

# Classifying and Addressing the Diversity of Errors in Retrieval-Augmented Generation Systems

Kin Kwan Leung Mouloud Belbahri Yi Sui Alex Labach  
Xueying Zhang Stephen Anthony Rose Jesse C. Cresswell

Layer 6 AI, Toronto, Canada

{kk, mouloud, amy, alex, chloe, stephen, jesse}@layer6.ai

## Abstract

Retrieval-augmented generation (RAG) is a prevalent approach for building LLM-based question-answering systems that can take advantage of external knowledge databases. Due to the complexity of real-world RAG systems, there are many potential causes for erroneous outputs. Understanding the range of errors that can occur in practice is crucial for robust deployment. We present a new taxonomy of the error types that can occur in realistic RAG systems, examples of each, and practical advice for addressing them. Additionally, we curate a dataset of erroneous RAG responses annotated by error types. We then propose an auto-evaluation method aligned with our taxonomy that can be used in practice to track and address errors during development. Code and data are available at <https://github.com/layer6ai-labs/rag-error-classification>.

## 1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020) has become the dominant paradigm for applying generative large language models (LLMs) in applications where outputs must incorporate knowledge from outside of the model’s training set. This is especially valuable for grounding generation in factual information to reduce fabricated content (Maynez et al., 2020; Shuster et al., 2021), and when non-public domain knowledge is required. RAG systems are already widely deployed in real-world applications to provide natural language interfaces to knowledge sources (Amugongo et al., 2024), but despite their merits they can still be error-prone in practice (Venkit et al., 2024; Magesh et al., 2025; Grant, 2024). Due to the greater complexity of RAG pipelines compared to direct LLM generation, these errors are diverse and their causes can be difficult to trace. Deploying a RAG pipeline, especially in critical industries like healthcare, re-

quires understanding the variety of errors that can occur in order to monitor and minimize them.

Existing work on RAG errors has generally not accounted for the complexity of real-world RAG systems and their failure modes. On the data side, widely used benchmark tasks for evaluating RAG systems are often overly simplistic, typically featuring multiple-choice questions (Mihaylov et al., 2018; Guinet et al., 2024; Yang et al., 2024) or requiring short factual answers for ease of validation (e.g., Rajpurkar et al. (2016); Yang et al. (2018); Joshi et al. (2017); Kwiatkowski et al. (2019)). In practice, users expect more sophisticated answers that thoroughly explain a topic, but as a result can fail in more subtle ways.

On the model side, real RAG systems use complex multi-step pipelines that go beyond a retriever-generator pair (Akkiraju et al., 2024), with additional data processing steps like adaptive chunking and reranking. Prior works that categorize RAG errors have used bare-bones pipelines benchmarked on simple datasets, and as a result have overlooked entire classes of errors, especially those associated with pre-generation steps (Barnett et al., 2024; Venkit et al., 2024). As a result, practitioners relying on these works may be left with an overly optimistic view of their RAG system’s performance, and fail to identify the scope and severity of errors afflicting them. To address these gaps, we use challenging public datasets and build a realistic RAG question-answering (QA) system reflective of those currently used in industry. We perform a deep analysis of RAG failure modes with illustrative examples, and practical advice for mitigating errors.

Finally, we develop auto-evaluation tools to classify error types according to our taxonomy. To validate these systems, we manually annotate errors producing a first-of-its-kind RAG error type dataset. The overall purpose of our taxonomy and auto-evaluation system is for practitioners to be able to identify weak links and common errors in

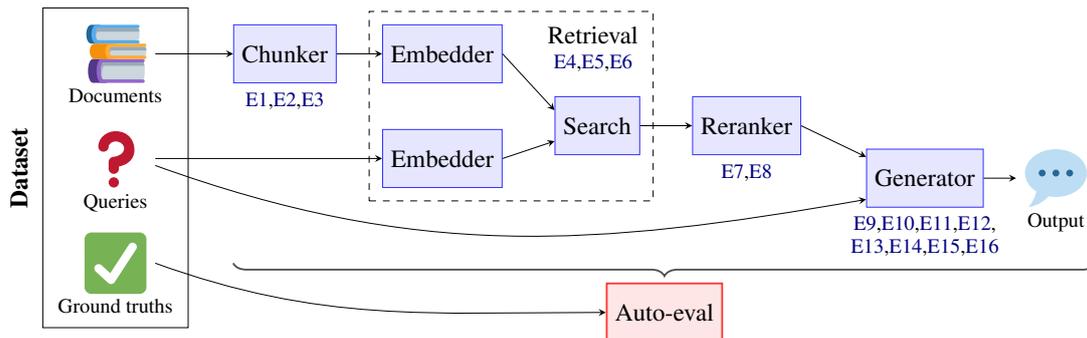


Figure 1: Diagram of our implemented RAG architecture, reflective of systems currently applied in industry. Components are annotated with error types that are caused by them.

their RAG pipelines. Hence, we provide an end-to-end demonstration of how auto-evaluation can be used by identifying the most common errors produced by our reference pipeline, and implementing targeted improvements.

Our main contributions are:

1. A novel taxonomy of errors made by RAG systems, along with examples of each and recommendations for addressing them;
2. An auto-evaluation system for identifying and classifying errors according to our taxonomy;
3. A dataset of RAG errors annotated by type.

## 2 Related Works

**RAG Error Taxonomies:** Barnett et al. (2024) provide the most comparable taxonomy of RAG errors to ours with seven types associated to select RAG components. Here, we describe a more comprehensive range of RAG errors relating to *every* step of a realistic RAG pipeline, along with specific examples and mitigation strategies, both of which are lacking in the prior work. Venkit et al. (2024) classify errors in public black-box QA systems from the perspectives of human evaluators without examining specific pipeline components or attributing errors to them. Agrawal et al. (2024) focus on RAG systems built on knowledge graphs. Yu et al. (2024a) survey RAG evaluation benchmarks and metrics but do not break down causes of errors beyond the high-level dichotomy of retrieval versus generation. Other works (Huang et al., 2025; Magesh et al., 2025) that specifically focus on hallucinations overlook the nuanced ways RAG can fail. Error taxonomies for non-RAG QA settings also exist (Rawte et al., 2023; Huang et al., 2025).

**RAG Auto-evaluation:** Simple QA benchmarks can be evaluated using exact match or overlap-based metrics, but realistic questions and answers

require more sophisticated evaluations based in natural language understanding. Previous works have applied metrics using LLM-as-a-Judge (Zheng et al., 2023) to automatically evaluate RAG outputs, including RAGChecker (Ru et al., 2024), ARES (Saad-Falcon et al., 2024), and RAGAs (Es et al., 2024). These works are each tailored to a specific error classification scheme, with narrower scope than ours. Other prior works have developed benchmark datasets with LLM-based auto-evaluation (e.g., Zhu et al. (2024); Liu et al. (2024)), but do not probe the internals of RAG systems to give detailed information on error causes.

## 3 RAG Implementation

To demonstrate errors that can occur in realistic, production-grade RAG systems, we designed a modular pipeline reflecting common architectural patterns used in industry (Bleiweiss, 2024; Alexander et al., 2025) (see Figure 1). We begin with a **chunking** stage, where documents are processed to optimize granularity and relevance. The chunks are indexed using an embedding model, then queried via dense **retrieval**, which returns top candidates based on similarity to the input query. Retrieved chunks are **reranked** using a separate language model given the original query as context, which can better highlight semantic relevance. Finally, the top-ranked chunks are fed to a **generator** LLM, which produces the final answer. This architecture supports component-level variation and allows for in-depth error analysis at each stage of the pipeline. We provide complete details on the options implemented and used at each stage in Appendix A.

**Datasets:** We use the DragonBall dataset, part of the RAGEval framework (Zhu et al., 2024), an evaluation suite designed to rigorously assess RAG systems across diverse domains and scenarios. The

dataset provides a broad benchmark for evaluating RAG performance in complex settings, spanning domains such as finance, law, and medicine. The dataset also comes with query type annotations in seven categories, which allows us to perform thorough analysis on error classification. We used both the English and Chinese portions of the dataset with 3108 and 3601 questions respectively. We also use the CLAPnq dataset for additional analysis (Rosenthal et al., 2025), which contains high-quality reference answers and passages for 4946 natural questions relating to Wikipedia documents.

## 4 Error Classification

We present a practically grounded taxonomy of errors that can be attributed to different parts of a RAG system. Errors are grouped by the pipeline stage that caused them: chunking, retrieval, reranking, or generation. For each stage and error type, we describe the nature of the error, explain why it occurs in RAG systems, provide real examples from our RAG implementation, and finally give advice for how to reduce their occurrence.

Error types are not mutually exclusive; in practice multiple error types often co-occur, as errors early in the pipeline beget later ones. Additionally, we do not claim that our taxonomy is exhaustive, since variations in RAG architecture will lead to different error types. Our taxonomy focuses on errors that occur within RAG systems, and hence we exclude failures caused by adversarial inputs, faults in the corpus, or similar anomalies.

### 4.1 Chunking

**E1 Overchunking:** Documents are split into excessively small or disjointed segments, causing incomplete coverage of topics. Individual chunks are fragmented or ambiguous. Errors cascade downstream when search fails to retrieve consecutive chunks.

**Query:** What platform did Sunrise Holidays introduce in April 2019?

**Chunk Excerpt:** *[Excluded from chunk: In April 2019, Sunrise Holidays introduced an **online booking platform**, which greatly improved its competitiveness.]* The launch of this user-friendly platform attracted more customers [...]

**Response:** The platform introduced by Sunrise Holidays in April 2019 is not specified.

**Ground Truth:** **An online booking platform.**

**Cause:** The retrieved chunks come from the correct document, but the description of it as an *on-*

*line booking* platform is cut off.

**E2 Underchunking:** Chunks are too large, covering multiple topics with mixed content. Irrelevant information dilutes keywords or phrases, lowering retrieval scores on the correct chunks. Chunks provided to the generator contain extraneous content that can confuse the model.

