



**Failure analysis of RAG in healthcare**  
**finding the most common failure modes of RAG systems with finetuning approaches**

**Nathaniel Apawti**

**Supervisor(s): Jie Yang, Yannick ter Heerdt**

**<sup>1</sup>EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,  
In Partial Fulfilment of the Requirements  
For the Bachelor of Computer Science and Engineering  
June 21, 2026

Name of the student: Nathaniel Apawti  
Final project course: CSE3000 Research Project  
Thesis committee: Yie Jang, Yannick ter heerdt, Pradeep Murukannaiah

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

## Abstract

This study introduces a systematic, metric-driven failure taxonomy to identify and quantify errors across document chunking, retrieval, and generation stages in RAG systems. We evaluate this framework on a benchmark derived from the *Nederlandse Huisartsen Genootschap* (NHG) protocols, comparing factual and clinical query settings. Our results show a substantial reduction in error-free performance when moving from factual tasks (137 error-free queries) to clinical scenarios (75 error-free queries). We further observe a shift in dominant failure modes: generation-level fabrications are most common in factual queries (14%), whereas clinical queries are dominated by missed retrieval (31%). Co-occurrence analysis reveals a strong association between retrieval failures and downstream generation errors, suggesting cascading effects across the pipeline. These findings highlight retrieval quality as the main bottleneck in clinical settings and motivate domain-specific retriever fine-tuning for safer deployment in Dutch primary care.

## 1 Introduction

Integrating Large Language Models (LLMs) into clinical decision support systems (CDSS) has shown to be a rapidly evolving and highly relevant area of research within healthcare informatics [1]. Recent advances show that LLM-based systems can achieve near-expert performance, with models such as Med-PaLM 2 producing answers that physicians prefer over those from other physicians in evaluation settings [2].

However, most medical questions in clinical practice require answers to adhere to local guidelines and protocols, which are often country- or institution-specific, like those of the 'Nederlandse Huisartsen Genootschap' (NHG). This specificity can result in LLM-generated responses that are generally correct but not aligned with local standards. To mitigate this risk, researchers have developed Retrieval-Augmented Generation (RAG) systems, which combine information retrieval with language generation to ground model output in relevant protocols and guidelines. These RAG-based approaches have been shown to significantly improve the faithfulness and clinical relevance of model output when grounded in structured guideline sources, compared to standard LLMs without retrieval augmentation [3; 4].

Despite these improvements, RAG models can still fail to reliably follow clinical guidelines, as errors can occur both in retrieving the relevant context and in incorrectly using that context while generating the answer. For example, a query using common terms like "heart attack" may fail to retrieve the relevant context if the guideline uses the technical term "myocardial infarction", leading to an incorrect or incomplete context grounding [3].

In order to address these limitations, this study investigates the vulnerabilities of retrieval-augmented models in a localized medical context. Specifically, this paper aims to answer the following primary research question:

**Main Research Question:** *what are the most common failure modes of RAG systems on NHG-based benchmarks and derives targeted fine-tuning strategies for Dutch primary care?*

To answer this question, the study systematically evaluates the performance of RAG systems on NHG-based benchmarks and categorizes incorrect responses using a predefined taxonomy of failure modes. The resulting error distributions are analyzed to identify the most prevalent sources of failure and to determine whether these patterns differ between factual and clinical question types. Based on the identified weaknesses, targeted fine-tuning strategies are investigated to assess their potential for mitigating dominant failure modes and improving the reliability of RAG systems in Dutch primary care settings.

The results are as follows: in the clinical benchmark, the most frequent error type was Missed Retrieval with 61 errors, whereas in the factual benchmark Fabricated Content was dominant with 27 errors. Across both settings, we observe a consistent relationship between retrieval failures and downstream generation errors, suggesting that missing or incomplete retrieval often contributes to incorrect or unsupported responses.

To summarize, this paper contributes to research by providing a systematic analysis of the most common failure modes of RAG systems on NHG-based benchmarks and proposing targeted fine-tuning strategies for these systems. This offers critical insights into where and why these systems fail, and how they can be practically improved in Dutch primary care settings.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 introduces the methodology. Section 4 reports the experimental results. Section 5 concludes the paper.

## 2 Related Work

### 2.1 Retrieval-Augmented Generation in Clinical Settings

Retrieval-Augmented Generation (RAG) has emerged as a promising approach for improving factual consistency and guideline adherence in medical large language model (LLM) systems. By retrieving relevant documents from external knowledge sources and generating responses grounded in the retrieved evidence, RAG systems reduce reliance on parametric memory and improve the traceability of generated outputs [5].

Several studies have demonstrated that RAG-based medical systems outperform standalone LLMs in terms of factual accuracy and clinical relevance, particularly on guideline-based question-answering tasks [2]. In healthcare-specific applications, retrieval from structured clinical guidelines has been shown to improve both faithfulness and interpretability, making RAG a promising foundation for clinical decision support systems [6; 7; 8]. These improvements are especially valuable in healthcare, where generated responses must remain strictly aligned with evidence-based medical practice.

However, many clinical RAG studies primarily report macro-level aggregate performance metrics such as accuracy or BLEU/ROUGE correctness. While helpful for benchmarking overall utility, these high-level assessments provide limited insight into the granular characteristics of incorrect outputs, masking the critical vulnerabilities that persist when moving from general domains to highly sensitive clinical settings.

## 2.2 Failure Modes of RAG-based Systems

To bridge this gap, recent research has increasingly focused on dissecting the specific technical limitations and failure modes of the RAG pipeline. Existing literature demonstrates that errors rarely occur uniformly; instead, they manifest at distinct stages of data processing, retrieval, and generation.

Retrieval-stage failures typically include missing relevant documents due to terminology mismatches, incomplete indexing, or embedding constraints, as well as suboptimal ranking where evidence-bearing fragments are pushed outside the LLM’s context window [9; 10]. On the other side, generation-stage failures occur when the LLM misinterprets the retrieved guidelines, yields to distractor text, contradicts its source material, or hallucinates despite having access to the correct evidence [11; 12; 13; 14]. These stage-specific errors are particularly concerning in healthcare, where unsupported or structurally broken outputs directly threaten clinical decision-making and patient safety [15].

Because traditional NLP metrics fail to capture these distinct errors [16], modern evaluation frameworks have shifted toward multi-dimensional grading, which incorporates factors like reasoning consistency, guideline adherence, and grounding quality [17; 18]. Furthermore, understanding that retrieval and generation errors are frequently interdependent and cascading [10] has motivated the development of structured error taxonomies. These taxonomies categorize outputs into discrete failure buckets (e.g., retrieval miss, partial grounding, chunking errors) [11; 13].

While taxonomy-based evaluation significantly improves error attribution and system interpretability compared to aggregate scores, existing approaches rely heavily on manual annotation, introducing evaluator subjectivity and limiting reproducibility. This limitation underscores the need for a more objective, metric-driven evaluation approach that pairs failure categories with automated quantitative metrics and strict thresholds to reliably analyze RAG robustness in specialized clinical applications.

## 2.3 Finetuning Strategies

Recent research has proposed a variety of adaptation strategies to improve the robustness of Retrieval-Augmented Generation systems. These adaptations target different stages of the RAG pipeline, including document chunking, retrieval, and generation.

At the document processing stage, advanced chunking methods have been developed to preserve semantic coherence and improve retrievability. While traditional fixed-size chunking may fragment important contextual information, newer approaches such as semantic chunking, STS-based

segmentation, and LLM-guided methods like LumberChunker and Meta-Chunking aim to identify more meaningful document boundaries, resulting in improved retrieval performance [19; 20].

At the retrieval stage, embedding models can be fine-tuned on domain-specific data to improve the alignment between user queries and relevant documents. Contrastive learning approaches and hard-negative mining have been shown to improve retrieval accuracy and reduce the retrieval of irrelevant context [21; 22].

At the generation stage, retrieval-aware fine-tuning methods such as RAFT, Finetune-RAG, and Self-RAG aim to improve grounding in retrieved evidence and reduce hallucinations [23; 24; 25]. These approaches train models to better utilize retrieved context, ignore distractor documents, and assess the reliability of generated outputs.

Although these adaptations improve retrieval quality and factual consistency, failures continue to occur in practical deployments. Consequently, understanding and systematically evaluating the failure modes that remain after these optimizations remains an important research challenge.

## 3 Methodology

This chapter presents the methodology used to evaluate failure modes in Retrieval-Augmented Generation (RAG) systems. The approach combines a structured failure taxonomy, quantitative metrics, and an automated evaluation pipeline to systematically identify and analyze errors across the RAG process.

### 3.1 Failure Taxonomy Overview

The taxonomy is derived from prior work on RAG evaluation and error classification [13; 12], and is adapted to the specific medical RAG system that was used. The failure modes are grouped according to the pipeline components: chunking, retrieval, and generation. Each failure category is associated with a quantitative metric, enabling systematic and reproducible error attribution.

A key characteristic of the proposed framework is that it supports multi-label failure attribution. A single response may exhibit multiple failure modes simultaneously. For example, a generated answer may be both incomplete and based on missing retrieved evidence. Therefore, failure categories are evaluated independently and are not mutually exclusive.

Table 1 provides an overview of the proposed taxonomy.

The following section describes the evaluation workflow used to compute these failure labels. Subsequently, Section 3.3 presents the metric definitions and thresholds associated with each category.

### 3.2 Overview of the Methodology

The proposed evaluation methodology consists of the following stages; an overview is shown in Figure 1:

1. **Data Layer and System Execution:** The RAG system is executed on a set of input queries. For each query, the retrieved chunks, the generated answer, the ground-truth answer, and the benchmarks results are stored and used. This forms the basis for all subsequent analysis.

