Evaluation of Retrieval-Augmented Generation: A Survey

Summary of the paper:
Yu et al. (2024)
`arXiv:2405.07437`

Michał Gromadzki, Jakub Piwko, Kacper Skonieczka, Grzegorz Zakrzewski

January 29, 2025

Outline

Introduction

What is Retrieval-Augmented Generation?

- **Motivation:** Large Language Models (LLMs) can produce coherent yet incorrect or "hallucinated" content.
- **RAG Solution:** *Retrieval-Augmented Generation* (RAG) reduces these factual errors by incorporating relevant information retrieved from external sources.
- **Key Benefit:** Ensures that generated text is both contextually rich and grounded in real data.

RAG Structure

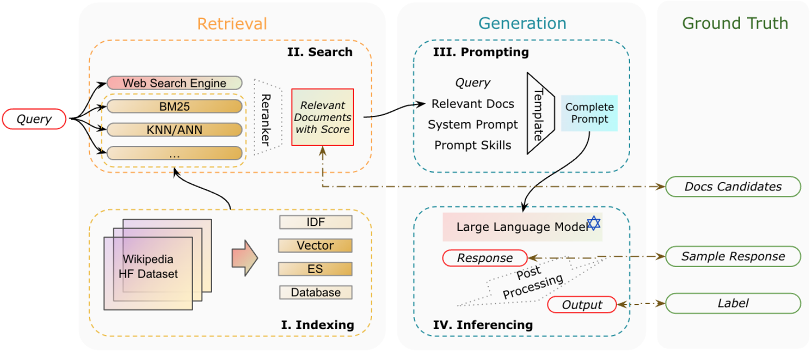


Figure 1: Overview of the RAG components

Challanges

- **Challenge 1:** Dynamically evolving external databases require flexible updates to indexes.
- **Challenge 2:** Balancing retrieval accuracy with generation quality when no ground-truth passages exist.
- **Challenge 3: Evaluating overall system performance (retrieval + generation) in diverse downstream tasks.**

Evaluation Frameworks

The authors investigated **12 distinct frameworks** that already focus on RAG evaluation.

What are the concerns?

- × Focus only on selected steps of RAG pipeline.
- × Unstandardized approach to evaluation.
- × Many metrics that assess exactly the same aspects.

This motivates the **unified framework** that addresses these multi-layered challenges.

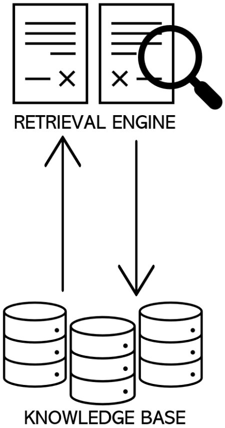
Contributions

Key Contributions:

1. *Challenge of Evaluation:* First work to classify RAG evaluation challenges by system components.
2. *A Unified Evaluation Process of RAG (Auepora)*: Proposes a structured framework to understand RAG benchmarks across multiple dimensions.
3. *Benchmark Analysis:* Comprehensive review of existing benchmarks, highlighting limitations and future directions for RAG evaluation.

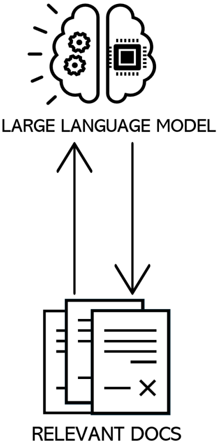
Challenges in Evaluating RAG Systems

Retrieval Challenges



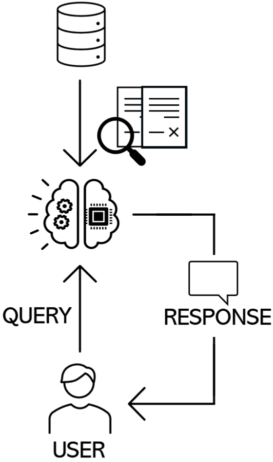
- **Dynamic Knowledge Bases:** Information may quickly become outdated or irrelevant.
- **Massive Sources:** Evaluations must handle diverse, unstructured data from large corpora or the entire web.
- **Quality Control:** Misleading or low-quality documents require robust filtering mechanisms.
- **Metric Limitations:** Precision and Recall alone cannot fully capture RAG's nuanced retrieval needs.