**Query:** What system does CleanCo Housekeeping Services have in place to safeguard assets and ensure financial accuracy?

**Chunk Excerpt:** [...] To mitigate risks such as increasing competition, regulatory changes, and economic uncertainties, CleanCo plans to implement **risk management** strategies through diversification and continuous monitoring. [...]

**Response:** CleanCo has a system in place to safeguard assets and ensure financial accuracy through its **risk management** framework.

**Ground Truth:** Unable to answer.

**Cause:** Retrieved chunks contain so much tangential information that the generator uses unrelated information to answer instead of abstaining.

**E3 Context Mismatch:** Chunks split text at arbitrary points, breaking contextual links by separating definitions from the information they support. This ambiguity causes failed retrieval downstream where keywords are missing.

**Query:** Why is **the Philippines** typhoon prone?

**Chunk A (Retrieved):** **The Philippines'** evident risk to natural disasters is due to its location. [...]

**Chunk B (Not Retrieved):** In addition, **the country** faces the Pacific Ocean where 60% of the world's typhoons are made. [...]

**Response:** Due to its geographical location, climate, and topography.

**Ground Truth:** It faces the Pacific Ocean where 60% of the world's typhoons are made.

**Cause:** Chunk B mentions "**the country**" but not **the Philippines** by name, which led to the retriever incorrectly assigning it low relevance.

**Improvement:** Straightforward heuristics include adjusting chunk size, where larger chunks help reduce E1, smaller chunks help reduce E2, while adding small overlaps mitigates E3 by preserving continuity (Safjan, 2023). However, these approaches require careful tuning and are often insufficient on their own. Adaptive chunking strategies handle real-world variability more effectively.

Structure-aware chunking preserves document

logic by splitting along units like paragraphs or section headers. This works especially well for corpora with consistent structure, such as financial reports (Yepes et al., 2024). Semantic chunking segments text based on meaning, such as by detecting topic shifts using cosine distance between sentence embeddings (Qu et al., 2025) or spikes in perplexity (Zhao et al., 2024). Hybrid strategies combine structural and semantic cues to balance coherence and topical focus. For example, S2 Chunking (Verma, 2025) integrates layout with embeddings and achieves strong results on document QA.

## 4.2 Retrieval

**E4 Missed Retrieval:** Relevant chunks are not retrieved, leading the generator to give incomplete answers, fabricate information to fill gaps, or abstain unnecessarily. Misses can occur when high-similarity chunks only address part of the question.

**Query:** What were the large-scale financing activities undertaken by Vanguard Media Group in 2019 and Energex Corp. in 2021?

**Response:** Vanguard Media Group’s large-scale financing activities in 2019 included the acquisition of Silver Screen Studios for \$200 million [...]

**Ground Truth:** Vanguard Media Group raised \$100 million in capital in 2019, while Energex Corp. conducted large-scale financing in 2021.

**Cause:** The query asks about two entities, but all top-ranked chunks relate to only one.

**E5 Low Relevance:** Retrieved chunks are only loosely related to the query. Causes include an unbalanced corpus with many chunks relating to one topic, and vocabulary mismatch where the query uses different terms than the corpus.

**Query:** In which year is the Corporate Governance Report for Green Fields Agriculture Co.?

**Response:** The Corporate Governance Report does not provide a specific year, but it appears to cover events and initiatives from 2018 and 2021.

**Ground Truth:** 2018.

**Cause:** Many chunks from the Report were assigned high similarity and retrieved, but did not relate to the Report’s publication year. The generator latches onto years discussed in the content.

**E6 Semantic Drift:** Retrieved documents match keywords, not the query’s intent, due to reliance on keyword matching rather than semantic relevance.

**Query:** According to the judgment of Hartford, Ashland, Court and Glenwood, Quailwood, Court, what sentences were given to the defendants M. Cooper and Y. Nelson?

**Response:** M. Cooper was sentenced to four years imprisonment.

**Ground Truth:** M. Cooper was sentenced to four years imprisonment; Y. Nelson was sentenced to three years imprisonment.

**Cause:** The retrieved chunks contain information about both defendants’ cases, but don’t include the sentence for Y. Nelson.

**Improvements:** Retrieval systems present the widest diversity of approaches out of the RAG stages, and hence can be tuned in many ways. First, query rewriting via keyword expansion, paraphrasing, or semantic reformulation can improve recall and precision by reducing vocabulary mismatch (Ma et al., 2023). Second, hybrid retrieval that combines sparse (e.g., BM25 (Robertson et al., 1994)) and dense (embedding-based) methods can improve robustness by capturing both lexical and semantic matches (Ni et al., 2022). Third, top-*k* retrieval is often insufficient when both E4 and E5 are frequent. Adaptive thresholds can better balance recall and precision (Sun et al., 2025). Fourth, domain-specific or fine-tuned embedding models can significantly boost retrieval accuracy and reduce hallucinations (Databricks, 2025). Finally, metadata filtering enhances precision by using structured fields (e.g., source or section) to prioritize relevant content (Poliakov and Shvai, 2025).

## 4.3 Reranking

**E7 Low Recall:** Although the necessary chunks are retrieved, they are reranked too low and not forwarded to the generation model. Low recall produces false negatives, where relevant data is available but unintentionally ignored.

**Query:** According to Greenfield, Fairview, Court and Vandalia, Bayside, Court, whose sentencing time is longer, H. Price or G. Torres?

**Response:** G. Torres’ sentencing time is longer, four years, compared to H. Price, whose sentence is not mentioned in the provided context.

**Ground Truth:** H. Price was sentenced to eight years of imprisonment, while G. Torres was sentenced to four years imprisonment.

**Cause:** The chunk with H. Price’s sentencing was retrieved, but was ranked below the reranker

threshold and not passed to the generator.

**E8 Low Precision:** Irrelevant chunks are ranked highly and forwarded to the generation model, leading to the generation model being confused by noise. Low precision leads to false positives, where the generator is overly confident in noisy context.

**Query:** On which stock exchange is Retail Emporium publicly traded?

**Response:** The **New York** Stock Exchange.

**Ground Truth:** Unable to answer.

**Cause:** Highly ranked but irrelevant chunks reference **New York**, leading the model to make an inappropriate conclusion.

**Improvements:** Modern RAG pipelines increasingly include a reranking stage to reorder retrieved chunks based on semantic relevance (Bleiweiss, 2024). The two dominant approaches are cross-encoders, which jointly encode query-chunk pairs (Déjean et al., 2024), and LLM-based rerankers, which use prompted models to assess candidates (Sun et al., 2023). Rerankers may still miss key context (E7) or over-rank irrelevant chunks (E8) when domain relevance is poorly understood. Fine-tuning on in-domain data is the most effective way to improve performance (Mora et al., 2024; Aarsen, 2025; Yu et al., 2024b), and even small amounts of labeled data can yield significant gains (Aarsen, 2025). More specifically, incorporating hard negatives, which are plausible but non-relevant passages, further helps the model reduce false positives (E8) (Yu et al., 2024b).

#### 4.4 Generation

**E9 Abstention Failure:** The model should have abstained, but instead answers incorrectly. Abstention is appropriate when the answer is not available in the corpus, or when the query is ambiguous.

**Query:** How much did Green Fields Co. raise through large-scale financing in December 2018?

**Response:** Green Fields Co. did not raise any funds in December 2018.

**Ground Truth:** Unable to answer

**Cause:** The retrieved context contains information about 2018, but not about fundraising. Hence, the model assumes fundraising did not occur, but this was not stated in the corpus.

**E10 Fabricated Content:** Although the query is answerable, the response includes unverifiable information not grounded in the retrieved chunks nor

supported by external evidence.

**Query:** According to the hospitalization records of Southport General Hospital, summarize the past medical history of K. Mendoza.

**Response:** K. Mendoza has a history of **Parkinson's disease**, with symptoms including tremors, bradykinesia, and postural instability.

**Ground Truth:** K. Mendoza was **generally healthy** prior to the onset of his symptoms.

**Cause:** The model fabricates a history of **Parkinson's disease** which is not in the context.

**E11 Parametric Overreliance:** The LLM relies on its internal (parametric) knowledge rather than retrieved documents. Internal knowledge may be factual, as opposed to fabricated, but irrelevant.

**Query:** What was the total amount of liabilities of Artistic Creations at the end of 2019?

**Response:** The total amount of liabilities of Artistic Creations at the end of 2019 is \$150M - \$20M (acquisition) - \$50M (financing) - \$10M (investment) + \$15M (restructuring) = \$85M.

**Ground Truth:** Unable to answer.

**Cause:** The model does a calculation based on parametric financial knowledge instead of acknowledging that it has insufficient information.

**E12 Incomplete Answer:** The response is from the corpus and correct, but misses critical details.

**Query:** How did the dividend in July 2017 reflect AMS Inc.'s **strong financial performance**?

**Response:** The dividend in July 2017 reflected AMS Inc.'s **strong financial performance**, which was driven by market demand and changes in product prices, resulting in an operating income of \$450 million and a net profit of \$60 million.

**Ground Truth:** The dividend demonstrated the company's **ability to generate sufficient profits and maintain a healthy cash flow**.

**Cause:** The response affirms the basis of the question, but does not analyze the situation.

**E13 Misinterpretation:** The generator misuses or misrepresents retrieved content.

**Query:** According to the judgments of Trenton, Springfield and Hamilton, Harrison Courts what sentences were given to **Q. Ruiz** and **M. Ward**?

**Response:** **M. Ward** was sentenced to four years imprisonment, and **Q. Ruiz** has the right to appeal this judgment within 30 days.

**Ground Truth:** M. Ward was sentenced to four years of imprisonment, and Q. Ruiz was sentenced to 5 years of imprisonment.

**Cause:** The model misinterprets facts about two defendants as being related and incorrectly combines them for the answer.

**E14 Contextual Misalignment:** The response is factual and comes from related information in the corpus, but does not address the query.