Table 1: Overview of the proposed RAG failure taxonomy and associated metrics.

Component	Failure Mode	Metric
Chunking	Overchunking	Sequential chunk ratio
Chunking	Underchunking	IoU
Chunking	Context Mismatch	Relation Recall
Retrieval	Missed Retrieval	Context Recall
Retrieval	Low Relevance	Context Precision
Retrieval	Low Ranked Retrieval	Context Precision@K
Generation	Abstention Failure	Context Recall + Generation
Generation	Fabricated Content	Faithfulness
Generation	Incomplete Answer	Answer Recall
Generation	Misinterpretation	Misinterpretation Rate

- 2. Failure Attribution via Taxonomy Metrics:** Each instance is evaluated using the proposed taxonomy (E1–E10). Metrics are computed independently per category using llm-as-judge, allowing multiple failure modes to be assigned to a single example. This includes chunking-level, retrieval-level, and generation-level diagnostics.
- 3. Aggregation and Analysis:** Finally, the distribution of failure modes are aggregated across the dataset. This allows identification of dominant error sources and systematic weaknesses in the RAG pipeline. These results will then be used to provide finetuning improvements for the current RAG-model.

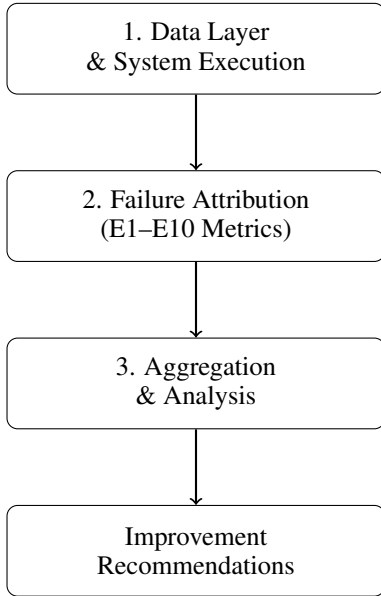


Figure 1: Overview of the proposed RAG failure analysis methodology.

### Experimental Setup

The evaluation dataset consists of clinical and factual question-answer pairs derived from the *Nederlandse Huisart-*

*sen Genootschap* (NHG) guidelines. Clinical questions are formulated as patient cases that require the interpretation of symptoms and contextual information (e.g., a patient presenting with shortness of breath and a relevant medical history), whereas factual questions focus on the retrieval of explicit guideline knowledge (e.g., diagnostic criteria or treatment recommendations). For all experiments, the top- $K$  retrieval setting was fixed at  $K = 5$  and the answers of the questions were generated using GPT-55.

### Methodological Motivation

The proposed methodology was selected to provide a reproducible and systematic evaluation framework for RAG failure analysis. By defining quantitative metrics for each failure category, the framework enables consistent error attribution across large benchmark datasets without relying solely on manual inspection. The use of an LLM-as-a-judge further supports scalability by allowing failure modes to be evaluated efficiently and consistently across hundreds of examples. While automated judging is not entirely free from bias, it reduces the subjectivity and time requirements associated with manual annotation, making large-scale error analysis feasible.

### 3.3 Failure Metrics

This section defines the quantitative metrics used to identify each failure category in the proposed taxonomy. Building on existing RAG evaluation frameworks such as RAGAS [26], the proposed metrics measure failures originating from chunking, retrieval, and generation. For each category, a metric and corresponding threshold are defined to enable systematic and reproducible failure attribution.

#### Chunking Errors (E1–E3)

**E1: Overchunking** occurs when semantically coherent information is fragmented across multiple chunks, making complete reconstruction difficult within the retrieval budget  $K$ . This is quantified by checking how many of the  $K$  retrieved documents are in a row. A high sequential chunk ratio means that a lot of chunks are 'wasted' on following chunks in a row, which could imply that there is overchunking for that specific section.

$$\text{Sequential Chunk Ratio} = \frac{\text{Number of Consecutive Chunks in a Row}}{K} \quad (1)$$

**Threshold:** Overchunking is assigned if Fragmentation Score  $> 0.9$ . 0.9 was chosen, because this would mean that more than 90% are sequential chunks and there is no space for other chunks about similar topics.

**E2: Underchunking** refers to large chunks containing irrelevant information. This is measured using concept-level Intersection-over-Union (IoU) [27]:

$$IoU = \frac{|G \cap R|}{|G \cup R|} \quad (2)$$

where  $G$  denotes gold concepts from the groundtruth answer and  $R$  denotes as golden chunk content.

**Threshold:** Underchunking is assigned if IoU Score  $< 0.1$ .

IoU is used because underchunking primarily means low semantic precision: retrieved chunks contain many additional concepts beyond those necessary for answering the question. The metric therefore evaluates the overlap between relevant answer concepts and the total concepts contained in the retrieved chunks. Lower IoU scores indicate that chunks include a lot of irrelevant information, suggesting that the chunk size is too large and not focused around the query-relevant content. Based on testing it was found that relevant chunks were not retrieved for an IoU score below 0.1,

**E3: Context Mismatch** occurs when chunk boundaries break logical relationships between entities and attributes. This is measured using Relation Recall, which uses triplets (subject, predicate, object) to form relationships, and then checks if all the relationships can be retrieved by looking at the chunks separately. [28]:

$$RelationRecall = \frac{\text{preserved relations}}{\text{gold relations}} \quad (3)$$

**Threshold:** Context mismatch is assigned if Relation Recall  $< 0.8$ .

Relation Recall is used because context mismatch primarily comes from broken semantic structure rather than missing individual entities or keywords. By measuring the proportion of gold relations preserved in the relevant chunks, the metric directly evaluates whether the semantic integrity of the original context remains intact after chunking and retrieval. Lower Relation Recall scores therefore indicate that important relationships between concepts have been disrupted. 0.8 was chosen as a threshold based on testing.

#### Retrieval Errors (E4–E6)

**E4: Missed Retrieval** is measured using Context Recall, as this error category concerns whether all claims required to construct the ground-truth answer are present in the retrieved context. [26]:

$$ContextRecall = \frac{\text{gold claims in retrieved context}}{\text{total gold claims}} \quad (4)$$

**Threshold:** Missed retrieval is assigned if Context Recall  $< 0.8$ .

Context Recall is used because missed retrieval errors occur when relevant claims or evidence is absent from the retrieved context. The metric therefore directly measures how much of the required ground-truth information is successfully retrieved. Lower Context Recall scores indicate that important claims needed to answer the query were not retrieved by the retrieval system. A threshold value of 0.8 was chosen as for clinical questions an opinion was asked which is never written in the retrieved context.

**E5: Low Relevance** is quantified using Context Precision, because this metrics checks for what proportion of the retrieved chunks is actually relevant for answering the question. [26]:

$$ContextPrecision = \frac{\text{relevant retrieved chunks}}{\text{retrieved chunks}} \quad (5)$$

**Threshold:** Low relevance is assigned if  $0 < \text{Context Precision} < 0.3$ .

**E6: Low ranked** is evaluated looking at Ranked Context Precision, this also takes into consideration the ranking. This is important as this in combination with low relevance can help identify if K needs to be decreased [26].

$$ContextPrecision@K = \frac{\sum_{k=1}^K (Precision@k \times v_k)}{\text{Total relevant items in top-K}} \quad (6)$$

**Threshold:** Low Ranked is assigned if  $0 < \text{Context Precision@K} < 0.3$ . This threshold was chosen, as values above 0.4 often still led to correct retrievals.

#### Generation Errors (E7–E10)

**E7: Abstention Failure** is assigned when the system attempts to generate a substantive response despite possessing inadequate evidence within the retrieved context. Rather than executing a necessary clinical abstention (e.g., stating that required data is missing), the system produces an ungrounded output. This failure mode is evaluated using an LLM-as-a-judge prompt which directly tells whether the retrieved text contains sufficient domain-specific information to completely fulfill the user’s query.

**E8: Fabricated Content** refers to responses that contain claims unsupported by the retrieved context. This error category is quantified using the *Faithfulness* metric, which measures the proportion of response claims that are grounded in the retrieved evidence.

$$Faithfulness = \frac{\text{supported claims}}{\text{total claims in response}} \quad (7)$$

**Threshold:** Fabricated Content is assigned if Faithfulness  $< 0.8$ .

A low Faithfulness score indicates that the generated response introduces hallucinated or fabricated information that cannot be verified from the retrieved documents. A threshold of 0.8 was selected, for non-clinical additions within the generated text. In multiple instances, responses achieving high grounding scores contained minor, ungrounded remarks, such as reassuring statements like ”this will comfort the patient”, which despite being absent from the retrieved guidelines, do not constitute a core failure.

**E9: Incomplete Answer** refers to responses that only partially address the information required by the query, despite the necessary evidence being present in the retrieved context. This error category is measured using *Answer Recall*, which evaluates how much of the relevant ground-truth information is included in the generated answer.

$$\text{Answer Recall} = \frac{\text{gold claims in answer}}{\text{total gold claims in context}} \quad (8)$$

**Threshold:** Incomplete answer is assigned if Answer Recall < 0.8.

Answer Recall is used because incomplete answers typically arise when the model omits important claims, facts, or supporting details that are available in the retrieved documents. By comparing the claims present in the generated response against the full set of gold claims contained in the context, this metric directly measures the completeness of the generated answer. Lower Answer Recall scores therefore indicate that relevant information was ignored or insufficiently incorporated into the response. A threshold of 0.8 was chosen based on testing with 30 test questions.

