its performance on the Romanian equivalent by 11 percentage points.

7 Discussion

The results illuminate the disparities between high-resource and low-resource evaluation pipelines, while also highlighting the nuance required in selecting the appropriate "Judge" model.

A primary observation from Table 1 is the consistent performance degradation when moving from English to Romanian. For instance, while Gemini 3 Pro achieved 94% alignment for Faithfulness in English, this dropped to 84% for Romanian. Similarly, Gemini 3 Flash's ability to assess Answer Relevancy dropped from 84% (English) to 76% (Romanian). This confirms our hypothesis that "LLM-as-a-Judge" systems are not language-agnostic, their reliability as evaluators is intrinsically tied to their pre-training exposure to the target language. Consequently, RAG pipelines for LRLs cannot blindly rely on default English-centric prompts and thresholds.

Our data suggests a bifurcation in the optimal evaluation strategy depending on the metric type. For objective metrics like **Faithfulness**, the granular Ragas-style decomposition consistently yielded the highest scores (e.g., 96% for Gemini 2.5 Pro). This suggests that breaking down claims into atomic statements helps the model overcome linguistic ambiguity in Romanian. Conversely, for **Answer Relevancy**, the Comparative Ranking method frequently outperformed the Ragas decomposition (e.g., Gemini 2.5 Pro scored 90% with Ranking vs. 85% with Ragas). This indicates that for nuanced, holistic judgments of "quality", simpler pairwise comparisons are easier for the model to process than complex embedding-based synthetic generation.

Unexpectedly, the older **Gemini 2.5 Pro** architecture outperformed the newer Gemini 3 variants in several Romanian benchmarks, particularly Faithfulness (96% vs. 84/87%). While Gemini 3 Pro demonstrated superior reasoning in Context Relevancy (93%), Gemini 2.5 Pro offered the most consistent all-around performance for the Romanian language.

Furthermore, the failure of **Gemini 3 Flash** in

Context Relevancy (scoring near 0%) serves as a cautionary tale. While "Flash" models are attractive for their speed and cost, they appear to lack the depth of reasoning required to discern subtle noise in retrieved Romanian contexts, making them unsuitable as judges for this specific metric.

8 Conclusion and Future Work

This study bridges a gap in the evaluation of RAG systems for Low-Resource Languages. By introducing **AdminRo-Eval** and benchmarking three automated evaluation strategies, we demonstrated that English-centric metrics cannot be blindly transferred to Romanian without performance degradation.

Our experiments reveal that the "LLM-as-a-Judge" paradigm is viable for Romanian, provided the methodology is tailored to the specific metric. We found that the granular decomposition of the Ragas framework is superior for objective metrics like **Faithfulness** (96% alignment with Gemini 2.5 Pro), whereas Comparative Ranking is more effective for nuanced semantic judgments like **Answer Relevance** (90% alignment). Furthermore, our results caution against the use of lightweight "Flash" models for complex reasoning tasks in LRLs, as evidenced by their inability to accurately assess Context Relevance. Ultimately, while newer models like Gemini 3 Pro excel in specific areas, the older Gemini 2.5 Pro architecture provided the most robust and consistent baseline for Romanian RAG evaluation.

To further advance RAG evaluation for underrepresented languages, we propose to expand the AdminRo-Eval dataset beyond public administration to include medical and legal domains, where hallucination risks carry higher stakes. Also, reliance on proprietary models (Gemini/GPT) limits accessibility for the LRL community, so we would want to assess the efficacy of open-weights models (e.g., Llama 3, Mistral) as judges also.

9 Limitations

Our study faces three primary limitations. First, the **AdminRo-Eval** dataset focuses exclusively on formal public administration documents. Findings may not generalize to open-domain or informal contexts where semantic ambiguity is higher. Second, the modest size of our manually annotated ground truth (100 samples) limits the statistical power. Third, since Gemini was used to generate the synthetic test data, the high agreement scores may partially reflect a 'self-enhancement' bias, where the model preferentially favors its own generation style.

10 Ethics Statement

This work utilizes the **AdminRo-Eval** dataset, which was constructed exclusively from public documents available on official Romanian government websites. We performed manual verification to ensure that no sensitive Personally Identifiable Information (PII) of private citizens was included in the corpus. Regarding the human evaluation, the native speaker annotators were volunteers. While our research aims to enhance the factual reliability of RAG systems, we acknowledge that the "Judge" models (Gemini) may still exhibit biases inherent in their pre-training data. We emphasize that automated evaluation tools for the public sector are intended to assist human decision-making, not to replace human oversight in administrative processes.