**Query:** According to the judgment of Preston, Lancaster, Court, summarize the facts of M. Harris's crimes.

**Response:** M. Harris was found guilty of homicide and was sentenced to life imprisonment.

**Ground Truth:** M. Harris committed the crime of intentional homicide by tampering with the brakes of her business partner.

**Cause:** The response contains factual information, but discusses sentences rather than the requested facts of the crimes.

**E15 Chronological Inconsistency:** The response presents events or facts in the wrong temporal order, or confuses past, present, or future timelines.

**Query:** Which company, Green Fields or JetWing, changed their board of directors earlier?

**Response:** Green Fields made changes to their board of directors earlier, in December 2021, while JetWing made changes in October 2021.

**Ground Truth:** JetWing made changes earlier in September 2021, while Green Fields made changes in December 2021.

**Cause:** The model provides a wrong date, and incorrectly considers Dec. earlier than Oct.

**E16 Numerical Error:** The response includes incorrect calculations, quantities, or misrepresents numerical data from the retrieved documents.

**Query:** Which company reduced its liabilities more, InnovateTech or Sparkling Clean?

**Response:** InnovateTech reduced its liabilities by a larger amount, a total of \$50 million, compared to Sparkling Clean's £500,000.

**Ground Truth:** InnovateTech reduced its liabilities by a larger amount (\$30 million) compared to Sparkling Clean (£500,000).

**Cause:** The model incorrectly added together two numbers from the context giving \$50 million.

**Improvements:** Generation-stage errors can be mitigated through abstention modeling, answer val-

idation, prompt engineering, and tool augmentation. For abstention failures (E9), models can explicitly use ambiguity detection (Cole et al., 2023; Kim et al., 2023), conformal abstention (Yadkori et al., 2024), or fine-tuning approaches like Trust-Align (Song et al., 2025). For content fabrication (E10), post-generation validation such as Chain-of-Verification (Dhuliawala et al., 2024) or critique modules (Asai et al., 2023) encourage fact-checking and grounding. Errors related to context misuse (E11-E14) often stem from noisy retrieval; refinement modules can filter context before generation (Jin et al., 2025; Chirkova et al., 2025; Wang et al., 2025). Query decomposition (Lin et al., 2023) helps reduce E12 by breaking complex queries into sub-questions, especially for comparisons, causal questions, or multi-step reasoning. For temporal and numerical errors (E15-E16), structured prompting (Wei et al., 2022), timeline reasoning (Bazaga et al., 2025), or tool-augmented generation with code modules like PAL (Gao et al., 2023) can boost reliability.

## 5 RAGEC: RAG Error Classification

To deepen our understanding of the RAG error taxonomy and aid practitioners in applying it, we developed a RAG Error Classification system (RAGEC). In this section we describe its design, curate a dataset with human annotations to validate it, and demonstrate its use via our reference RAG system with standard research datasets.

The primary objective of RAGEC is to identify weak links in the RAG pipeline by classifying the *stage* responsible for errors. More granular error *types*, which can co-occur and overlap, are classified secondarily. Identifying the *first* stage where an error occurred enables the developer to implement targeted improvements on the RAG pipeline.

**Design.** Given a RAG pipeline and dataset, RAGEC consists of 3 steps: answer evaluation, stage classification, and error type classification. For *answer evaluation*, we prompt an LLM on each datapoint individually to determine whether the generated answer is incorrect. Included as context are the original query, ground-truth answer and documents according to the dataset, and the generated answer. This evaluation yields a subset of  $N_{\text{err}}$  examples identified as incorrect. For *stage classification* on an erroneous example, we apply a rules-based approach to cascade over the stages, using specialized information to determine if an error

Table 1: Stage classification agreement matrix

Human Annotation	RAGEC Stage Classification			
	Chunking	Retrieval	Reranking	Generation
Chunking	49	13	5	10
Retrieval	24	85	12	40
Reranking	6	8	22	10
Generation	4	21	6	62

is present. Starting with the generation stage, we check whether the generator had sufficient information to answer the query by computing the chunk-level recall and comparing to a threshold. If so, we conclude the generation stage caused the error. Otherwise, we continue and check if the reranker recall dropped compared to the retrieval recall, indicating that the reranker caused the error. If not, we move to the final comparison between retrieval and chunking. To differentiate these two we adapt the idea from Matton et al. (2025) to extract key concepts in the query. We then prompt the LLM as to whether each concept is included in the ground truth chunks. If query concepts fail to appear in the chunks, it indicates the chunking stage caused the error. Otherwise, by process of elimination, we conclude the retriever was responsible. This design is empirically motivated, and provided the highest alignment with our human-annotated results. Other designs we tested are described in Appendix B.4.

For *error type classification*, we take information relevant to the identified stage, and conduct detailed analysis using LLM-as-a-Judge (Zheng et al., 2023). For example, on a chunking error we only provide the query, ground truth answer, and chunks from the ground truth document. Other information, like the retrieved chunks or generated answer are irrelevant for classifying a chunking error type. To account for uncertainty in the classification, each erroneous example is annotated  $K$  times (Wang et al., 2023). We then compute the modal vote over error types and metrics that capture the self-consistency of the LLM’s predictions.

Further details on the implementation of RAGEC, including the complete methodologies, inputs used for each step, and LLM prompts, are included in Appendix B.

**Data Curation.** No existing dataset details the types of errors which can occur in RAG systems. Hence, to validate RAGEC, we manually annotated 406 erroneous responses generated by our reference system on the DragonBall dataset as to the responsible stage, and error types present. Annotators

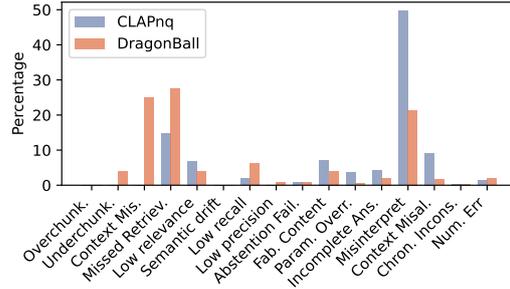


Figure 2: Modal error type distribution per dataset.

used all available context including ground truth answer, generated response, retrieved and reranked chunks, and the corpus. First, annotators selected the earliest stage in the pipeline which demonstrated erroneous behaviour, since errors propagate through RAG systems. Subsequently, one or more error types were selected within that stage. The repo at <https://github.com/layer6ai-labs/rag-error-classification> contains our annotated data.

**Results & Analysis.** In the first step of RAGEC, *answer evaluation*, 27% (25%) of responses for the DragonBall-EN (CLAPnq) dataset were classified as incorrect, giving  $N_{\text{err}} = 832$  (1222) responses to analyze. Of these, 406 queries were manually annotated, and 377 were confirmed as erroneous, reaching an agreement rate of 92.9%.

For *stage classification*, RAGEC achieves a 57.8% agreement rate with human annotations. Details of the results are shown in Table 1. We find that human annotations more often identify the retrieval stage as causing errors, whereas RAGEC favors blaming the generator. While RAGEC’s agreement rate is not close to perfect, we tested several other methods to engineer the context for LLM-based auto-evaluation that all performed worse (described in Appendix B.4). Hence, RAG error classification appears to be a challenging problem due to the highly complex set of intermediate outputs produced by RAG pipelines. We encourage the community to further explore this topic by releasing our reference pipeline and annotated error data.

For *error type classification*, we query the LLM  $K = 10$  times for each datapoint and extract the most and second-most frequently predicted error types. Compared to human-annotated error types, RAGEC achieves an accuracy of 40.3%. Figure 2 shows the distribution of the modal error type for DragonBall-EN and CLAPnq. Table 2 shows more details for the DragonBall-EN dataset (CLAPnq is

Table 2: Error-type distributions for DragonBall-EN. Mode and 2<sup>nd</sup> Mode show the original RAG pipeline, while Impr. shows the mode after targeted improvements to the RAG pipeline.

Error Classification	Mode	2 <sup>nd</sup> Mode	Impr.
E1 Overchunking	0	19	0
E2 Underchunking	32	92	10
E3 Context Mismatch	207	32	78
<b>Percentage Chunking</b>	<b>29.7%</b>	<b>32.2%</b>	<b>16.4%</b>
E4 Missed Retrieval	229	27	122
E5 Low Relevance	33	107	8
E6 Semantic Drift	0	7	0
<b>Percentage Retrieval</b>	<b>31.5%</b>	<b>31.8%</b>	<b>24.3%</b>
E7 Low Recall	53	5	28
E8 Low Precision	6	21	0
<b>Percentage Reranking</b>	<b>7.1%</b>	<b>5.9%</b>	<b>5.2%</b>
E9 Abstention Failure	6	0	5
E10 Fabricated Content	34	37	53
E11 Parametric Overreliance	4	6	0
E12 Incomplete Answer	17	22	26
E13 Misinterpretation	178	42	164
E14 Contextual Misalignment	14	18	8
E15 Chronological Inconsistency	3	3	13
E16 Numerical Error	16	6	19
<b>Percentage Generation</b>	<b>32.7%</b>	<b>30.2%</b>	<b>53.9%</b>
<b>Total</b>	<b>832</b>	<b>444</b>	<b>534</b>

shown in Table 6 of Appendix C.1). These distributions differ, showing that the nature of the dataset can affect the types of errors committed by a given RAG pipeline. This type of information can direct diagnosis and triaging of problems in RAG. For example, CLAPnq exhibited no chunking errors, as its source documents were manually curated by the dataset annotators into self-contained paragraphs designed to answer the corresponding queries. In contrast, DragonBall-EN almost never shows Parametric Overreliance (E11) for both English and Chinese (See Appendix C.2 for details). This can be attributed to the dataset’s domain-specific and context-dependent queries, which are unlikely to be covered by the LLM’s parametric knowledge. RAGEC also demonstrates that some error types dominate their stage, like E3 for chunking, while others like E5 primarily occur as a secondary error along with another type. Notably, *Fabricated Content* (E10)—often referred to as hallucination—is relatively rare in our analysis. This observation stands in contrast to the focus on hallucination in recent research and public discourse surrounding LLMs (Ji et al., 2023). Our findings suggest that, in the context of RAG, other types of errors such as retrieval or chunking issues are more prevalent and may warrant greater attention.