**E10: Misinterpretation** occurs when the generated response includes claims that misunderstand the meaning of the retrieved context. While the claims may appear related to the source material, their interpretation does not accurately reflect the original intent or factual meaning. This error category is evaluated using the *Misinterpretation Rate*.

$$\text{Misinterpretation Rate} = \frac{\text{Misinterpreted Claims}}{\text{Supported Claims}} \quad (9)$$

**Threshold:** Misinterpretation is assigned if Misinterpretation Rate > 0.2.

This metric is used because misinterpretation errors comes from the model that generates statements that are grounded in retrieved content but incorrectly interpreted. Measuring the proportion of supported claims that contain semantic distortions allows us to quantify how often the model misunderstands or incorrectly reformulates retrieved evidence. Higher values indicate that the generated response frequently alters the intended meaning of the original context. A threshold of 0.2 was established to account for minor semantic variations. In several instances, claims were not fundamentally misinterpreted but were phrased differently than the source text, leading the LLM judge to falsely flag them as logical distortions.

## 4 Experimental Results & interpretation

This section presents the results of the evaluation pipeline described in Section 3.2. First the distribution of failure modes on the clinical benchmark is analyzed, followed by the factual benchmark. Finally, we compare the two benchmarks to identify differences in error patterns and uncover the dominant sources of failure within the RAG pipeline.

### 4.1 Clinical Benchmark Failure Mode Distribution

To establish a baseline understanding of pipeline vulnerabilities, we first analyzed the absolute frequency of each error category across the clinical benchmark. Figure 2 illustrates the distribution of failure modes  $E_1$  through  $E_{10}$ .

The analysis indicates that retrieval and generation errors constitute the dominant sources of failure, whereas chunking-related issues are comparatively rare. In particular, E4:

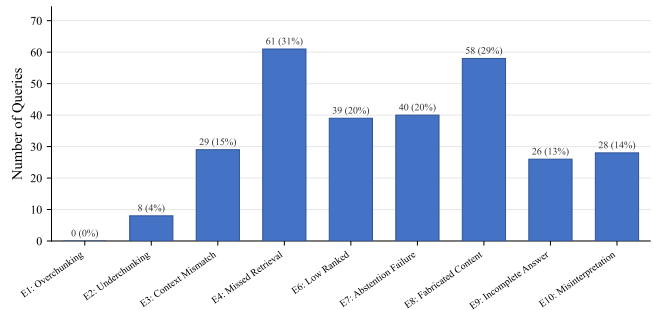


Figure 2: Total number of queries triggering each specific failure mode in the clinical benchmark.

*Missed Retrieval* emerges as the most prevalent failure mode, affecting 61 queries (31% of the evaluated set). In these cases, the retrieved context over-indexes on general diseases or symptom descriptions rather than isolating the specific operational parameters requested by the query. A representative example is detailed in Appendix A.1, where the context focuses on influenza characteristics instead of concrete safety-netting protocols.

A similar pattern is observed in E8: *Fabricated Content*, which accounts for 29% (58 queries) of the evaluated cases. Here, the pipeline fails at the granular level: when confronted with high-level source text, the generator model activates its parametric memory to fill in missing logistical or clinical parameters (e.g., specific dosages) not provided by the retriever, as illustrated in Appendix A.2. Additionally, E7: *Abstention Failure* (40 queries, 20%) and E6: *Low Ranked* (39 queries, 20%) represent substantial portions of the error distribution. In contrast, structural chunking issues are negligible, with E1: *Overchunking* observed in 0 queries and E2: *Underchunking* appearing in only 8 queries (4%).

Overall, these results suggest that the current chunking strategy is largely effective in preserving semantic coherence during document segmentation. The primary limitations are in the retrieval stage, where embedding-based matching fails to consistently surface relevant context, and in the generation stage, where the language model occasionally produces unsupported or hallucinated content despite receiving appropriate context.

### Error Co-occurrence and Cascading Failures

A central hypothesis of this study is that failures in Retrieval-Augmented Generation (RAG) systems are rarely isolated events. Instead, errors occurring in earlier stages of the pipeline, particularly retrieval, tend to propagate downstream and induce additional generation failures. To investigate this effect, we quantify the number of concurrent error types per query and analyze their co-occurrence patterns.

As illustrated in Figure 3, 75 queries execute without any detected errors. However, among failing cases, multiple concurrent failure modes are common. In particular, 26 queries exhibit two simultaneous errors, while 22 queries contain four concurrent error types. This indicates that failures tend to cluster rather than occur in isolation. To understand how these errors interact, a co-occurrence matrix was generated

(Figure 4).

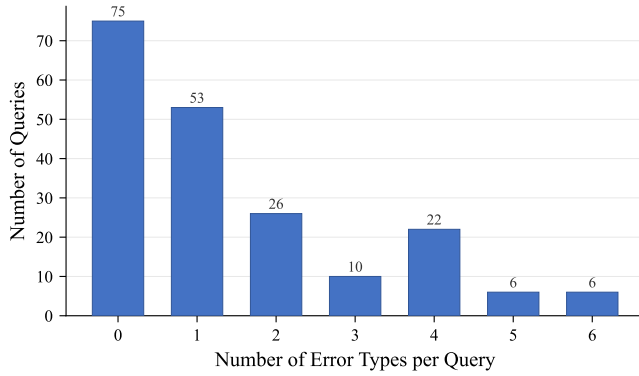


Figure 3: Distribution of the number of concurrent error types per query.

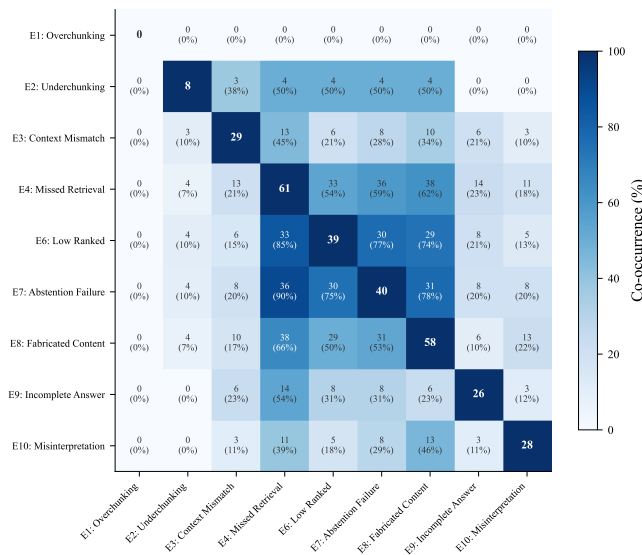


Figure 4: Heatmap of co-occurrence for clinical benchmark

The resulting patterns which can be seen in Figure 4 provide strong evidence of cascading failure behavior, especially in E4: *Missed Retrieval*. Among the 61 queries affected by E4, a large proportion also exhibit downstream generation issues, including E8: *Fabricated Content* in 38 cases. Similarly, 36 of the E4-affected queries also show E7: *Abstention Failure*, while 33 concurrently suffer from E6: *Low Ranked* retrieval outcomes. An example can be seen in Appendix A.3, here a missed retrieval downstreams to multiple other failures such as E7 (Abstention failure) and E8 (Fabricated content).

These results suggest a strong dependency between retrieval quality and generation behavior. When the retriever fails to surface relevant supporting evidence (E4), the generator is forced to operate under insufficient grounding and the model either continues generating despite the lack of ad-

equately evidence, resulting in fabricated outputs, or fails to appropriately recognize the absence of sufficient context.

## 4.2 Factual Benchmark Failure Mode Distribution

To evaluate the pipeline’s performance on standard informational queries, we analyzed the absolute frequency of each error category across the factual benchmark. Figure 5 illustrates the distribution of failure modes  $E_1$  through  $E_{10}$ .

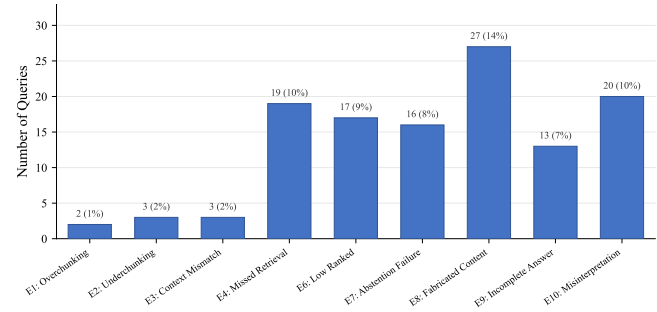


Figure 5: Total number of queries triggering each specific failure mode in the factual benchmark.

The results shown in Figure 5 indicate that failures in the factual benchmark are primarily concentrated in the generation stage, followed by retrieval-related errors. In particular, E8: *Fabricated Content* is the most frequent failure mode, occurring in 27 queries (14% of the evaluated set). Parallel to its behavior in the clinical trials, this vulnerability originates from the model expanding upon narrow context windows with ungrounded assertions pulled from its parametric weights, as shown in Appendix B.1.

This is closely followed by E10: *Misinterpretation* with 20 queries (10%). Errors within this category manifest when the generator correctly identifies the primary clinical entities but misinterprets the conditional logic, triage rules, or risk thresholds specified in the guidelines (see Appendix B.2 for an example). Followed by E4: *Missed Retrieval* with 19 queries (10%), E6: *Low Ranked* (17 queries, 9%) and E7: *Abstention Failure* (16 queries, 8%) also occur with notable frequency. In contrast, chunking-related errors are comparatively rare. E1: *Overchunking* appears in 2 queries (3%), while both E2: *Underchunking* and E3: *Context Mismatch* occur in only 3 queries each (2%)

Overall, the results indicate that chunking contributes relatively little to the overall error profile compared to retrieval and generation components. Generation-related failures, particularly fabricated content and misinterpretation, account for the largest proportion of errors in the factual benchmark.

### Error Co-occurrence and Cascading Failures

To determine whether errors occur independently or as part of cascading failure patterns, we analyze the number of concurrent error types per query, as shown in Figure 6.