11 Acknowledgments

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P7-7.1-PTE-2024-0046, within PNCDI IV.

References

Florentina Anghel, Craciun Petru-Theodor, Claudiu Creanga, and Sergiu Nisioi. 2025. Rals: Resources and baselines for romanian automatic lexical simplification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Gemini Team, Google DeepMind. 2025. Gemini 3 technical report. Technical report, Google DeepMind. Released November 18, 2025.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, and 1 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Teodor-George Marchitan, Claudiu Creanga, and Liviu P Dinu. 2024. Transformer and hybrid deep learning based models for machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 403–411.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12076–12100.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and Ion P. Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

A Annotation Pipeline

The following figures illustrate the annotation interfaces used for human evaluation of the three metrics.

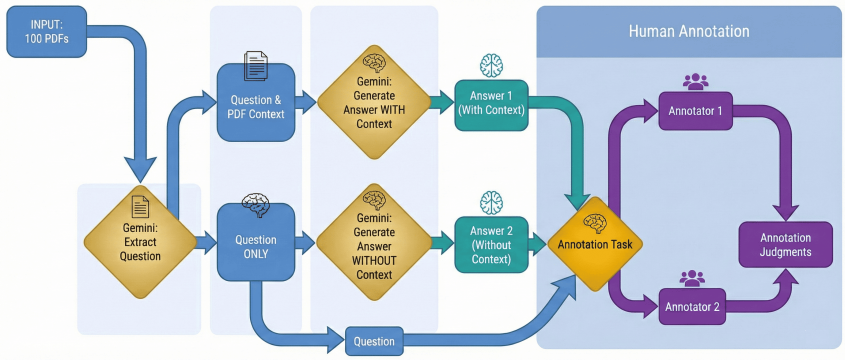

Figure 10: Interface for Faithfulness annotation. Reviewers examined the question, context, and two candidate answers to determine which response was strictly supported by the source text.

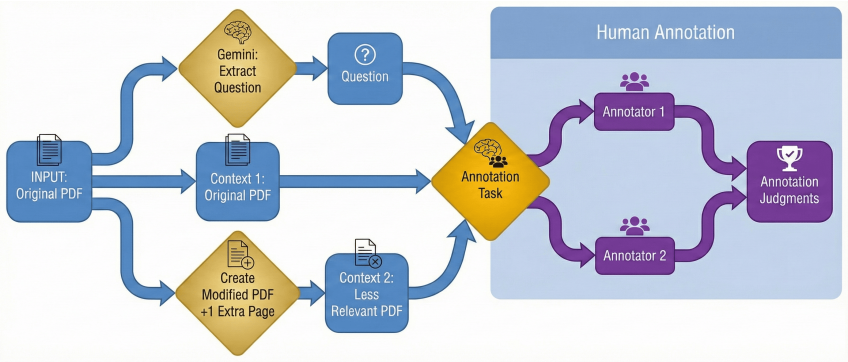

Figure 11: Interface for Context Relevance annotation. Annotators selected the most useful passage for answering the prompt, choosing between the original context and a version modified with extraneous text.

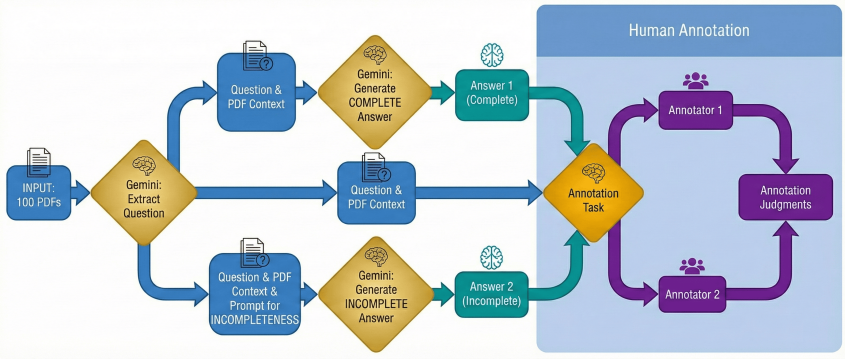

Figure 12: Interface for Answer Relevance annotation. The task required selecting the most helpful response by comparing a complete, direct answer against an intentionally partial or evasive one.