Generation Challenges



- **Faithfulness to Retrieved Content:** Ensuring responses accurately reflect retrieved information.
- **Contextual Relevance:** Aligning responses with user queries.
- **Subjectivity in Outputs:** Creative or open-ended tasks complicate "correctness" criteria.

System-Level Evaluation



- **Holistic Assessment:** Evaluating retrieval and generation separately overlooks cross-component interdependencies
- **Added Value of Retrieval:** Measuring how effectively retrieved data improves final answers.
- **Practical Constraints:** Latency, ambiguity handling, and user satisfaction are crucial real-world factors.

A Unified Evaluation Process of RAG (Auepora)

Auepora: Three Core Questions

A Unified Evaluation Process of RAG (Auepora) aims to streamline and clarify how we assess RAG systems by addressing three fundamental questions:

- **What to Evaluate?** (Target)
- **How to Evaluate?** (Dataset)
- **How to Measure?** (Metric)

Overview of Auepora

- **Holistic Scope:** Evaluates every stage of RAG, from retrieval to generation, within real-world constraints.
- **Modular Structure:**
  1. *Target* — determining evaluation direction.
  2. *Dataset* — comparing data constructions used in benchmarks.
  3. *Metrics* — linking specific targets and datasets with appropriate evaluation measures.
- **Practical Insights:** Identifies specific weaknesses, guiding targeted improvements.

Evaluation Target (What to Evaluate?)

Target modular

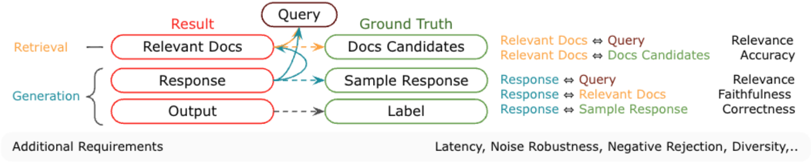


Figure 2: The Target modular of the Auepora

Retrieval Targets

Focus: Evaluable outputs are the *relevant documents* a system retrieves.

- **Relevance (Docs** $\leftrightarrow$ Query): Measures precision and specificity; checks how well retrieved docs match the user's needs.
- **Accuracy (Docs** $\leftrightarrow$ Candidate Docs): Assesses correctness of the retrieval among a pool of candidate documents, ensuring higher ranking for relevant items.

Generation Targets

Focus: Outputs are the generated text or structured responses.

- **Relevance (Response** $\leftrightarrow$ Query): Checks alignment with the query's topic and specific requirements.
- **Faithfulness (Response** $\leftrightarrow$ Relevant Docs): Evaluates consistency between the generated content and the retrieved documents.
- **Correctness (Response** $\leftrightarrow$ Sample Response): Gauges factual accuracy compared to a reference or ground-truth answer.

Tools & Benchmarks

Various frameworks emphasize different RAG aspects.

- **Retrieval Focus:** RAGAs, ARES measure doc relevance.
- **Generation Focus:** RGB, MultiHop-RAG prioritize correctness of responses.
- **Mixed Focus:** Some tools (e.g., Databricks Eval) and benchmarks (e.g., DomainRAG) assess both retrieval outputs and final answers.

Evaluation Dataset (How to Evaluate?)

Dataset Strategies & Existing Resources

Key Observations:

- **Variety of Approaches:** Benchmarks use both well-known datasets and newly generated data.
- **Common Choices:**
  - KILT-based benchmarks (e.g., Natural Questions, HotpotQA, FEVER).
  - SuperGLUE subsets (e.g., MultiRC, ReCoRD).
- **Limitation:** Datasets from static resources may not address dynamic, real-world information changes.