The results show a high level of robustness in the factual benchmark. A total of 137 queries execute without any detected errors. Among the remaining cases, errors are most often isolated, with 29 queries exhibiting exactly one failure mode. Still, multi-error interactions are present, with 8

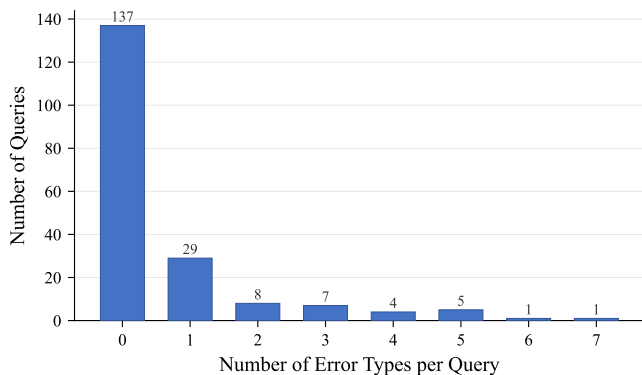


Figure 6: Distribution of the number of concurrent error types per query within the factual benchmark.

queries showing two concurrent errors, 7 queries showing three errors, and a small number of cases exhibiting up to seven simultaneous failure types.

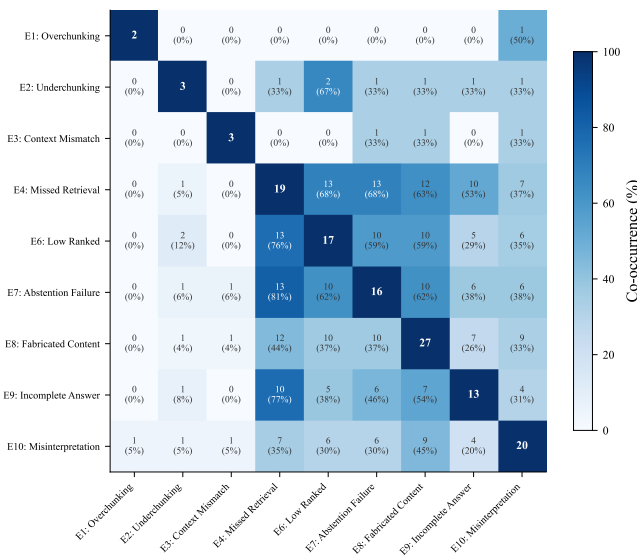


Figure 7: Heatmap detailing the co-occurrence percentages and absolute counts of failure modes in the factual benchmark.

The co-occurrence analysis shown in Figure 7 provides evidence of cascading failure behavior. Among the 19 queries affected by E4: *Missed Retrieval*, 13 also exhibit E6: *Low Ranked*, 13 co-occur with E7: *Abstention Failure*, and 12 co-occur with E8: *Fabricated Content*. An example of such a cascading failure is provided in Appendix B.3, where underchunking propagates through the pipeline and results in errors across both the retrieval and generation stages.

Overall, the factual benchmark demonstrates a high level of reliability, with 137 queries executing without any detected failure modes. Nevertheless, similar to the clinical benchmark results, retrieval failures frequently co-occur with generation failures, indicating that missing or incomplete retrieval results remain an important source of downstream errors.

### 4.3 Cross-Benchmark Comparison: Factual vs. Clinical Queries

To determine if different type of input query impact pipeline stability, we compared the results of the factual benchmark against the clinical reasoning benchmark.

Table 2: Macro Comparison of Top Failure Modes Between Benchmarks

Failure Mode	Factual Dataset	Clinical Dataset
Perfect Queries (No Errors)	137 (70.6%)	75 (38.5%)
E4: Missed Retrieval	19 (9.8%)	61 (31.3%)
E8: Fabricated Content	27 (13.9%)	58 (29.7%)
E10: Misinterpretation	20 (10.3%)	28 (14.4%)

A high level analysis immediately reveals that the factual benchmark exhibits a much higher baseline of system reliability than the clinical benchmark. As shown in Table 2, the factual benchmark contains less errors in general, with a striking 137 queries executing flawlessly without triggering a single failure mode. In contrast, only 75 queries in the clinical benchmark achieved an error-free execution. This represents an approximate 47% drop in absolute baseline reliability when moving from standard informational requests to complex clinical reasoning scenarios.

Furthermore, the structural distribution of errors differs notably between the two benchmarks:

- Shift in dominant error profiles:** The primary bottleneck shifts from generation in the factual benchmark to retrieval in the clinical benchmark (Table 2). This suggests that increased query complexity places greater demands on the retrieval stage than on generation. One possible explanation is that clinical queries frequently describe patient symptoms, risk factors, and clinical scenarios, whereas the corresponding guideline passages are organized around diagnoses, treatment protocols, or decision criteria. This semantic mismatch can reduce the effectiveness of embedding-based similarity search, as the relationship between the query and the relevant evidence is often indirect rather than explicit. In contrast, in the factual benchmark, retrieved evidence is often more directly aligned with the query, making retrieval less error-prone. However, generation errors may arise because the prompt does not sufficiently enforce grounding in the retrieved context or explicitly encourage abstention when the answer is not supported by the provided evidence.
- Persistence of cascading failures:** Despite this shift in dominant error modes, both benchmarks exhibit consistent evidence of retrieval-driven error propagation. In both settings, retrieval failures are associated with downstream generation issues. In particular, missed retrieval frequently co-occurs with abstention failures (E7) and fabricated content (E8). This indicates that insufficient retrieval context reduces the reliability of subsequent generation, leading either to unsupported outputs or inappropriate abstention.

Overall, the comparison indicates a clear shift in system behavior across benchmarks. While factual queries exhibit

higher overall reliability, clinical queries introduce a stronger retrieval bottleneck and a higher rate of reasoning-related errors. Despite these differences in magnitude, both benchmarks exhibit a consistent pattern in which retrieval failures propagate into downstream generation issues. This suggests that retrieval quality is a shared limiting factor across domains, although its impact becomes more visible in clinical settings.

#### 4.4 Implications for Targeted Fine-Tuning

The distribution of bottlenecks provides a direct roadmap for optimizing RAG systems in dutch primary care. Given that E4: *Missed Retrieval* is the primary failure bottleneck in clinical queries, optimizing the retriever is the most critical priority. The baseline embedding model clearly struggles to bridge the semantic gap between a patient’s casual symptom descriptions and the dense, diagnosis-oriented layout of the NHG guidelines. To resolve this, domain-specific retriever fine-tuning via contrastive learning, utilizing frameworks like REMED or CERA as outlined in Section 2.3 should be deployed to realign the embedding space on Dutch primary care datasets.

Furthermore, our co-occurrence analysis exposes a severe grounding vulnerability when the retriever fails. Here, the LLM falls back on its internal parametric weights, triggering E8: *Fabricated Content* and E7: *Abstention Failures* errors. This behavior demonstrates that prompting alone is insufficient to enforce safe boundaries. Mitigating these cascading errors requires the integration of retrieval-aware generation fine-tuning strategies such as RAFT or Self-RAG (see Section 2.3). Explicitly training the generator to detect context deficits, ignore noisy distractors, and safely abstain when evidence is missing is vital for clinical deployment.

### 5 Conclusions and Future Work

This study presented a systematic, metric-driven failure analysis of Retrieval-Augmented Generation (RAG) systems applied to Dutch primary care guidelines. By moving beyond aggregate evaluation metrics, we introduced a structured error taxonomy with quantitative thresholds to identify failures across retrieval and generation components.

#### 5.1 Limitations of the Current Framework

While the proposed evaluation framework enables fine-grained and largely automated identification of failure modes without relying on large-scale manual annotation, several limitations should be acknowledged.

First, the project was conducted within a constrained development period of approximately ten weeks and involved dependencies on parallel contributions from other team members. As a result, there was limited opportunity to perform extensive manual validation or human annotation of the generated error labels. This restricts the ability to fully verify the absolute correctness of all automatically assigned failure modes.

Second, the evaluation framework relies entirely on an LLM-as-a-judge to classify and attribute failure modes. While this automated judging enables scalable analysis across

hundreds of queries, it cannot replicate the nuance, absolute accuracy, and clinical judgment of expert human error checking. Consequently, the system remains vulnerable to model bias and interpretation variance, particularly when analyzing high contextual ambiguity or overlapping error categories.

#### 5.2 Conclusions

Our results show a clear shift in failure behavior between factual and clinical benchmarks. In the factual setting, the system is largely stable, with most errors occurring in generation, particularly E8: *Fabricated Content*, alongside smaller proportions of E10: *Misinterpretation* and E4: *Missed Retrieval*. This suggests that retrieved evidence is generally well-aligned with queries, and that remaining errors are mainly due to insufficient grounding during generation.

In contrast, the clinical benchmark is dominated by retrieval failures, with E4: *Missed Retrieval* becoming the most frequent error type. This reflects a semantic mismatch between symptom-oriented queries and guideline-structured documents, which reduces retrieval effectiveness and leads to incomplete context being passed to the generator. As a result, downstream generation errors such as hallucination and incorrect abstention become more frequent.

Overall, the findings highlight a strong dependency between retrieval and generation: retrieval failures often propagate into multiple downstream error types, making retrieval quality the key bottleneck in complex clinical settings. Since retrieval-related failures emerge as the primary driver of downstream errors, targeted retriever fine-tuning appears to be the most promising mitigation strategy. In particular, domain-specific retrieval adaptation could improve alignment between clinical queries and NHG guideline content, thereby reducing missed and low-ranked retrieval cases. Additionally, retrieval-aware generation approaches may improve grounding and abstention behavior, reducing fabricated content and other generation-related failures.