Next we demonstrate how insights can be gathered into a RAG pipeline’s behaviour through fine-

grained analysis with RAGEC. Table 3 shows the distribution of the error types across knowledge domains and query types defined in DragonBall-EN.

**Error Types across Knowledge Domains.** We see that *finance* queries lead to the most diversity in error types, but also skew more towards generation errors. Misinterpretation (E13) is the single biggest challenge. For *law* queries three of the RAG stages contribute roughly the same amount of errors, and overall law is the most difficult domain. Numerical errors (E16) are more common for law than any other domain which may reveal an underlying deficiency in the generation model around calculations that are typically done in law, as opposed to finance. *Medical* queries are by far dominated by chunking errors, with context mismatch (E3) having an outsized presence. A practitioner seeing this could simply increase chunk size, return entire documents corresponding to the top chunks, or invest in more complex solutions like expanding chunked text to include explicit definitions.

**Error Types across Query Types.** For *factual questions* three error types dominate (E3, E4, E13), whereas *unanswerable questions*, where the correct response is to abstain, show the highest diversity in error types. As expected, the abstention failure (E9) error type is used nearly exclusively for unanswerable questions. *Summarization* queries, the most difficult type, are dominated by chunking errors (E2, E3). Multiple semantically related passages are typically needed to produce a coherent and complete summary, and these connections can be lost because of chunking. *Numerical comparison* queries show fewer generation errors than other types, and numerical errors (E16) are not particularly troublesome. Errors are instead coming from context mismatch (E3) and missed retrieval (E4). Finally, as expected, *multi-hop reasoning* is one of the more challenging types and displays a wide range of error types for this RAG implementation. This range accurately reflects the more complicated nature of multi-hop reasoning queries where many things can go wrong, and demonstrates that the proposed error categories are expressive enough to capture the diversity of errors in complex query types.

**Classification Consistency.** Finally, in Table 4 we examine the consistency of classifications via mode frequency, the number of the  $K$  runs that agreed with the modal error type. For the majority of the queries, RAGEC is quite consistent in determining error categories. Appendix C.3 discusses the co-occurrence of the error categories.

Table 3: Modal error type distributions for DragonBall-EN across knowledge domains and query types. Query type labels are provided by the dataset.

Error Classification	Knowledge Domain			Query Type						
	Finance	Law	Medicine	Factual Question	Unanswerable Question	Numerical Comparison	Information Integration	Temporal Sequence	Multi-hop Reasoning	Summarization
E1 Overchunking	0	0	0	0	0	0	0	0	0	0
E2 Underchunking	2	19	11	1	3	1	0	1	7	19
E3 Context Mismatch	19	71	117	19	5	22	43	12	24	82
<b>Percentage Chunking</b>	<b>5.9%</b>	<b>32.8%</b>	<b>62.4%</b>	<b>36.4%</b>	<b>16.3%</b>	<b>29.1%</b>	<b>18%</b>	<b>14.6%</b>	<b>27%</b>	<b>49%</b>
E4 Missed Retrieval	94	76	59	22	7	29	87	35	20	29
E5 Low Relevance	25	6	2	1	0	4	20	1	3	4
E6 Semantic Drift	0	0	0	0	0	0	0	0	0	0
<b>Percentage Retrieval</b>	<b>33.7%</b>	<b>29.9%</b>	<b>29.8%</b>	<b>41.8%</b>	<b>14.3%</b>	<b>41.8%</b>	<b>44.8%</b>	<b>40.4%</b>	<b>20%</b>	<b>16%</b>
E7 Low Recall	28	23	2	0	2	3	25	10	8	5
E8 Low Precision	4	2	0	1	1	2	0	1	0	1
<b>Percentage Reranking</b>	<b>9.1%</b>	<b>9.1%</b>	<b>1%</b>	<b>1.8%</b>	<b>6.1%</b>	<b>6.3%</b>	<b>10.5%</b>	<b>12.4%</b>	<b>7%</b>	<b>2.9%</b>
E9 Abstention Failure	5	1	0	0	5	0	1	0	0	0
E10 Fabricated Content	30	4	0	0	4	2	5	0	8	15
E11 Parametric Overreliance	4	0	0	0	0	0	0	0	1	3
E12 Incomplete Answer	8	8	1	0	0	0	7	0	8	2
E13 Misinterpretation	116	50	12	10	16	11	46	29	26	40
E14 Contextual Misalignment	11	2	1	0	4	2	3	0	0	5
E15 Chronological Inconsistency	2	1	0	1	0	0	1	0	0	1
E16 Numerical Error	5	11	0	0	2	3	1	0	10	0
<b>Percentage Generation</b>	<b>51.3%</b>	<b>28.1%</b>	<b>6.8%</b>	<b>20%</b>	<b>63.3%</b>	<b>22.8%</b>	<b>26.8%</b>	<b>32.6%</b>	<b>46.1%</b>	<b>32%</b>
<b>Total Errors</b>	<b>353</b>	<b>274</b>	<b>205</b>	<b>55</b>	<b>49</b>	<b>79</b>	<b>239</b>	<b>89</b>	<b>115</b>	<b>206</b>
<b>Total Queries</b>	<b>1574</b>	<b>782</b>	<b>752</b>	<b>535</b>	<b>233</b>	<b>297</b>	<b>737</b>	<b>241</b>	<b>532</b>	<b>533</b>
<b>Error Rates</b>	<b>22.4%</b>	<b>35%</b>	<b>27.3%</b>	<b>10.3%</b>	<b>21%</b>	<b>26.6%</b>	<b>32.4%</b>	<b>36.9%</b>	<b>21.6%</b>	<b>38.6%</b>

Table 4: Distribution of mode frequency ( $K = 10$ ).

Mode Frequency	DragonBall	CLAPnq
3	1	4
4	7	32
5	41	91
6	74	123
7	74	135
8	94	149
9	153	172
10	388	516

**Summary.** RAGEC provides an entry point for debugging and improvement of RAG pipelines, which can be complex and opaque to developers. Instead of manually sifting through individual generations, our system automatically captures intermediate information like retrieved and reranked chunks, and distills the information into high-level descriptive statistics. Hence, our focus is not on numerical metrics of performance, but on understanding the types and causes of errors so that developers can intelligently prioritize aspects to improve.

**Using RAGEC to Improve RAG.** We conduct a case study on how RAGEC can be used in practice to improve a RAG pipeline. We focus on DragonBall-EN where 832 errors were originally identified (73.3% correct). RAGEC indicates that chunking, retrieval, and generation all caused roughly the same proportion of errors (Table 2, Mode), and hence are the primary candidates for improvement. Since most chunking errors come from Context Mismatch (E3) where con-

textual links are broken across chunks, we follow the recommendation in Section 4.1 and modify the chunking strategy from fixed-length segmentation to a recursive, sentence-level segmentation of approximately the same size. In addition, RAGEC indicates that most retrieval and reranking errors come from Missed Retrieval (E4) and Low Recall (E7) which are addressed by increasing the number of chunks retrieved and passed to the generator.

With these improvements, we reran the RAG pipeline and RAGEC. The improved pipeline resulted in fewer errors overall, only 534 (82.8% correct). In Table 2 (column Impr.) we observe that chunking-, retrieval-, and reranker-related errors were significantly reduced, while generator errors stay similar as no intervention was done.

## 6 Conclusion

Our contributions in this work provide a framework for practitioners working with RAG systems to understand, categorize, track, and fix model errors. Our findings also indicate the wide range of possible avenues for improving the robustness of RAG systems in future works. While our RAG error classification method, RAGEC, achieves a useful level of agreement to human error annotations, it is far from perfect, indicating that error classification remains a challenging problem for LLM-based systems. To promote future research in this area, we release our annotated dataset of error types as a resource for the community.

## Limitations

Our RAG error taxonomy is more expansive than prior works in this vein, but still is not exhaustive of all possible errors that could occur, especially if further stages were to be added to the RAG pipeline. We focused on the most prevalent types observed in practice.

The RAG pipeline we used for demonstrating auto-evaluation for error classification reflects the general architecture of modern pipelines used in real-world applications. However, it was not highly tuned for the DragonBall and CLAPnq datasets, for example by implementing the many improvements we listed in Section 4. Our aim was to show a starting point, where practitioners can understand errors in their system before making changes.

Additionally, our study focuses on single-turn textual queries. Expanding the evaluation framework to handle multi-turn conversations or multi-modal inputs remains an important direction for future work.

Finally, we found RAG error classification according to our taxonomy to be a challenging prediction problem. Part of this stems from the difficulty of collecting labeled data. The error type labels we collected require annotators to be experts in RAG systems so that they comprehend the role and operation of each stage, and can disambiguate the error types. Hence, crowdsourcing the annotation work was not possible, and all annotation was done manually by the authors to ensure the highest quality labels. Still, LLM-as-a-Judge systems were not fully capable of reasoning over all the intermediate information created throughout the RAG pipeline, leading to rather low stage-classification accuracy. We expect this to increase as LLMs become more proficient with multi-step reasoning and assimilating information over long contexts.