Dynamic & Custom Datasets

- **LLMs for Dataset Construction:**
  - Authors can design queries and ground truths for specific evaluation targets.
  - Some benchmarks use online news to test a system's adaptability to real-world information.
- **DomainRAG:** Combines single-doc, multi-doc, single- and multi-round QA from annually updated college websites.

Evaluation Metric (How to Quantify?)

Evaluation Metrics

Metrics translate qualitative goals (e.g., relevance, correctness, faithfulness) into measurable criteria.

- **Complex Landscape:** Aligning metrics with human preferences is challenging.
- **Component-Specific:** Retrieval and generation each require tailored evaluations.
- **Practical Factors:** Metrics should reflect real-world usage and address system robustness.

Retrieval Metrics

Key Focus: Quantify how effectively the system fetches relevant information from a vast or dynamic corpus.

- **Relevance, Accuracy, Diversity, Robustness:** Reflects precision in fetching pertinent documents and resilience to misleading or evolving data.
- **Specialized Benchmarks:** Some frameworks introduce custom metrics (e.g., Misleading Rate, Error Detection Rate) to capture nuances of real-world contexts.
- **LLMs as Judges:** Certain approaches use large language models themselves to evaluate retrieval quality.

Non-Rank Based Retrieval Metrics

- **Accuracy:** Proportion of true results (positives and negatives) among all examined cases.
- **Precision:** Fraction of retrieved documents that are relevant.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall@k:** Fraction of relevant items successfully retrieved within the top-k results.

$$\text{Recall@k} = \frac{|RD \cap TopK|}{|RD|}$$

Rank-Based Retrieval Metrics

- **Mean Reciprocal Rank (MRR):** Averages the reciprocal of the rank position of the first correct result across multiple queries.

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$

- **Mean Average Precision (MAP):** Averages the average precision scores for each query.

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} AP_q = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\sum_{k=1}^{n} \left(\mathrm{P@k} \times \mathrm{rel}(k)\right)}{|\text{relevant docs}_q|}$$

Example: Precision Calculation

Scenario: A retrieval system searches a corpus of 100 documents. Among them:

- 20 documents are relevant (**ground truth**).
- The system retrieves 15 documents, out of which 12 are relevant.

Calculations:

Precision: Fraction of retrieved documents that are relevant.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{TP} + \text{FP}}$$

Substituting values:

$$\text{Precision} = \frac{12}{12+3} = \frac{12}{15} = 0.8$$

Example: Recall Calculation

Scenario: A retrieval system searches a corpus of 100 documents. Among them:

- 20 documents are relevant (**ground truth**).
- The system retrieves 15 documents, out of which 12 are relevant.

Calculations:

Recall: Fraction of relevant documents successfully retrieved.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{Total Relevant Documents}}$$

Substituting values:

$$\text{Recall} = \frac{12}{20} = 0.6$$

Example: Mean Reciprocal Rank (MRR) Calculation

Scenario: A system is queried three times, producing ranked results. The rank position of the first correct result for each query is:

- Query 1: Correct result at rank 2.
- Query 2: Correct result at rank 1.
- Query 3: Correct result at rank 4.