#### 5.3 Future Work

The next step is to implement and evaluate the mitigation strategies identified through the failure analysis. Since retrieval failures were found to be the primary source of downstream errors, future work should first focus on improving the retrieval component of the pipeline.

Domain-specific retriever fine-tuning using contrastive learning methods such as REMED and CERA could improve alignment between clinical queries and NHG guideline passages. Future evaluation should determine whether these approaches reduce missed retrieval and low-ranked retrieval failures.

In addition, retrieval-aware generation approaches such as RAFT, Finetune-RAG, and Self-RAG should be explored to improve grounding and abstention behavior. These methods may help reduce fabricated content and other generation-related errors by encouraging stronger reliance on retrieved evidence.

Overall, jointly improving retrieval and generation is expected to increase the robustness and safety of RAG systems for Dutch primary care applications.

## 6 Responsible Research

### 6.1 Ethical Considerations

The research conducted in this study does not raise ethical concerns. All experiments are performed using publicly available benchmark datasets consisting of predefined clinical question-answering tasks. No patient records, personal information, or other sensitive data are used during data collection, model evaluation, or analysis. As a result, the study does not involve privacy risks, informed consent requirements, or the processing of personally identifiable information.

Furthermore, the objective of this work is to evaluate and analyze the failure modes of Retrieval-Augmented Generation (RAG) systems rather than to deploy a clinical decision support system in practice. The proposed methodology is therefore intended to improve transparency and understanding of system limitations, which may ultimately contribute to the development of safer and more reliable medical AI systems.

### 6.2 Reproducibility

The reproducibility of the evaluation methodology is high, because the proposed metric-based evaluation framework relies on an LLM-as-a-Judge approach. Given the same inputs, prompts, and evaluation procedure, this approach provides a highly reproducible method for identifying the failure mode. Especially since the temperature is set to zero, making sure the same output is retrieved every time given the same input. The code itself can be found on `github`: `'https://github.com/Nuffs/RP_NHG_2026'` in the `failure_analysis` branch, where the `'RP_NHG_2026.analysis.all.py'` file triggers the failure analysis.

### 6.3 Use of Generative AI

Generative AI tools were used in accordance with TU Delft guidelines. These tools were primarily used to improve the clarity, structure, and readability of written text. They were not used to generate research results, conduct experiments, perform data analysis, or draw scientific conclusions.

All AI-generated suggestions were critically reviewed, verified, and revised by the author before inclusion in this report. The author remains fully responsible for the content, methodology, analysis, and conclusions presented in this work.

## References

- [1] C. Lin and C.-F. Kuo, "Roles and potential of large language models in healthcare: A comprehensive review," *Biomedical Journal*, pp. 1–8, 2025. [Online]. Available: <https://doi.org/10.1016/j.bj.2025.100868>
- [2] K. Singhal, S. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, J. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachany, C. Semigran, C.-C. Liu, Y. Makuuchi, G. Barral, V. Natarajan, D. Scoccia, M. Karton, K. Prabhu, J. Schrouff, A. Loh, N. Howell, G. S. Corrado, E. Ortiz, S. Azizi, A. Karthikesalingam, and Y. Liu, "Towards expert-level medical question answering with large language models," *Nature Medicine*, vol. 30, no. 12, pp. 3521–3531, 2024. [Online]. Available: <https://www.nature.com/articles/s41591-024-03423-7>
- [3] A. Bora, H. Cuayáhuil et al., "Retrieval-augmented generation (rag) in healthcare: A comprehensive review," *AI*, vol. 6, no. 1, pp. 226–250, 2025. [Online]. Available: <https://www.mdpi.com/2504-4990/6/4/116>
- [4] G. Xiong, Q. Jin, Z. Wang et al., "Benchmarking retrieval-augmented generation for medicine," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 6233–6252. [Online]. Available: <https://aclanthology.org/2024.findings-acl.372/>
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. [Online]. Available: <https://arxiv.org/pdf/2005.11401>
- [6] K. Wołk, "Evaluating retrieval-augmented generation variants for clinical decision support: Hallucination mitigation and secure on-premises deployment," *Electronics (MDPI)*, vol. 14, no. 21, p. 4227, 2025. [Online]. Available: <https://www.mdpi.com/2079-9292/14/21/4227>
- [7] S. Zhang, E. Phan, P. E. Velmovitsky, Q. Pham, and S. Sanner, "Retrieval-augmented generation for medical question answering on a heart failure dataset: Performance analysis," *JMIR Formative Research*, vol. 10, no. 1, p. e84932, 2026. [Online]. Available: <https://formative.jmir.org/2026/1/e84932>
- [8] B. G. Collaco, P. Srinivasagam, C. A. Gomez-Cabello, S. A. Haider, A. Genovese, N. G. Wood, S. Bagaria, M. A. Lifson, and A. J. Forte, "Integrating fine-tuning and retrieval-augmented generation for healthcare ai systems: A scoping review," *Bioengineering (MDPI)*, vol. 13, no. 2, p. 225, 2026. [Online]. Available: <https://www.mdpi.com/2306-5354/13/2/225>
- [9] T. Miao, R. Zhou, S. He, K. Liu, and J. Zhao, "Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2410.07176>
- [10] T. Cao, N. Bhandari, A. Yerukola, A. Asai, and M. Sap, "Out of style: Rag's fragility to linguistic variation," 2025. [Online]. Available: <https://arxiv.org/abs/2504.08231>
- [11] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Mindful-rag: A study of points of failure in retrieval augmented generation," *arXiv preprint arXiv:2407.12216*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.12216>
- [12] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," *arXiv preprint arXiv:2401.05856*, 2024. [Online]. Available: <https://arxiv.org/abs/2401.05856>

- [13] K. K. Leung, M. Belbahri, Y. Sui, A. Labach, X. Zhang, S. A. Rose, and J. C. Cresswell, "Classifying and addressing the diversity of errors in retrieval-augmented generation systems," 2026. [Online]. Available: <https://arxiv.org/abs/2510.13975>
- [14] L. Wong, A. Ali, R. Xiong, S. Shen, Y. Kim, and M. Agrawal, "Retrieval-augmented systems can be dangerous medical communicators," *arXiv preprint arXiv:2502.14898*, 2025, proceedings of the 42nd International Conference on Machine Learning (ICML). [Online]. Available: <https://arxiv.org/abs/2502.14898>
- [15] M. G. S. A. A. S. F. B. P. V. Jean-Philippe Corbeil, Minseon Kim, "Medriskeval: A structured medical risk evaluation benchmark for llms," 2025. [Online]. Available: <https://arxiv.org/abs/2507.07248>
- [16] Y. Q. Aswini Sivakumar, Vijayan Sugumaran, "Rag-x: Systematic diagnosis of retrieval-augmented generation for medical question answering," 2026. [Online]. Available: <https://arxiv.org/abs/2603.03541>
- [17] M. T. I. R. R. S. M. A. R. M. Rahman, "Noise, distraction, and mitigation: An analysis of rag failure modes in medical question answering," 2026. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/11454362>
- [18] H. Z. X. W. C. L. Y. W. Yuelyu Ji, Min Gu Kwak, "Medragchecker: A claim-level verification and diagnostic framework for biomedical rag," 2026. [Online]. Available: <https://arxiv.org/abs/2601.06519>
- [19] A. V. Duarte, J. D. S. Marques, M. Graça, M. Freire, L. Li, and A. L. Oliveira, "Lumberchunker: Long-form narrative document segmentation," *arXiv preprint arXiv:2406.17526*, 2024. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.377.pdf>
- [20] J. Zhao, Z. Ji, Y. Feng, P. C. Qi, S. Niu, B. Tang, F. Xiong, and Z. Li, "Meta-chunking: Learning text segmentation and semantic completion via logical perception," *arXiv preprint arXiv:2410.12788*, 2024. [Online]. Available: <https://openreview.net/forum?id=gh563RwulS>
- [21] T. Pang, K. Tan, Y. Yao, X. Liu, F. Meng, C. Fan, and X. Zhang, "Remed: Retrieval-augmented medical document query responding with embedding fine-tuning," in *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/10651011/>
- [22] F. Vargas, J. Robiatti, D. Alves, L. P. Valem, M. Seeth, S. Ferrada, A. Agrawal, D. Pedronette, and A. Freitas, "Beyond topical similarity: Contrastive evidence retrieval with interpretable attention alignment in rag," *arXiv preprint arXiv:2606.01482*, vol. v1, 2026. [Online]. Available: <https://arxiv.org/abs/2606.01482v1>
- [23] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "Raft: Adapting language model to domain specific rag," *arXiv preprint arXiv:2403.10131*, 2024. [Online]. Available: <https://arxiv.org/html/2403.10131v1>
- [24] Z. P. Lee, A. Lin, and C. Tan, "Finetune-rag: Fine-tuning language models to resist hallucination in retrieval-augmented generation," *arXiv preprint arXiv:2505.10792*, 2025. [Online]. Available: <https://arxiv.org/abs/2505.10792>
- [25] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2023. [Online]. Available: <https://arxiv.org/abs/2310.11511>
- [26] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," 2023. [Online]. Available: <https://arxiv.org/abs/2309.15217>
- [27] B. Smith and A. Troynikov, "Evaluating chunking strategies for retrieval," Chroma, Tech. Rep., July 2024. [Online]. Available: <https://www.trychroma.com/research/evaluating-chunking#chunking-evaluation-dataset>
- [28] L. Pommeret, T. Gerald, P. Paroubek, S. Ghannay, C. Servan, and S. Rosset, "LLM-based atomic propositions help weak extractors: Evaluation of a propositioner for triplet extraction," in *Proceedings of the Workshop on Knowledge Graphs and Large Language Models (KG-LLM@LREC)*. Palma de Mallorca, Spain: European Language Resources Association (ELRA), May 2026. [Online]. Available: <https://arxiv.org/abs/2604.02866>

## A Examples of clinical failures

this provides examples of individual failure modes

### A.1 Case Study: Missed Retrieval Failure (E4) clinical benchmark

This example demonstrates a classic Context Recall (E4) failure from our evaluation dataset. The user query explicitly requests actionable, guideline-based clinical instructions regarding when a patient should re-contact the primary care practice. However, because the system's generated retrieval query focuses overly on n general descriptions of the illness rather than the specific operational intent, the retrieved chunks contain general diagnostic criteria but completely omit the core safety-netting thresholds required to safely guide the patient.