## References

- Tom Aarsen. 2025. Training and finetuning reranker models with sentence transformers v4. <https://huggingface.co/blog/train-reranker>.
- Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. *Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation*. In *2nd International Conference on Foundation and Large Language Models*, pages 607–611.
- Rama Akkiraju, Anbang Xu, Deepak Bora, Tan Yu, Lu An, Vishal Seth, Aaditya Shukla, Pritam Gundecha, Hridhay Mehta, Ashwin Jha, Prithvi Raj, Abhinav Balasubramanian, Murali Maram, Guru Muthusamy, Shivakesh Reddy Annepally, Sidney Knowles, Min Du, Nick Burnett, Sean Javiya, and 19 others. 2024. Facts about building retrieval augmented generation-based chatbots. *arXiv:2407.07858*.
- John Alexander, Kristine Toliver, and Schofield McLean. 2025. Build advanced retrieval-augmented generation systems. <https://learn.microsoft.com/en-us/azure/developer/ai/advanced-retrieval-augmented-generation>. Accessed: 2025-05-01.
- Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Geoffrey Brooks, Stefan Doering, and Jan Seidel. 2024. *Retrieval augmented generation for large language models in healthcare: A systematic review. Preprints*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.
- Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 194–199.
- Adrián Bazaga, Rexhina Biloshmi, Bill Byrne, and Adrià de Gispert. 2025. Learning to reason over time: Timeline self-reflection for improved temporal reasoning in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Amit Bleiweiss. 2024. Enhancing RAG Pipelines with Re-Ranking. <https://developer.nvidia.com/blog/enhancing-rag-pipelines-with-re-ranking/>. Accessed: 2025-05-01.
- Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stéphane Clinchant. 2025. *Provence: efficient and robust context pruning for retrieval-augmented generation*. In *The Thirteenth International Conference on Learning Representations*.
- Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. *Selectively answering ambiguous questions*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543.
- Databricks. 2025. *Improving Retrieval and RAG with Embedding Model Fine-Tuning*. Accessed: 2025-05-01.

- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578.
- Hervé Déjean, Stéphane Clinchant, and Thibault Formal. 2024. [A thorough comparison of cross-encoders and llms for reranking splade](#). *Preprint*, arXiv:2403.10407.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 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*.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [PAL: Program-aided Language Models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 10764–10799.
- Nico Grant. 2024. Google’s AI Search errors cause a furor online. *The New York Times*. Accessed: 2025-05-01.
- Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. [Automated evaluation of retrieval-augmented language models with task-specific exam generation](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions](#). *ACM Trans. Inf. Syst.*, 43(2).
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2025. [Hierarchical Document Refinement for Long-context Retrieval-augmented Generation](#). arXiv:2505.10413.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.
- Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joon-suk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural Questions: A Benchmark for Question Answering Research](#). *Transactions of the Association for Computational Linguistics*, 7.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Kevin Lin, Kyle Lo, Joseph Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5521–5533.
- Jintao Liu, Ruixue Ding, Linhao Zhang, Pengjun Xie, and Fie Huang. 2024. [CoFE-RAG: A Comprehensive Full-chain Evaluation Framework for Retrieval-Augmented Generation with Enhanced Data Diversity](#). arXiv:2410.12248.
- Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. [Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools](#). *Journal of Empirical Legal Studies*, 22(2):216–242.
- Katie Matton, Robert Ness, John Gutttag, and Emre Kıcıman. 2025. [Walk the talk? measuring the faithfulness of large language model explanations](#). In *The Thirteenth International Conference on Learning Representations*.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On Faithfulness and Factuality in Abstractive Summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

- Gabriel de Souza P Moreira, Ronay Ak, Benedikt Schifferer, Mengyao Xu, Radek Osmulski, and Even Oldridge. 2024. Enhancing Q&A Text Retrieval with Ranking Models: Benchmarking, fine-tuning and deploying Rerankers for RAG. *arXiv:2409.07691*.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Mykhailo Poliakov and Nadiya Shvai. 2025. Multi-Meta-RAG: Improving RAG for Multi-hop Queries Using Database Filtering with LLM-Extracted Metadata. In *Information and Communication Technologies in Education, Research, and Industrial Applications*. Springer Nature Switzerland.
- Renyi Qu, Ruixuan Tu, and Forrest Sheng Bao. 2025. Is semantic chunking worth the computational cost? In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Vipula Rawte, Swagata Chakraborty, Agnih Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. [The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference*, pages 109–126. NIST Special Publication.
- Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025. CLAPnq: Cohesive Long-form Answers from Passages in Natural Questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.
- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. [RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 21999–22027.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. [ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.
- Krystian Safjan. 2023. [From Fixed-Size to NLP Chunking – A Deep Dive into Text Chunking Techniques](#). Accessed: 2025-05-01.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval Augmentation Reduces Hallucination in Conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*.
- Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2025. [Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse](#). In *The Thirteenth International Conference on Learning Representations*.
- Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. 2025. [DynamicRAG: Leveraging Outputs of Large Language Model as Feedback for Dynamic Reranking in Retrieval-Augmented Generation](#). *arXiv:2505.07233*.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. [Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.
- Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2024. [Search Engines in an AI Era: The False Promise of Factual and Verifiable Source-Cited Responses](#). *arXiv:2410.22349*.
- Prashant Verma. 2025. [S2 Chunking: A Hybrid Framework for Document Segmentation Through Integrated Spatial and Semantic Analysis](#). *arXiv:2501.05485*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Zilong Wang, Zifeng Wang, Long Le, Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2025. [Speculative RAG: Enhancing retrieval augmented generation through drafting](#). In *The Thirteenth International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In

- Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating LLM Hallucinations via Conformal Abstention. *arXiv:2405.01563*.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024. [CRAG - Comprehensive RAG Benchmark](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 10470–10490.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial Report Chunking for Effective Retrieval Augmented Generation. *arXiv:2402.05131*.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024a. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.
- Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiakuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. In *Advances in Neural Information Processing Systems*, volume 37.
- Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Meta-Chunking: Learning Efficient Text Segmentation via Logical Perception. *arXiv:2410.12788*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework. *arXiv:2408.01262*.

## A RAG Pipeline and Dataset Details

This section outlines the configurable components of our RAG pipeline and provides details on the specific configurations used to obtain the experimental results presented in the main paper. In Table 5, we list important hyperparameters used on the DragonBall and CLAPnq datasets. Note that the DragonBall dataset is publicly available under a CC BY-NC-SA 4.0 license. The CLAPnq dataset is publicly available under an Apache 2.0 license.

**Chunker.** We implemented multiple chunking strategies, including fixed-length chunking (with and without overlap) and recursive chunking at the sentence level. These methods enable flexible pre-processing of documents depending on the context granularity needed. When available, our implementation also handles pre-defined semantic chunks provided by the dataset. For instance, the CLAPnq dataset provides pre-chunked documents aligned with natural discourse boundaries, which we use directly in our experiments. For the DragonBall dataset, we follow the original paper’s best configuration and apply fixed-length chunking with a window size of 128 tokens and an overlap of 25 tokens (Zhu et al., 2024).

**Retrieval.** For dense retrieval, we used models from the Hugging Face repository (e.g., gte-large-en-v1.5) to generate embeddings of the queries and chunks. Retrieval computes vector similarity (using cosine similarity) between the query and document chunks and selects the top- $k$  most relevant candidates. The value of  $k$  is treated as a hyperparameter. We fixed  $k = 8$  and  $k = 5$  for the DragonBall and CLAPnq datasets respectively.

**Reranker.** For the reranking stage, we use specialized LLM reranker models fine-tuned for ranking tasks (e.g., rank-zephyr-7b-v1-full), also sourced from Hugging Face. Given a set of retrieved candidate chunks and the original query, this model assigns a relevance score to each chunk. We then select the top  $k'$  candidates (with  $k' < k$  from retrieval) to ensure that only the most relevant context is passed to the generator. The reranking step plays a critical role in filtering noisy or marginally relevant content returned by dense retrieval. This two-stage process is meant to enhance retrieval precision, thereby improving final answer quality. We fixed  $k' = 5$  and  $k' = 3$  for the DragonBall and CLAPnq datasets respectively. The system prompt is reproduced in Listing 1.

Table 5: RAG Implementation Hyperparameters

Parameter	DragonBall	CLAPnq
Chunker	Fixed-Length size=128, overlap=25	Semantic
Embedder	gte-large-en-v1.5	gte-large-en-v1.5
Search	Top- $k=8$	Top- $k=5$
Reranker	rank-zephyr-7b-v1-full Top- $k'=5$	rank-zephyr-7b-v1-full Top- $k'=3$
Generator	Llama-3-8B-Instruct	Llama-3-8B-Instruct

```

"""
System:
You are RankLLM, an intelligent
assistant that can rank passages
based on their relevancy to the
query.

User:
I will provide you with {{K}} passages,
each indicated by a numerical
identifier []. Rank the passages
based on their relevance to the
search query: {{query}}

[1] {{chunk1}}
[2] {{chunk2}}
...
[K] {{chunkK}}

Search Query: {{query}}.

Rank the {{K}} passages above based on
their relevance to the search query.
All the passages should be included
and listed using identifiers, in
descending order of relevance. The
output format should be [] > [], e.g.
., [2] > [1]. Only respond with the
ranking results, do not say any word
or explain.

Assistant:
"""

```

Listing 1: System Prompt Used for our RAG Reranker

**Generator.** The final stage involves generating answers using an LLM. The reranked top  $k'$  chunks are concatenated and provided as context to the model alongside the query. Note that only the retrieved chunks (not full documents) are used as context. Our implementation is modular enough so that one can experiment with several LLMs from Hugging Face and select the one that best suits the needs. In our experiments, we used Meta-Llama-3-8B-Instruct. The system prompt is reproduced in Listing 2.

```

"""
System:
You are an assistant for answering
queries. You are given a list of
context (extracted parts of some
documents) and a query. Based on the

```

```
given context, provide an answer to the query. Please be concise and to the point. If you don't know the answer say 'I don't know!' Don't make up an answer. Cite the document id used. The output format should be answer with citations. Only respond with the answer, do not explain.
```

```
User:
Query:
Where is the capital of France?
Context:
{document id: 1, content: Paris is the capital of France, the largest country of Europe with 550 000 km2}
.{document id: 2, content: France is a country in Europe.}

Assistant:
Answer: Paris. [1]

User:
Query:
{{query}}
Context:
{{Context}}
"""
```

Listing 2: System Prompt Used for our RAG Generator

## B Automatic Error Classification Details

This section provides additional materials that support our main auto-evaluation analysis from section 5. We begin with an extended description of our auto-evaluation pipeline, RAGEC, including details such as the system prompts we used. Additionally, we present alternative auto-evaluation methods which empirically had lower agreement with our human annotations.