Calculation:

- MRR formula:

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$

  Substituting values:

$$\mathrm{MRR} = \frac{1}{3} \left(\frac{1}{2} + \frac{1}{1} + \frac{1}{4} \right) = \frac{1}{3} \times 1.75 = 0.583$$

Example: Average Precision (AP) Calculation

Scenario: A system is queried three times, producing ranked results. Each query retrieves exactly 4 documents. The relevant documents and the retrieved results are:

- Query 1: Relevant = {D2, D4}; Retrieved = [D1, D2, D3, D4]
- Query 2: Relevant = {D3, D4}; Retrieved = [D3, D1, D2, D4]
- Query 3: Relevant = {D1, D5}; Retrieved = [D1, D2, D5, D3]

Step 1: Calculate Average Precision (AP) for Each Query

- **AP formula (for a single query):**

$$AP = \frac{\sum_{k=1}^{n} \left(\mathrm{P@k} \times \mathrm{rel}(k)\right)}{|\text{relevant docs}|}$$

- **Query 1:**

$$\mathrm{P@2} = \frac{1}{2}, \quad \mathrm{P@4} = \frac{2}{4}$$
$$\mathrm{AP}_1 = \frac{1}{2}\left(\mathrm{P@2} + \mathrm{P@4}\right) = \frac{0.5 + 0.5}{2} = 0.5$$

Example: Average Precision (AP) Calculation

Scenario: A system is queried three times, producing ranked results. Each query retrieves exactly 4 documents. The relevant documents and the retrieved results are:

- Query 1: Relevant = {D2, D4}; Retrieved = [D1, D2, D3, D4]
- Query 2: Relevant = {D3, D4}; Retrieved = [D3, D1, D2, D4]
- Query 3: Relevant = {D1, D5}; Retrieved = [D1, D2, D5, D3]

Step 1: Calculate Average Precision (AP) for Each Query

- **Query 2:**

$$\mathrm{P@1} = 1.0, \quad \mathrm{P@4} = \frac{2}{4} = 0.5$$

$$\mathrm{AP}_2 = \frac{1}{2}(\mathrm{P@1} + \mathrm{P@4}) = \frac{1.0 + 0.5}{2} = 0.75$$

- **Query 3:**

$$\mathrm{P@1} = 1, \quad \mathrm{P@3} = \frac{2}{3} = 0.67$$

$$\mathrm{AP}_3 = \frac{1}{2}(\mathrm{P@1} + \mathrm{P@3}) = \frac{1.0 + 0.67}{2} = 0.833$$

Example: Mean Average Precision (MAP) Calculation

Step 2: Calculate Mean Average Precision (MAP)

- Formula:

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \mathrm{AP}_q$$

- Substituting values:

$$\mathrm{MAP} = \frac{1}{3}(0.5 + 0.75 + 0.833) = \frac{2.083}{3} = 0.694$$

Generation Metrics

- **Beyond Accuracy:** Evaluating generated responses requires assessing coherence, relevance, fluency, and user alignment.
- **Traditional Metrics:** BLEU, ROUGE, and F1 Score remain cornerstone methods, highlighting precision and recall.
- **Newer Approaches:** Metrics like Misleading Rate, Mistake Reappearance Rate, and Error Detection Rate address RAG-specific challenges.
- **Evolving Landscape:** Growing emphasis on *factual correctness, readability, and user satisfaction* in real-world scenarios.

Human Evaluation & LLM as Judge

- **Human Evaluation:**
  - Provides a gold standard comparison for model outputs.
  - Captures subjective qualities (e.g., fluency, clarity) not always quantifiable by automated metrics.
- **LLM as an Evaluative Judge:**
  - Versatile, automatic method that can assess outputs in context where reference answers may be missing.
  - Uses prediction-powered inference (PPI) and context relevance scoring to gauge text quality.
  - Detailed prompt templates standardize evaluation (scale from 1 to 5, etc.), aligning with human preferences.

Traditional Metrics: ROUGE & BLEU

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Compares system-generated text with human-generated reference summaries.
- Variants measure n-gram overlap (ROUGE-N), longest common subsequence (ROUGE-L), and more.
- Indicates content overlap, but does not fully capture readability or fluency.

BLEU (Bilingual Evaluation Understudy)

- Evaluates machine-translated text by comparing n-gram precision against reference.
- Includes a brevity penalty to discourage overly short outputs.
- Can miss nuances of fluency and grammar.