#### Retrieved Context Provided to Generator Component

The following raw segments represent the top-ranked context blocks passed to the downstream language model for generation, demonstrating a complete absence of safety-netting contact rules:

```
[Chunk 1] (doc: influenza_0042)
Respiratory viruses (such as influenza virus) can cause pneumonia and/or bacterial pneumonia can occur as a complication in the case of an influenza-like illness (see NHG Standard Acute Cough).

[Chunk 2] (doc: influenza_0024)
- The presentation of an influenza virus infection varies widely. The severity of symptoms depends on the (sub)type of virus and the patient's immunity. - An infection with the seasonal influenza virus is frequently subclinical, especially in adults with (partial) immunity. - An estimated 15-75% of infections are asymptomatic; the wide variation is due to differences in the type of examination performed. A patient with an influenza virus infection may develop symptoms consistent with an influenza-like illness: acute onset of symptoms and >= 1 of the following 4 systemic symptoms: fever or feeling of fever general malaise headache muscle pain. With influenza, the temperature often rises sharply and suddenly... especially in older patients, the temperature may be less elevated or not elevated at all.

[Chunk 3] (doc: influenza_0025)
- The clinical picture of an infection with an influenza virus can vary from an asymptomatic course to a severe course with complications. - Usually, people recover spontaneously within a few days to a week; sometimes symptoms can last longer or be more severe. - A number of patient groups have an indication for vaccination in the run-up to the flu season... Due to the lack of evidence of effectiveness, we do not recommend antivirals for the treatment of patients with suspected influenza.

[Chunk 4] (doc: influenza_0026)
Influenza is suspected in the following cases: an influenza-like illness: acute onset of symptoms AND >= 1 of the following 4 symptoms: (feeling of) fever; general malaise; headache; muscle pain AND >= 1 of the following 3 respiratory symptoms: cough; sore throat; shortness of breath... A positive PCR test confirms suspicion.

[Chunk 5] (doc: influenza_0027)
- If an infection with the influenza virus is symptomatic, it usually results in a short-term respiratory infection that resolves on its own. - In rare cases, usually in risk groups, an infection with the influenza virus can cause serious complications that may lead to hospitalization and/or death. The most important and common complications are primary viral infections such as viral pneumonia and myocarditis, and secondary bacterial infections, such as bacterial pneumonia. Additionally, an infection with the influenza virus can disrupt existing comorbidities, such as diabetes mellitus, cardiac and pulmonary conditions, depression, and dementia. The risk of complications increases with age.
```

**Failure Analysis Takeaway:** As shown in Table 3, this case represents a widespread failure mode observed across our benchmarking runs: the generated retrieval queries over-index on descriptive disease characteristics and comprehensive symptom lists (e.g., pulling segments discussing general risk factors, diagnostic variations, or pneumonia com-

plications). Crucially, they fail to isolate the specific action-oriented logic of the query, namely the exact thresholds dictating when a patient must re-engage with primary care services (safety-netting guidelines). Consequently, 75% of the core facts required to resolve the clinical scenario are missing from the context before the generation stage even begins.

### A.2 Case Study: Fabricated Content Failure (E8)

This case study illustrates a Faithfulness (E8) failure, commonly referred to as an ungrounded hallucination. While the generator's response is clinically accurate and adheres perfectly to standard medical practice, it violates strict RAG constraints by introducing specialized details absent from the retrieved text.

#### Retrieved Context Provided to Generator Component

The following source segments represent the text blocks the language model used to synthesize its answer:

```
[Chunk 1] (doc: hand-en-polsklachten_0340)
Advise pain medication if necessary in accordance with the policy for acute nociceptive pain in the NHG Standard Pain (paracetamol and/or a (dermal) NSAID).

[Chunk 2] (doc: hand-en-polsklachten_0273)
Advise pain medication if necessary in accordance with the policy for acute nociceptive pain in the NHG Standard Pain (paracetamol and/or a (dermal) NSAID).

[Chunk 3] (doc: hand-en-polsklachten_0329)
Advise pain medication if necessary in accordance with the policy for acute nociceptive pain in the NHG Standard Pain (paracetamol and/or a (dermal) NSAID).

[Chunk 4] (doc: hand-en-polsklachten_0349)
Advise pain medication if necessary in accordance with the policy for acute nociceptive pain in the NHG Standard for Pain (paracetamol and/or a (dermal) NSAID).

[Chunk 5] (doc: hand-en-polsklachten_0339)
- Advise keeping the hand elevated and sparing it pending further diagnostics, but continuing to use the hand and wrist within the pain limits to prevent unnecessary stiffness - Provide a sling if necessary.
```

**Failure Analysis & Elaboration:** As outlined in Table 4, this case reveals a distinct, knowledge-leakage failure mechanism typical of large language models operating within specialized clinical frameworks. The retrieved context correctly identifies the high-level answer, pointing to paracetamol and dermal NSAIDs as the standard policy for acute nociceptive pain. However, the source text leaves the practical implementation details entirely open.

To bridge this gap, the generator model activates its internal parametric memory to inject specific clinical parameters, including the standard adult dosing regimen for paracetamol (*1000 mg up to 4 times daily, max 4000 mg/day*) and a crucial medical safety-check (*evaluating contraindications before prescribing NSAIDs*).

From a purely medical standpoint, these additions are completely correct and represent high-quality care. However, from a RAG engineering architecture evaluation perspective, they are classified as "fabricated fabrication failure" because they cannot be cross-referenced or audited using only the text snippets that were retrieved. This case demonstrates that the generator model often leaks external training knowledge to expand on high-level source text, making automated verification difficult unless the underlying guideline databases are fully expanded with exact dosing charts.

Table 3: Detailed Breakdown of an Automated Context Recall (E4) Failure Analysis

Metric Component	Content Value
Query ID	63d4ebcb
User Query	Mr. Jansen, 67 years old, comes to your consultation with flu-like symptoms. He now reports shortness of breath and feels confused. You have previously given him advice. What do you instruct Mr. Jansen regarding contacting the practice again?
Generated Query	<i>influenza-achtig ziektebeeld contact opnemen huisarts complicaties</i>
Ground Truth (GT) Answer	Mr. Jansen should get in touch if his symptoms increase, or if he experiences shortness of breath or confusion. These symptoms could indicate complications.
Gold Claims (from GT)	<ol style="list-style-type: none"> <li>1. Mr. Jansen must contact you if symptoms worsen.</li> <li>2. Mr. Jansen must contact you if he experiences shortness of breath.</li> <li>3. Mr. Jansen must contact you if he becomes confused.</li> <li>4. Worsening symptoms, shortness of breath, or confusion may indicate complications.</li> </ol>
Supported Claims	<ol style="list-style-type: none"> <li>4. Worsening symptoms, shortness of breath, or confusion may indicate complications.</li> </ol>
Missed Claims	<ol style="list-style-type: none"> <li>1. Mr. Jansen must contact you if symptoms worsen.</li> <li>2. Mr. Jansen must contact you if he experiences shortness of breath.</li> <li>3. Mr. Jansen must contact you if he becomes confused.</li> </ol>
Context Recall Score	<b>0.2500</b>

Table 4: Detailed Breakdown of an Automated Fabricated content (E8) Failure Analysis

Metric Component	Content Value
Query ID	84f5bcda
User Query	Mrs. Bakker, 46 years old, bruised her wrist after falling off her bicycle. She is in a lot of pain and is asking for advice on pain relief. Which medication is preferred?
Generated Response	For a sprained wrist, <b>paracetamol is the preferred first choice</b> for pain relief. Paracetamol <b>1000 mg up to 4 times a day</b> as needed, maximum <b>4000 mg per day</b> . If this is insufficient, an <b>NSAID</b> may be added or used, preferably locally/dermally, such as <b>diclofenac gel</b> , provided there are no contraindications. <b>So: start with paracetamol; optionally supplement with a dermal NSAID.</b>
Response Claims	<ol style="list-style-type: none"> <li>1. For a bruised wrist, paracetamol is preferred as the first choice for pain relief.</li> <li>2. Paracetamol 1000 mg up to 4 times a day as needed.</li> <li>3. The maximum dosage of paracetamol is 4000 mg per day.</li> <li>4. If paracetamol is insufficient, an NSAID may be added or used.</li> <li>5. Preferably, an NSAID is used locally/dermally, such as diclofenac gel.</li> <li>6. The use of an NSAID is permitted provided there are no contraindications.</li> <li>7. The advice is: start with paracetamol; optionally supplement with a dermal NSAID.</li> </ol>
Supported Claims	Claims 1, 4, 5, and 7 (Grounded in the general mention of paracetamol and dermal NSAIDs).
Fabricated Claims	<ol style="list-style-type: none"> <li>2. Paracetamol 1000 mg up to 4x per day as needed.</li> <li>3. The maximum dosage is 4000 mg per day.</li> <li>6. The use of an NSAID is permitted provided there are no contraindications.</li> </ol>
Faithfulness Score	<b>0.5714</b> (4 out of 7 claims supported)

### A.3 Case Study: Multi-Error Cascading Pipeline Failure (E4, E7, E8)

This case study demonstrates a systemic pipeline failure characterized by an empty retrieval context (retrieved\_context : []). This upstream retrieval collapse triggers a cascading sequence of errors downstream, simultaneously registering as a complete Context Recall Failure (E4), low ranked retrieval (E6), a severe Abstention Failure (E7), and a completely ungrounded Fabricated Content Failure (E8).