### B.1 Answer Evaluation System Prompts

The first step of RAGEC, answer evaluation uses an LLM to determine if a generated response is correct, given the query, and ground-truth answer and documents. In practice, we used GPT-4o with the prompt in Listing 3.

```
"""
You are an expert evaluator. Your task is to evaluate if a proposed answer matches the ground truth answer for a given question.

You will be provided with information between special tags :
1. The original question <question>
2. The ground truth (correct) answer <ground_truth>
```

```
3. A proposed answer to evaluate <proposed_answer>
4. (Optional) the proposed answer's cited information <ground_truth_citations>
```

Please evaluate the proposed answer based on the following criteria:

- Abstain: If the proposed answer is "I don't know" or "I don't have enough information to answer this question", then the label is 'abstain'.
- Accuracy: Does it contain the same key information as the ground truth?
- Completeness: Does it cover all important points from the ground truth?
- Correctness: Are there any factual errors compared to the ground truth?

Provide your evaluation as a JSON object with the following fields :

```
{
  "label": "\"correct\" | \"possible_correct\" | \"incorrect\" | \"abstain\"",
  "reasoning": string // A very brief explanation of your evaluation
}
```

Here are some examples:

```
<...>

<question>
{question}
</question>

<ground_truth>
{ground_truth}
</ground_truth>

<proposed_answer>
{proposed_answer}
</proposed_answer>

{citations}

Your evaluation:
"""
```

Listing 3: System Prompt Used for Answer Evaluation

### B.2 Error Stage Classification

This section describes the algorithm for the identification of which stage in the RAG pipeline caused the error.

#### B.2.1 Determining the ground truth chunks

The first step is to determine which chunks contain the ground truth information from the corpus. For

CLAPnq, the dataset provides short ground truth chunks along with each query; thus we use the dataset’s chunks directly.

For DragonBall, only the document containing the ground truth is given with the query, but these documents are long, so we perform the chunking ourselves. We leverage an LLM to determine which chunks are necessary to answer the query. We provide GPT-4o-mini the query, ground truth answer, and all the chunks from the ground truth documents, and ask for the IDs of chunks that contain information necessary to answer the query. We perform this process 10 times and select chunks appearing more than 8 times as ground truth chunks. The repetition accounts for the stochasticity of using an LLM, and better reflects its confidence. The prompt is shown below in Listing 4.

```
"""
You are an expert evaluator for
Retrieval Augmented Generation (RAG)
systems. You will help identify
chunking-related errors in RAG
systems by analyzing the
relationship between queries, ground
truth answers, and document chunks.
Your task is to identify which
chunks from the ground truth
document are relevant to answering
the query.

## Instructions
1. Review ALL available chunks from the
ground truth document
2. Identify ALL chunks containing
information relevant to answering
the query
3. Consider both direct and indirect
relevance to the query
4. Select chunks that together provide
sufficient information to answer the
query

## Context
**Query**: {query}
**Ground Truth Answer**: {ground_truth}

**Available Chunks from Ground Truth
Document**:
{chunks}

## Output Format
Provide your answer in the following
format:
Relevant Chunks: [45_1, 45_4, 45_10]
"""
```

Listing 4: System prompt for determining the chunks containing the ground truth on the DragonBall dataset

## B.2.2 Determining errors in the generation stage

Given the ground truth chunks, if we know that the chunks being passed to the generator are sufficient to determine the answer, we can conclude the error happened in the generation stage. Thus we look at the fraction of the ground truth chunks being passed to the generator. However, we must account for the possibility that multiple chunks contain the ground truth. In these cases, using only a subset of the ground truth chunks may be sufficient to determine the correct answer. To this end, we specify that if more than half of the ground truth chunks are passed to the generator, it is deemed a generator error.

If there are no ground truth chunks extracted, it is likely that the error is an abstention error as the question may be unanswerable. Thus it is also deemed a generator error in this case.

## B.2.3 Determining errors in the reranking stage

An error occurs in the reranking stage if the reranker filters out some ground truth chunks which the retriever had returned. Thus if there is any ground truth chunk that is filtered out at this stage, we deem the error a reranker error.

Note that it could be the case that the reranker filters out redundant chunks containing the ground truth, as described in the above step. However, in practice, if both chunks containing the correct answer to the question are retrieved, it is not likely that the reranker would treat them differently. We found the described method to be the most effective in determining whether there is a reranker error.

## B.2.4 Distinguishing between chunking errors and retrieval errors

Chunking errors and retrieval errors are difficult to distinguish, as the efficacy of the retriever is tightly coupled to the quality of chunks. Chunking error types E1, E2, and E3 all relate to cases where important information is missing or overshadowed in the chunks. Thus we aim to check if each “concept” in each query is preserved in at least one of the chunks containing the ground truth.

To extract the concepts in each query, we follow closely the method from Matton et al. (2025). We ask GPT-4o-mini to extract the concepts from each query, then for each concept, we ask the LLM which ground truth chunks contain that concept. We then compute the fraction of all concepts which

are included in at least one of the ground truth chunks. If this fraction is less than 0.8, we deem the error to be a chunking error. Otherwise it is deemed a retrieval error.

The parameter 0.8 was chosen to reflect that if the number of concepts extracted from the query is high, it is likely that not all the concepts are needed to answer the question. Thus we allow some leeway for the cases where the number of concepts is high (namely greater than 6).

The prompt for determining the concepts in the query is given in Listing 5, and for determining whether a concept is contained in the ground truth chunks is given in Listing 6.

```

"""
Consider the following questions. Your
task is to list the set of distinct
concepts, or high-level pieces of
information, in the 'Context' that
could possibly influence someone's
answer to the question. Each concept
should appear word-to-word in the
question, or a very minor rewording.
Here are three examples.

Example 1:
Question:
Compare the debt restructuring efforts
of Company A in 2018 and Company B
in 2021. Which company reduced more
liabilities through debt
restructuring?

Concept List:
Debt restructuring efforts
Company A
2018
Company B
2021
Reducing liabilities through debt
restructuring

{examples 2 and examples 3}

Please fill out the 'Concept List' for
the fourth example by providing a
numbered list. You should not
restate the 'Concept List' header.
You should not put dash ('-') or
numbers before each item in the list
.

Example 4
{query}
"""

```

Listing 5: System prompt used for determining the concepts in each query

```

"""
You will be given a list of excerpts and
a concept. Your job is to determine
whether the concept given is
contained in each excerpt. Output a
line for excerpt, output "True" or "
False".
"""

```

```

Example 1:
**Concepts**: Peter
**Excerpts**:
[45_3] John is running
[45_6] Peter is walking
**Answer**:
[45_3] False
[45_6] True

Example 2:
**Concepts**: running
**Excerpts**:
[45_3] John is running
[45_6] Peter is walking
**Answer**:
[45_3] True
[45_6] False

Question:
**Concepts**: {concept}
**Excerpts**:
{excerpts}
**Answer**:
"""

```

Listing 6: System prompt used for determining whether a concept is contained in the ground truth chunks

### B.3 Error Type Classification

After determining the RAG stage responsible for the error, we then prompt GPT-4o-mini  $K = 10$  times to judge the error type, and use the distribution of results to evaluate the LLM's confidence. Each stage uses a different system prompt and different contextual information relevant to that stage.

#### B.3.1 Chunking error type

If stage classification determined the error to be a chunking error, then classifying which type of chunking error only requires the query, ground truth answer, and chunks from the ground truth document. The prompt for chunking errors is given in Listing 7.

```

"""
You are an expert Retrieval Augmented
Generation (RAG) system evaluator.
Background:
Given a query, and a list of documents
possibly containing the answer to
the query, a RAG model has tried to
answer the query using the following
4 steps.
- Chunking: Each document in the
document list is split into smaller
chunks.
- Retrieval: A specified number of
chunks closest to the query will be
retrieved.
- Reranking: The retrieved chunks are
ranked for the second time,
according to how relevant the chunks
are to the query. Then, only a

```

```

specified number of reranked chunks
are retained.
- Generation: The query and reranked
chunks are passed to the generator
LLM to generate an answer.

You will be given a query and the ground
truth answer, as well as all the
chunks that belong to the document
containing the ground truth answer.
As described in the background, each
chunk will be seen as independent
text blocks by the RAG model. Given
there is an error in the chunking
step, your job is to determine the
best description of the error from
the list below.
- Overchunking: Document is split into
excessively small chunks, causing
important context to be lost.
Individual chunks appear incomplete
or ambiguous.
- Underchunking: Chunks are too large,
covering multiple topics with mixed
content. Individual chunks can be
confusing and crucial information is
diluted.
- Context Mismatch: Chunks are split at
arbitrary boundaries, disrupting the
logical structure of the document.
Key contextual links are separated
from the information they link to

Context:
**Query**: {query}
**Ground Truth**: {ground_truth}
**Document Chunks**:
{ground_truth_chunks}

**Output format**:
Your answer should only contain one of
the following,
Overchunking, Underchunking, Context
Mismatch
"""

```

Listing 7: System prompt used for determining chunking error types

### B.3.2 Retrieval error type

For retrieval errors, the LLM only needs the query, the ground truth answer, the ID for the ground truth documents, and the retrieved chunks. The prompt for retrieval errors is given in Listing 8.

```

"""
You are an expert Retrieval Augmented
Generation (RAG) system evaluator.
Background:
Given a query, and a list of documents
possibly containing the answer to
the query, a RAG model has tried to
answer the query using the following
4 steps.
- Chunking: Each document in the
document list is split into smaller
chunks.