Modern Metrics: BertScore & LLM as a Judge

BertScore

- Leverages contextual embeddings to evaluate semantic similarity.
- Generates precision, recall, and F1 scores at the token level.
- More robust to paraphrasing and meaning shifts than n-gram methods.

LLM as a Judge

- Uses large language models to rate responses on coherence, relevance, fluency, and more.
- Often operates in zero-shot or few-shot settings, or with fine-tuning on human annotations.
- Reduces reliance on strict reference comparisons, capturing deeper context alignment.

Additional Metrics & Methods

Latency

- Time from query to final response (mean or percentile).
- Critical in interactive systems (e.g., chatbots).

Diversity

- Evaluates variety of retrieved or generated content.
- Lower cosine similarity among outputs $\rightarrow$ higher diversity.

Noise Robustness

- Ability to maintain quality despite misleading or irrelevant info.
- Measured by error or misleading rates in responses.

Negative Rejection

- System refrains from answering when data is insufficient.
- Rejection rate indicates prudence in uncertain contexts.

Counterfactual Robustness

- Detection of incorrect or contradictory text.
- Error detection rate for flagged counterfactuals.

Significance in Real-World Scenarios

- **User-Centric Design:** Faster response times, diverse outputs, and accurate handling of uncertain data enhance user trust.
- **System Reliability:** Noise resilience and counterfactual detection mitigate the risk of misinformation.
- **Holistic Evaluation:** These additional metrics complement retrieval and generation scores, offering a comprehensive view of RAG performance.

Discussion

QA Datasets vs. RAG Complexity

- **Traditional QA Format:**
  - Commonly used to verify RAG capabilities.
  - Strong LLMs can solve QA tasks effectively, obscuring the impact of retrieval.
- **Need for Specialized Benchmarks:**
  - Multi-hop or multi-document queries.
  - Single- to multi-round dialogues.
  - Structural outputs, content moderation, and hallucination checks.
- **Additional Requirements:**
  - Noisy documents, latency, and diverse outputs.
  - Emphasis on real-world challenges beyond simple question-answer pairs.

Dataset Construction & Evolution

- **Tailored Datasets:**
  - Custom-built for specific RAG targets (e.g., news, structured databases).
  - Increased development overhead but allow thorough evaluation.
- **Dynamic Updates:**
  - Automated QA pair generation on a daily or frequent basis.
  - Prevents "cheating" by LLMs and tests system adaptability.
- **Diverse Sources:**
  - From Wikipedia expansions to domain-specific content.
  - Reflects broad scenarios where RAG must adapt to changing data.

Metrics, LLM Judges, & Resource Constraints

- **LLM as Judge:**
  - Offers deeper reasoning but requires consistent prompts and scoring scales.
  - Challenges in aligning automated ratings with human judgment.
- **Standardization Gaps:**
  - No universal reference or grading for LLM-based evaluation.
  - Human preferences differ across tasks, making uniform guidelines elusive.
- **Resource Constraints:**
  - Running large-scale LLMs is costly and time-intensive.
  - Need for compact yet reliable evaluation methods that balance thoroughness and feasibility.

Conclusion

Conclusion

- **Holistic Evaluation:** RAG systems require dedicated benchmarks that capture retrieval accuracy, generative quality, and real-world practicalities.
- **Auepora Framework:** Targets, datasets, and metrics form a structured lens for RAG assessment, highlighting both technical and user-centric dimensions.
- **Remaining Gaps:** Current methods often overlook dynamic data, nuanced user needs, and robust ways of measuring noise handling and factual correctness.

References

References

Qiu, W., Dong, M., Long, Y., Long, Q., He, K., Pan, H., & Qin, Z. (2023).
Chain-of-Domains for Natural Language Interfaces.
https://arxiv.org/abs/2405.07437

Q&A

Thank you!

Questions?

Take the Quiz!

Scan the QR code below to participate.