#### Retrieved Context Provided to Generator Component

The following block represents the total context available to the system prior to generation, illustrating a critical system disconnect:

```
[Empty Context Space]
retrieved_context: []
```

**Failure Analysis & Elaboration:** This example in Table 5 illustrates how a single retrieval failure can trigger multiple downstream errors. Because no relevant context was retrieved, the system first gets a complete Context Recall Failure (E4). The absence of supporting evidence then prevents the model from grounding its response, leading to a Fabricated Content Failure (E8). Furthermore, despite the empty context, the model does not abstain and instead generates a detailed answer, resulting in an Abstention Failure (E7). This case demonstrates that retrieval failures are not isolated events but can propagate through the pipeline, causing multiple generation-level errors from a single upstream defect.

Table 5: Detailed Breakdown of a Systemic Multi-Error Cascading RAG Pipeline Failure

Metric Component	Content Value
Query ID	92a1c8ef
User Query	Mr. Jansen, 45 years old, has been suffering from depressive symptoms since experiencing problems at work. He wonders whether he should call in sick. What do you advise him regarding his work?
Ground Truth (GT) Answer	Advise him to continue working, unless work is the primary cause of the complaints, because this is often better than staying home.
Generated Response	Advise Mr. Jansen not to <b>automatically report sick</b> , but to <b>continue working as much as possible</b> , possibly with adjustments. Specifically: Work can provide structure, rhythm, and social contact... Discuss problems with his <b>supervisor</b> and/or <b>company doctor</b> ... See if temporary adjustments are possible... If working is truly not possible, a <b>partial or temporary sick leave</b> may be necessary, but preferably with a plan to maintain contact... In short: <b>remain active and connected to work if possible, but involve the company doctor.</b>
Context Recall (E4)	<b>Score: 0.0000</b> (0 out of 3 Gold Claims supported). <i>Missed Claims:</i> All ground truth assertions regarding continuing work versus staying home were lost due to zero document retrieval.
Faithfulness (E8)	<b>Score: 0.0000</b> (0 out of 12 Response Claims supported). <i>Fabricated Claims:</i> Every single one of the 12 generated claims, spanning operational workplace adjustments, structural benefits of work, and company doctor occupational health integration is classified as a total hallucination relative to the empty context.
Abstention Failure (E7)	<b>Score: 0.0000</b> (Verdict: INSUFFICIENT — Confidence: 1.00). <i>Judge Reasoning:</i> No chunks were retrieved. The system possesses zero underlying evidence to formulate an answer, yet completely failed to execute a necessary clinical abstention.

## B Factual benchmark failure examples

### B.1 Case Study: Fabricated Content Failure (E8)

This case study illustrates a nuanced Faithfulness (E8) failure mode driven by logical guessing. Although the system achieves a relatively high overall grounding score (0.7273), the automated judge correctly isolates critical operational assertions regarding a highly specialized medical procedure (*renal replacement therapy*) that are completely unsupported by the high-level primary care guidelines retrieved.

#### Retrieved Context Provided to Generator Component

The following source segments represent the text blocks the language model used to synthesize its answer:

[Chunk 1] (doc: chronische_nierschade_0090) Should the prescriber determine kidney function in elderly patients whose current kidney function is unknown and to whom they prescribe medication where impaired kidney function is a factor?
[Chunk 2] (doc: chronische_nierschade_0068) Protein restriction of 0.8 g/kg ideal body weight per day is advised for an eGFR < 30 ml/min/1.73 m2. Should such a dietary intervention not already be advised for patients with chronic kidney damage with an eGFR >= 30 ml/min/1.73 m2, particularly if they also have elevated albuminuria?
[Chunk 3] (doc: chronische_nierschade_0128) In the 2012 multidisciplinary guideline on Polypharmacy in the Elderly, impaired renal function is mentioned as a risk factor... in the staging of chronic renal damage, this cutoff value is not applied, and chronic renal damage is assumed at an eGFR < 60 ml/min/1.73 m2... Additionally, the working group considers it important that current kidney function is known for patients with polypharmacy.
[Chunk 4] (doc: chronische_nierschade_0113) Is treatment with a statin (and/or ezetimibe) indicated in patients with chronic kidney damage to reduce the risk of cardiovascular and renal endpoints and mortality?
[Chunk 5] (doc: chronische_nierschade_0047) The treatment of chronic kidney damage is aimed at reducing the increased risk of cardiovascular morbidity/mortality and preventing or limiting the progression of kidney damage. - The general practitioner can manage a large proportion of patients themselves. This applies in any case to patients with: - eGFR 30 ml/min/1.73 m in combination with normal albuminuria - eGFR 30 ml/min/1.73 m in combination with normal albuminuria - eGFR 45 ml/min/1.73 m in combination with moderately elevated albuminuria

- Determine the management based on:
- the risk category
- the cause and course of the chronic kidney damage
- factors such as age and comorbidity
- the risk category
- the cause and course of the chronic kidney damage
- factors such as age and comorbidity
- A weighing of the pros and cons of strict (pharmacological) treatment is necessary, especially in patients with a limited life expectancy and/or extensive comorbidity. Make this assessment together with the patient.
  - If desired, treat patients with a significantly increased risk (red in Table 1) yourself who have not been referred to the internist-nephrologist due to a limited life expectancy and/or extensive comorbidity. In these patients, the emphasis is on identifying and usually in consultation with the internist-nephrologist pharmacologically treating metabolic complications (see section Monitoring). - For a summary of the management strategy by risk category, see Appendix 1. - eGFR 30 ml/min/1.73 m2 in combination with normal albuminuria - eGFR 45 ml/min/1.73 m2 in combination with moderately elevated albuminuria - the risk category - the cause and course of chronic kidney damage - factors such as age and comorbidity

**Failure Analysis & Elaboration:** This example in table 6 illustrates a fabricated content failure. Although the retrieved context contains relevant information about the general management of chronic kidney disease, it does not contain explicit guidance on renal replacement therapy. The generator therefore extrapolates from broadly related evidence and supplements the response with information drawn from its parametric knowledge.

As a result, most response claims remain grounded in the retrieved context, producing a relatively high faithfulness score (72.7%). However, several clinically plausible statements, including criteria related to renal replacement therapy and eGFR cutoffs, are unsupported by the retrieved evidence and are therefore classified as fabricated content. This case demonstrates how partially relevant retrieval can lead to subtle generation errors that are difficult to detect from the overall response quality alone.

### B.2 Case Study: Misinterpretation Failure (E10)

This case study illustrates a Misinterpretation Failure (E10). In this scenario, the generator model correctly identifies the

Table 6: Detailed Breakdown of an Automated Faithfulness (E8) Partial Failure Analysis

Metric Component	Content Value
Query ID User Query Generated Response	chronische nierschade_0013_0001 Which criteria are important when deciding whether or not to propose renal replacement therapy? When deciding whether or not to propose renal replacement therapy, such as dialysis or transplantation, no single criterion is decisive. According to the guideline, the assessment should primarily be based on: The severity/risk category... The cause and course... Age, frailty, and comorbidity... Life expectancy and expected treatment benefit... The burden and disadvantages... Presence of complications of kidney failure that may necessitate treatment... Wishes, goals, and preferences of the patient... In short: renal replacement therapy should not be proposed based solely on an eGFR cutoff value, but after an individual assessment... In cases of very limited life expectancy or severe comorbidities, conservative treatment may be more appropriate.
Response Claims	<ol style="list-style-type: none"> <li>1. No single criterion is decisive when deciding whether or not to propose renal replacement therapy.</li> <li>2. The assessment must be based on the severity/risk category of the chronic kidney damage (eGFR and albuminuria).</li> <li>3. The assessment must be based on the cause and course of the kidney damage.</li> <li>4. The assessment must be based on age, frailty, and comorbidity.</li> <li>5. The assessment must be based on life expectancy and expected treatment benefit.</li> <li>6. The assessment must be based on the burden and disadvantages of treatment versus the benefits.</li> <li>7. The assessment must be based on the presence of complications of renal failure.</li> <li>8. The assessment must be based on the wishes, goals, and preferences of the patient.</li> <li>9. The decision should be made together with the patient.</li> <li>10. Renal replacement therapy should not be proposed solely on the basis of an eGFR cutoff value.</li> <li>11. In cases of very limited life expectancy or severe comorbidity, conservative treatment may be more appropriate.</li> </ol>
Supported Claims	Claims 2, 3, 4, 5, 6, 8, 9, and 11 (Semantically mapped from general management guidelines).
Fabricated Claims	<ol style="list-style-type: none"> <li>1. No single criterion is decisive when deciding whether or not to propose renal replacement therapy.</li> <li>7. The assessment must be based on the presence of complications of renal failure.</li> <li>10. Renal replacement therapy should not be proposed solely on the basis of an eGFR cutoff value.</li> </ol>
Faithfulness Score	0.7273 (8 out of 11 claims supported)

core clinical conditions but actively alters a specific clinical guideline by transforming a conditional safety boundary into an absolute restriction. This introduces a logical distortion that can affect subsequent clinical decision-making.

