

FusionRAG-Ex: An Interactive RAG Framework with Retrieval and Confidence-Aware Explanations

Chandana Sree Mala[1(✉)], Gizem Gezici[2], Sezer Kutluk[2], and Fosca Giannotti[2]

[1] University of Pisa, Pisa, Italy
c.mala@studenti.unipi.it
[2] Scuola Normale Superiore, Pisa, Italy
{gizem.gezici,sezer.kutluk,fosca.giannotti}@sns.it

Abstract. RAG improves factual grounding in LLMs, yet diagnosing whether outputs are faithfully supported by retrieved evidence remains challenging. We present FusionRAG-Ex, an interactive framework for explainable and faithfulness-aware RAG analysis. FusionRAG-Ex integrates retrieval relevance and token-level confidence into explanation generation, producing highlights that distinguish grounded from weakly supported responses. The system enables real-time inspection of retrieval scores, answer correctness, and explanation signals across retrievers, LLMs, and datasets, supporting systematic analysis and development of more trustworthy RAG systems.

Keywords: RAG · Explainable AI · Trustworthy AI

1 Introduction

Large Language Models (LLMs) generate fluent responses but may produce unsupported or incorrect outputs, raising reliability concerns. Retrieval-Augmented Generation (RAG) mitigates this by grounding responses in external evidence; however, diagnosing whether outputs are faithfully supported by retrieved content remains challenging [2]. RAG-Ex [3], later extended to retrieval explanations [4], provides perturbation-based unit-level explanations for RAG outputs. However, these approaches assume a fixed retrieved context and do not explicitly incorporate retrieval relevance or model confidence into explanation generation, treating retrieved evidence uniformly and limiting faithfulness diagnostics.

We introduce FusionRAG-Ex, an interactive framework for explainable and faithfulness-aware RAG analysis. FusionRAG-Ex integrates retrieval scores and token-level confidence into explanation generation, producing retrieval- and confidence-aware highlights that distinguish *grounded* from *weakly supported*

The live demo is available at: https://huggingface.co/spaces/SNS-KDDLab/FusionRAG-Ex.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Baratchi et al. (Eds.): ECML PKDD 2026, LNAI 16950, pp. 315–319, 2027.
https://doi.org/10.1007/978-3-032-37685-5_25

responses. The system supports multi-level faithfulness diagnostics within a unified RAG pipeline, enabling real-time inspection of retrieval relevance, answer correctness, and confidence-weighted explanations across retrievers, LLMs, and datasets. Our main contributions are as follows.

- **Modular RAG Framework:** Interactive system supporting multiple retrievers, LLMs, and benchmark datasets.
- **Retrieval and Confidence Augmented Explanations:** Extension of RAG-Ex with retrieval weighting and confidence fusion for evidence-grounded explanations.
- **Faithfulness Diagnostics:** Multi-level analysis integrating retrieval relevance, answer correctness, and confidence signals, benchmarked against RAG-Ex.

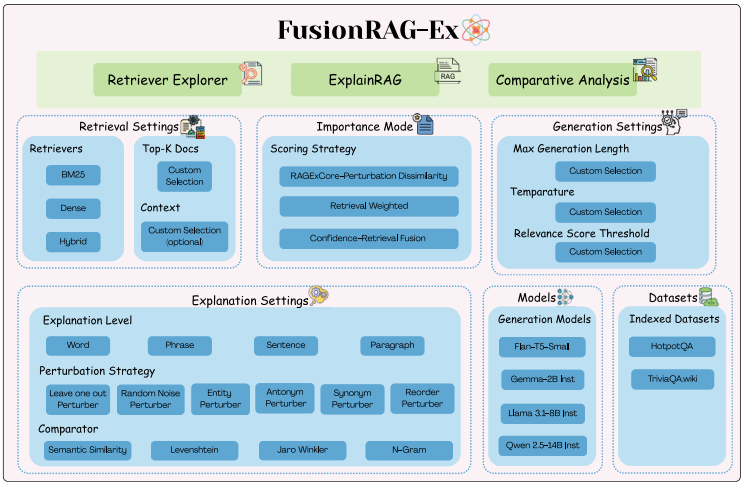


Fig. 1. Architecture of FusionRAG-Ex. Light blue boxes denote modules; dashed boxes indicate grouped functional components. (Color figure online)

2 FusionRAG-Ex Architecture

FusionRAG-Ex is an interactive and modular framework that integrates a complete end-to-end RAG pipeline with retrieval and confidence augmented explanations and systematic faithfulness diagnostics (Fig. 1). Unlike the isolated explainer in RAG-Ex, it enables controlled experimentation across retrieval, generation, and explanation components within a unified system. The framework supports sparse, dense, and hybrid retrieval. Sparse retrieval uses BM25,

while dense retrieval leverages MPNet[1]. Hybrid retrieval combines sparse and dense candidates with BGE re-ranking[2]. Multiple LLMs are supported, including Gemma[3] and Llama[4]. The modular design enables controlled comparison across retrievers, LLMs, and datasets.