```

```

- Retrieval: A specified number of
chunks closest to the query will be
retrieved.
- Reranking: The retrieved chunks are
ranked for the second time,
according to how relevant the chunks
are to the query. Then, only a
specified number of reranked chunks
are retained.
- Generation: The query and reranked
chunks are passed to the generator
LLM to generate an answer.

You will be given a query and the ground
truth answer. You will also be
given IDs of the documents
containing the ground truth and a
selection of document chunks
retrieved.
Given there is an error in the retrieval
step, your job is to determine the
best description of the error from
the list below.
- Missed Retrieval: Retrieved chunks do
not contain the relevant information
to answer the query from the ground
truth documents
- Low Relevance: Retrieved chunks are
only loosely related to the query
- Semantic Drift: Retrieved chunks
appear to match keywords but do not
align with the query's intent

Context:
**Query**: {query}
**Ground Truth**: {ground_truth}
**Ground Truth Document ID**: {
ground_truth_doc_ids}
**Retrieved Chunks**:
{retrieved_chunks}

**Output format**:
Your answer should only contain one of
the following:
Missed Retrieval, Low Relevance,
Semantic Drift
"""

```

Listing 8: System prompt used for determining retrieval error types

### B.3.3 Reranking error type

To determine the reranking error type, the LLM only needs the query, the ground truth answer, the retrieved chunks, and the reranked chunks. The prompt for reranking error is given in Listing 9.

```

"""
You are an expert Retrieval Augmented
Generation (RAG) system evaluator.
Background:
Given a query, and a list of documents
possibly containing the answer to
the query, a RAG model has tried to
answer the query using the following
4 steps.
- Chunking: Each document in the
document list is split into smaller

```

- chunks.
- Retrieval: A specified number of chunks closest to the query will be retrieved.
- Reranking: The retrieved chunks are ranked for the second time, according to how relevant the chunks are to the query. Then, only a specified number of reranked chunks are retained.
- Generation: The query and reranked chunks are passed to the generator LLM to generate an answer.

You will be given a query and the ground truth answer. You will also be given a selection of document chunks retrieved. The retrieved chunks will be reranked so that only { num\_reranked\_chunks} chunks are further selected. Given there is an error in the reranking step, your job is to determine the best description of the error from the list below.

- Low Recall: Necessary chunks are retrieved but reranked too low and not forwarded to the generator
- Low Precision: Irrelevant chunks are reranked highly and forwarded to the generator, with greater importance than the truly relevant chunks

Context:

```

**Query**: {query}
**Ground Truth**: {ground_truth}
**Ground Truth Document ID**: {
    ground_truth_doc_ids}
**Reranked Chunks**:
{reranked_chunks}

```

**\*\*Output format\*\*:**

```

Your answer should only contain one of
the following:
Low Recall, Low Precision
"""

```

Listing 9: System prompt used for determining reranking error types

### B.3.4 Generation error type

To determine the generation error type, the LLM only needs the query, the ground truth answer, incorrect answer by RAG model and the reranked chunks. The prompt for generation error types is as follows.

```

"""
You are an expert Retrieval Augmented
Generation (RAG) system evaluator.
Background:
Given a query, and a list of documents
possibly containing the answer to
the query, a RAG model has tried to
answer the query using the following
4 steps.

```

- Chunking: Each document in the document list is split into smaller chunks.
- Retrieval: A specified number of chunks closest to the query will be retrieved.
- Reranking: The retrieved chunks are ranked for the second time, according to how relevant the chunks are to the query. Then, only a specified number of reranked chunks are retained.
- Generation: The query and reranked chunks are passed to the generator LLM to generate an answer.

You will be given a query, the ground truth answer and an incorrect answer to the query generated by the RAG model. You will also be given { num\_reranked\_chunks} document chunks. Your job is to determine the reason why the model outputs the incorrect answer, from the list below.

- Abstention Failure: The model should have abstained but provided an incorrect answer
- Fabricated Content: The response includes information not present in the source document chunks and is unverifiable
- Parametric Overreliance: The response depends on the LLM's internal knowledge rather than the source document chunks
- Incomplete Answer: The response is correct but missing critical details
- Misinterpretation: The generator misuses or misrepresents the source document chunks
- Contextual Misalignment: The response is correct but does not address the query
- Chronological Inconsistency: The response presents events or facts in the wrong temporal order, or confuses past, present, or future timelines
- Numerical Error: The response includes incorrect calculations, quantities, or misrepresents numerical data from the retrieved documents

Context:

```

**Query**: {query}
**Ground Truth**: {ground_truth}
**Incorrect Answer**: {
    incorrect_rag_answer}
**Reranked Chunks**:
{reranked_chunks}

```

**\*\*Output format\*\*:**

```

Your answer should only contain one of
the following:
Abstention Failure, Fabricated Content,
Parametric Overreliance, Incomplete
Answer, Misinterpretation,
Contextual Misalignment,
Chronological Inconsistency,

```

```
Numerical Error
"""
```

Listing 10: System prompt used for determining generation error types

## B.4 Alternative Auto-evaluation Approaches

So far this appendix has described in detail RAGEC, our proposed error classification method. While designing RAGEC, we explored many other systems for LLM-based error classification, but RAGEC empirically had the best agreement rates with our human-annotated data. For comparison, in this section we describe two alternative approaches to error classification that we tested. For each alternative, we used GPT-4o-mini, the same model as RAGEC above for fair comparison.

### B.4.1 Single-step error type classification

The most straightforward approach to error type classification is to directly prompt an LLM to output one of the 16 defined error types given all available contextual information, including the query, ground truth response from the dataset, generated response, retrieved chunks, and reranked chunks. Consistently with other methods, the 16 error types are defined and described in the system prompt. We apply this method directly after the *answer evaluation* step of RAGEC, such that the same set of 832 potential errors from DragonBall are annotated.

Compared to RAGEC which achieved 57.8% stage classification agreement, and 40.3% error-type classification agreement, single-step prompting had only 41.1% and 31.1% agreement rates, respectively. Note that stage classification is done implicitly, taking the stage of the modal error type. Single-step prompting overloads the LLM with information which must be digested and synthesized in a single shot. Breaking the classification up by stage helps to guide the LLM’s reasoning towards a smaller set of possibilities, potentially with a more focused collection of contextual information.

### B.4.2 Stage-sequential error type classification

As above, we begin with the *answer evaluation* stage of RAGEC to identify which RAG responses need to be classified. After this, we proceed sequentially over the stages of the RAG pipeline in order, starting with chunking. At each stage, an LLM is prompted with context to determine if an error occurred at this stage, and if so, which of the error types occurred within that stage. Once the earliest error is identified, the evaluation stops.

In more detail, for the chunking stage, the context contains the query, ground truth response from the dataset, and all chunks from the ground truth document. Notably, we do not include information like the generated response or retrieved chunks, because these are not available during the chunking step and hence are not relevant to determining chunking errors. Next, for the retrieval stage we provide the query and ground truth response, but now include the retrieved chunks and the ground truth document ID. Continuing to reranking if necessary, the retrieved chunks are replaced by the reranked chunks. Because of the sequential nature of the pipeline, retrieved chunks are not strictly necessary here. Finally, if no error is identified in the first three stages, we assume the generator caused the error by process of elimination. The LLM is prompted to classify the error type given the query, ground truth response, generated response, and reranked chunks.

Stage-sequential error type classification performed better than single-step with 47.5% stage classification agreement, and 36.3% error-type classification, but this still falls short of RAGEC by a considerable margin. We analysed the performance of stage-sequential error type classification compared to human annotations and found that it lacked the ability to correctly classify any chunking errors. Without context on the generated response or how chunks were used downstream in the RAG pipeline, determining that chunking was the primary culprit is extremely difficult. This informed how we designed RAGEC, which also iterates over stages sequentially, but in a backward fashion, starting from generation and proceeding back to chunking. We also provided RAGEC more information for the chunking stage by generating ground truth chunk labels with concept extraction.

## C Additional Results

In this section we present extended results on the co-occurrence patterns of error types, which offer further insight into the structure and dependencies among different failure modes observed in RAG systems beyond what was discussed in Section 5.

### C.1 CLAPnq Error Type Distribution

Dataset characteristics can influence the failure modes of RAG systems. In Table 6, we present the distribution of the most and second most frequent error types as identified by RAGEC on the CLAPnq

Table 6: Distribution of most and second most common error types across the CLAPnq dataset.

Error Classification	Mode	Second Mode
E1 Overchunking	0	0
E2 Underchunking	0	0
E3 Context Mismatch	0	0
<b>Percentage Chunking</b>	<b>0%</b>	<b>0%</b>
E4 Missed Retrieval	180	69
E5 Low Relevance	84	99
E6 Semantic Drift	1	16
<b>Percentage Retrieval</b>	<b>21.69%</b>	<b>26.10%</b>
E7 Low Recall	23	1
E8 Low Precision	1	3
<b>Percentage Reranking</b>	<b>1.96%</b>	<b>0.57%</b>
E9 Abstention Failure	9	23
E10 Fabricated Content	87	125
E11 Parametric Overreliance	46	98
E12 Incomplete Answer	51	38
E13 Misinterpretation	607	136
E14 Contextual Misalignment	112	87
E15 Chronological Inconsistency	4	6
E16 Numerical Error	17	4
<b>Percentage Generation</b>	<b>76.35%</b>	<b>73.33%</b>
<b>Total</b>	<b>1222</b>	<b>705</b>

dataset. Compared to the DragonBall-EN dataset (see Table 2 in Section 5), the CLAPnq error distribution is much more heavily weighted towards generation errors, with few reranking errors. The lack of chunking errors is because the dataset comes pre-chunked with ground truth properly included.

## C.2 DragonBall-CN Error Type Distribution

To demonstrate that our method extends to languages other than English, we tested our method on the Chinese subset of the DragonBall dataset. Note that the documents and the queries are in Simplified Chinese. As in Table 5, the parameters used for chunking, embedding, retrieval, reranking, and generation are all the same as DragonBall-EN, with the exception that chunking uses 512 characters instead of 128 tokens, which is more suitable for the structure of Chinese. We also express the generator prompt in Chinese, as listed below:

在此任务中，你需要总结出回答问题所必需的Key Points，并使用这些Key Points来帮助你回答问题。

请按以下格式列出Key Points:

1. ...
2. ...

依此类推，根据需要增加序号，但不超过10个。

每个Key Points必须附带引用来源。随后，利用这些Key Points，在“Answer:”之后生成最终答案。

最终答案中也要引用来源。

Table 7: Distribution of most and second most common error types across the DragonBall-CN dataset.

Error Classification	Mode	Second Mode
E1 Overchunking	1	27
E2 Underchunking	8	75
E3 Context Mismatch	185	9
<b>Percentage Chunking</b>	<b>15.11%</b>	<b>18.32%</b>
E4 Missed Retrieval	296	19
E5 Low Relevance	25	89
E6 Semantic Drift	0	6
<b>Percentage Retrieval</b>	<b>25.00%</b>	<b>18.81%</b>
E7 Low Recall	378	19
E8 Low Precision	21	155
<b>Percentage Reranking</b>	<b>31.07%</b>	<b>28.71%</b>
E9 Abstention Failure	16	15
E10 Fabricated Content	93	57
E11 Parametric Overreliance	0	7
E12 Incomplete Answer	3	5
E13 Misinterpretation	213	77
E14 Contextual Misalignment	30	42
E15 Chronological Inconsistency	4	1
E16 Numerical Error	11	3
<b>Percentage Generation</b>	<b>28.82%</b>	<b>34.16%</b>
<b>Total</b>	<b>1284</b>	<b>705</b>

确保每个Key Point覆盖不同的内容。

所有答案必须使用中文进行回答。

示例问题:

新修订的《公司法》有哪些重大变化?

示例回答:

Key Points:

1. 修订强化了对公司治理的监管，明确了董事会和监事会的具体职责。[12-1]
2. 引入了强制性的ESG报告披露要求。[14-2]
3. 调整了公司资本制度，降低了最低注册资本要求。[12-1]
4. 为中小企业引入了专项扶持措施。[15-1][13-2]

Answer:

2023年修订的《公司法》引入了几项重大变化。首先，此次修订强化了对公司治理的监管，具体明确了董事会和监事会的职责。[12-1]其次，引入了强制性的环境、社会及公司治理（ESG）报告披露要求。[14-3]此外，修订还调整了公司资本制度，降低了最低注册资本要求。[12-2]最后，此次修订为中小企业引入了专项扶持措施，以促进其发展。[14-2][12-1]

## C.3 Error-types Co-occurrence Analysis

To understand whether certain error types tend to appear together, we analyze the relationship between the mode and second most frequent error types for each example for each of the RAG stages.

Figure 3 shows the co-occurrence counts between generation error types for DragonBall-EN and CLAPnq. We observe that *Misinterpretation*

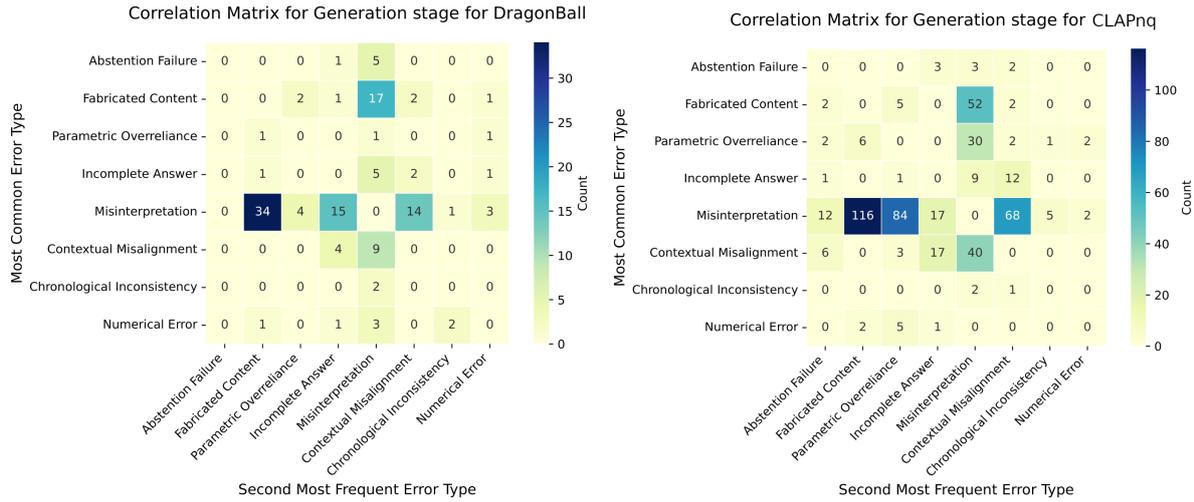


Figure 3: Joint distribution of first and second mode error categories for the generation stage. **Left:** Dragonball-EN **Right:** CLAPnq

(E13) often co-occurs with *Fabricated Content* (E10) and *Contextual Misalignment* (E14). Interestingly, for CLAPnq, *Misinterpretation* (E13) also frequently co-occurs with *Parameter Overreliance* (E11), whereas this pattern is not observed in DragonBall-EN. This can be explained by the fact that CLAPnq queries are natural questions that may exist in the model’s parametric knowledge acquired during pre-training, while DragonBall contains domain-specific queries that are unlikely to appear in the pre-training data.

Despite both datasets exhibiting significant  $\chi^2$  test results, we observe a notable difference in the density of their error co-occurrence patterns. The contingency table for the DragonBall-EN dataset is considerably sparser, indicating that error types tend to co-occur in fewer, more specific combinations. In contrast, the CLAPnq dataset presents a denser co-occurrence matrix, with a wider variety of error type pairs appearing more frequently. This difference may reflect that CLAPnq surfaces more complex and intertwined error patterns, reinforcing the need for fine-grained diagnostic evaluation across diverse RAG applications.

#### C.4 Detailed Error Classification Results

Detailed error classification results comparing RAGEC to human annotations are shown in Figure 4. For RAGEC, we show the most frequent error type out of the  $K = 10$  repeated LLM calls per query. Note that humans sometimes annotated more than one error type for a single query (though only from a single stage), and did not indicate an ordering between types. In such cases, each human

annotated error type appears in the counts in the figure. We find that some specific error types are preferred by humans or RAGEC, such as *Missed Retrieval* (E4) which was heavily used by humans, or *Context Mismatch* (E3) which was overrepresented in the RAGEC labels.

#### D AI assistant usage

In this research, we used ChatGPT to assist paper editing, literature search, and Github Copilot for coding assistance.

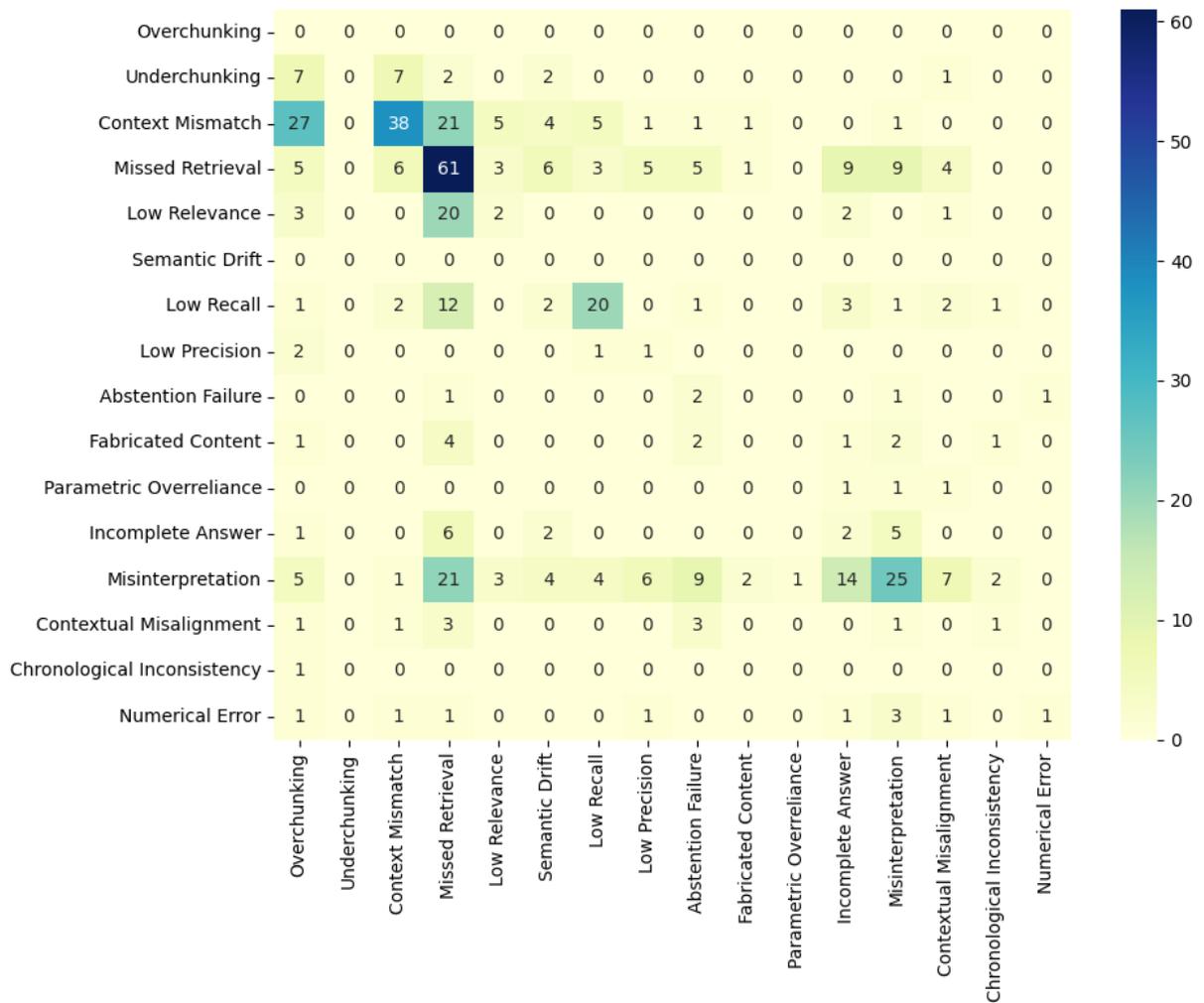


Figure 4: Joint distribution of RAGEC auto-evaluation error type classifications and human annotations for DragonBall-EN