FusionRAG-Ex integrates retrieval relevance and model confidence into explanation generation. Retrieval scores are normalized to weight unit importance, and token-level confidence is used to further modulate these weights. We extend the original RAG-Ex formulation by combining the document retrieval score and the model confidence score in a weighted manner. This joint modeling yields information-aware explanations that distinguish faithfully grounded from weakly supported response segments. The system is publicly available[5] (Fig. 2).

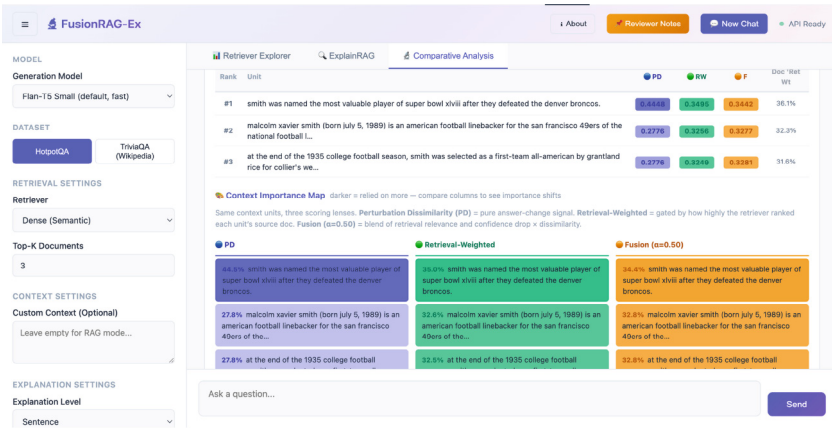


Fig. 2. FusionRAG-Ex interface illustrating retrieval scores, answer similarity, and explanations weighted by retrieval relevance and model confidence.

2.1 Faithfulness Diagnostics

FusionRAG-Ex enables multi-level faithfulness diagnostics by jointly analyzing retrieval, generation, and explanation signals. The interface exposes document relevance scores and retrieval recall (when gold evidence is available), reports answer correctness (for datasets with ground-truth answers), and highlights influential context units reflecting confidence-weighted alignment with relevant

[1] `sentence-transformers/all-mpnet-base-v2`.
[2] `BAAI/bge-reranker-base`.
[3] https://ai.google.dev/gemma.
[4] https://ai.meta.com/llama/.
[5] Source code: https://github.com/chandanasreemala/XAI_RAG. Use case Demonstration video: https://www.youtube.com/watch?v=DpALH3pYBq4.

evidence. By correlating these signals across retrieval (evidence ranking), generation (answer similarity to ground truth), and explanation (confidence-modulated alignment), the system supports systematic identification of faithfully grounded and weakly supported responses within an end-to-end RAG pipeline. Explanations are evaluated against the RAG-Ex baseline.

2.2 System Overview

FusionRAG-Ex targets researchers developing trustworthy RAG systems. The interactive and customizable web interface enables real-time query submission, retrieval inspection (sparse, dense, hybrid), explanation visualization, and confidence-aware diagnostics. Users can dynamically switch retrievers, LLMs, and datasets, and can additionally upload custom (`jsonl`) datasets for BM25 and FAISS[6] indexing. Unlike RAGAS (offline metrics) [1] and RAG-Ex (static explanations) [3], FusionRAG-Ex offers interactive diagnostics that jointly account for retrieval quality and model confidence within a unified framework.

2.3 Evaluation

With a retrieval pool of 1,000 documents ($k = 3$), Hybrid achieves the highest NDCG@3 on HotpotQA (0.7794), while Dense performs best on TriviaQA (0.8076). These results suggest that hybrid retrieval is advantageous for multi-hop reasoning, whereas dense retrieval is more effective for single-hop factual queries. Using GPT-5[7] as a judge on 10 instances, FusionRAG-Ex improves evidence coverage over RAG-Ex but is slightly weaker at diagnosing relational errors, resulting in a small drop in perceived usefulness. This may be due to similar explanations under sentence-level leave-one-out perturbations on HotpotQA.

3 Conclusion and Future Work

We introduce FusionRAG-Ex, an interactive, modular and open-source framework for explainable and faithfulness-aware RAG analysis, unifying retrieval and generation, alongside retrieval and confidence augmented explanations within a single interactive pipeline. Future work will explore additional fusion signals and complementary approaches to enhance hallucination detection and faithfulness diagnostics in RAG systems.

Acknowledgments. This work has been supported by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", and by the European Union under ERC-2018-ADG GA 834756 ("XAI") and under grant agreement no. 101120763 ("TANGO").

Disclosure of Interests. The authors declare no competing interests.

[6] https://faiss.ai/.
[7] https://openai.com/it-IT/index/introducing-gpt-5/.

References

1. Es, S., James, J., Anke, L.E., Schockaert, S.: 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 (2024)
2. Huang, L., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. **43**(2), 1–55 (2025)
3. Sudhi, V., Bhat, S.R., Rudat, M., Teucher, R.: Rag-ex: a generic framework for explaining retrieval augmented generation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2776–2780 (2024)
4. Sudhi, V., Bhat, S.R., Rudat, M., Teucher, R., Flores-Herr, N.: Towards end-to-end model-agnostic explanations for rag systems (2025). arXiv preprint arXiv:2509.07620