### Retrieved Context Provided to Generator Component

The following source segments represent the text blocks the language model used to synthesize its answer:

```
[Chunk 1] (doc: hartfalen_0156)
- Severe renal impairment (eGFR < 15ml/min/1.73m2) - Type 1 diabetes mellitus and other situations that increase the risk of euglycemic ketoacidosis, such as alcoholism, malnutrition, intermittent fasting, diet with < 70 grams of carbohydrates per day - Active foot ulcer (relative contraindication: history of foot ulcer). Temporarily discontinue the SGLT2 inhibitor if a foot ulcer develops during treatment.

[Chunk 2] (doc: hartfalen_0168)
The use of SGLT2 inhibitors is well acceptable for most patients.
Contraindications for use include severe renal impairment, an active foot ulcer (exercise caution in patients with a history of foot ulcers), recurrent genital fungal infections, type 1 diabetes mellitus...

[Chunk 3] (doc: hartfalen_0158)
- Inform the patient about the necessity to temporarily discontinue the SGLT2 inhibitor and to contact the general practitioner in the following situations:... a non-healing wound on the foot... An SGLT2 inhibitor likely leads to fewer hospital admissions in patients with heart failure...

[Chunk 4] (doc: diabetes_0215)
- Continue the SGLT2 inhibitor if the eGFR decreases during use and discontinue it only at the start of renal dialysis... Inform the patient about the need to discontinue the SGLT2 inhibitor discontinue and contact your GP in case of: a non-healing wound on the foot...

[Chunk 5] (doc: hartfalen_0157)
```

- Be alert for side effects in patients at increased risk of dizziness and falls due to dehydration... If the eGFR decreases during use of the SGLT2 inhibitor, it is advised to continue it and only discontinue it at the start of dialysis...

**Failure Analysis & Elaboration:** This case which can be seen in Table 7 illustrates a Misinterpretation Failure (E10) where the generator preserves the correct medical concepts but distorts their logical strength. Although the retrieved context explicitly defines an active foot ulcer as a *relative* contraindication, the model upgrades it into an absolute contraindication.

This shift from conditional to categorical reasoning introduces an overly conservative interpretation of the guideline. As a result, a clinically nuanced recommendation is transformed into a stricter rule than intended, despite most surrounding claims being correctly grounded. This highlights that misinterpretation can arise not from missing information, but from incorrect preservation of logical constraints within retrieved evidence.

### B.3 Case Study: Cascading Multi-Error Structural Failure (E2, E4, E6, E7, E8, E9, E10)

This case study illustrates a systemic, multi-error cascading pipeline failure originating from a structural database defect:

Table 7: Detailed Breakdown of an Misinterpretation (E10) Analysis on SGLT2 Triage Guidelines

Metric Component	Content Value
Query ID	diabetes_0009_0005
User Query	State the clinical criteria and precautions required when initiating or managing an SGLT2 inhibitor in relation to foot ulcers.
Generated Response	Caution with an SGLT2 inhibitor is indicated in the following cases: <b>Active foot ulcer:</b> this is considered a contraindication; preferably do not start or discontinue temporarily. <b>History of foot ulcer:</b> this is a relative contraindication; weigh use with extra care. <b>Development of a non-healing wound/foot ulcer during treatment:</b> temporarily discontinue the SGLT2 inhibitor and contact the GP.
Response Claims	<ol style="list-style-type: none"> <li>Active foot ulcer is considered a contraindication, dictating that clinicians preferably do not initiate or temporarily discontinue the drug.</li> <li>A history of a foot ulcer is a relative contraindication, requiring clinical usage to be weighed with extra care.</li> <li>The development of a non-healing wound or foot ulcer during treatment requires a patient to temporarily discontinue the SGLT2 inhibitor and contact their GP.</li> </ol>
Accurate Claims	Claims 2 and 3 ( <i>Perfectly grounded in multiple adjacent context blocks across the heart failure and type 2 diabetes standards</i> ).
Misinterpreted Claims	<b>1. Active foot ulcer is considered a contraindication...</b> <i>Distortion Mechanism:</i> The generator model actively escalates a safety tier. The retrieved guidelines explicitly define an active foot ulcer as a <i>relative</i> contraindication. By presenting it as an absolute, unqualified contraindication (" <i>preferably do not start</i> "), the system introduces a logical distortion that restricts standard therapeutic parameters.
Misinterpretation Rate	<b>0.3333</b> (1 out of 3 total evaluated claims verified as structurally distorted).

underchunking (E2). This upstream fragmentation keeps a lot of different information in one chunk, causing an immediate domino effect that registers simultaneously as failures in Low ranked retrieval (E6), Missed Retrieval (E4), fabricated content (E8), incomplete answer (E9), misinterpretation (E10), and abstention failure (E7).

### Retrieved Context and Upstream Chunk Fragmentation

The following segments showcase the structural data division passed to the model, highlighting how the target context block was severed mid-sentence:

<p>[Chunk 1] (doc: chronische_nierschade_0090) Should the prescriber determine kidney function in elderly patients whose current kidney function is unknown and to whom they prescribe medication where impaired kidney function is a factor?</p>
<p>[Chunk 2] (doc: chronische_nierschade_0068) Protein restriction of 0.8 g/kg ideal body weight per day is advised for an eGFR &lt; 30 ml/min/1.73 m2. Should such a dietary intervention not already be advised for patients with chronic kidney damage with an eGFR &gt;= 30 ml/min/1.73 m2, particularly if they also have elevated albuminuria?</p>
<p>[Chunk 3] (doc: chronische_nierschade_0128) In the 2012 multidisciplinary guideline on Polypharmacy in the Elderly, impaired renal function is mentioned as a risk factor... in the staging of chronic renal damage, this cutoff value is not applied, and chronic renal damage is assumed at an eGFR &lt; 60 ml/min/1.73 m2... Additionally, the working group considers it important that current kidney function is known for patients with polypharmacy.</p>
<p>[Chunk 4] (doc: chronische_nierschade_0113) Is treatment with a statin (and/or ezetimibe) indicated in patients with chronic kidney damage to reduce the risk of cardiovascular and renal endpoints and mortality?</p>
<p>[Chunk 5] (doc: chronische_nierschade_0047) The treatment of chronic kidney damage is aimed at reducing the increased risk of cardiovascular morbidity/mortality and preventing or limiting the progression of kidney damage. - The general practitioner can manage a large proportion of patients themselves. This applies in any case to patients with:            - eGFR 30 ml/min/1.73 m in combination with normal albuminuria            - eGFR 30 ml/min/1.73 m in combination with normal albuminuria            - eGFR 45 ml/min/1.73 m in combination with moderately elevated albuminuria            - Determine the management based on:            - the risk category            - the cause and course of the chronic kidney damage</p>

- factors such as age and comorbidity
- the risk category
- the cause and course of the chronic kidney damage
- factors such as age and comorbidity
- A weighing of the pros and cons of strict (pharmacological) treatment is necessary, especially in patients with a limited life expectancy and/or extensive comorbidity. Make this assessment together with the patient.
  - If desired, treat patients with a significantly increased risk (red in Table 1) yourself who have not been referred to the internist-nephrologist due to a limited life expectancy and/or extensive comorbidity. In these patients, the emphasis is on identifying and usually in consultation with the internist-nephrologist pharmacologically treating metabolic complications (see section Monitoring). - For a summary of the management strategy by risk category, see Appendix 1. - eGFR 30 ml/min/1.73 m2 in combination with normal albuminuria - eGFR 45 ml/min/1.73 m2 in combination with moderately elevated albuminuria - the risk category - the cause and course of chronic kidney damage - factors such as age and comorbidity

### Golden chunk

"- For comprehensive recommendations regarding decision-making concerning renal replacement therapy, see the multidisciplinary guideline Renal Replacement Therapy (link is external). For care, it is important that: collaboration between primary and secondary care, with accessible consultation in case of doubt, can take place for sound decision-making, together with the patient, regarding whether or not to start renal replacement therapy; age is not a criterion for proposing renal replacement therapy, but comorbidity is; timely referral (eGFR < 30 ml/min/1.73 m2) to secondary care improves the outcomes of renal replacement therapy; patients with severe renal loss (eGFR < 30 ml/min/1.73 m2) receive information about the different forms of treatment for end-stage renal failure (renal replacement therapy by dialysis or kidney transplantation and conservative treatment), including the advantages and disadvantages of the interventions; - collaboration between primary and secondary care, with accessible consultation in case of doubt, can take place for sound decision-making, together with the patient, regarding whether or not to start renal replacement therapy - age is not a criterion, but comorbidity is for proposing renal replacement therapy - timely referral (eGFR < 30 ml/min/1.73 m2) to secondary care improves the outcomes of renal replacement therapy - patients with severe renal loss (eGFR < 30 ml/min/1.73 m2) receive information about the different forms of treatment for end-stage renal failure (renal replacement therapy by dialysis or kidney transplantation and conservative treatment), including the advantages and disadvantages of the interventions - If desired, use the consultation cards Options for renal replacement therapy(link is external) and Renal replacement therapy or conservative treatment?(link is external) during the conversation with the patient. - If foregoing renal replacement therapy is being considered, it is important that the patient and their family know what conservative treatment entails. This means that although patients do not receive renal replacement

Table 8: Systemic Multi-Error Diagnostic Matrix: Interlocking Upstream and Downstream Failures

Evaluation Dimension	Metric Values, Classification, and Grounding Status
<b>Query ID &amp; Context</b> <b>User Query</b> <b>Generated Response</b>	chronische nierschade.0013.0001 Which criteria are important when deciding whether or not to propose renal replacement therapy? When deciding whether or not to propose renal replacement therapy, such as dialysis or transplantation, no single criterion is decisive. The assessment should be based on: severity/risk category (eGFR/albuminuria), cause/course, age, frailty, comorbidity, life expectancy, expected treatment benefit, burden/disadvantages, complications, and patient wishes...
<b>Upstream Retrieval</b>	<ul style="list-style-type: none"> <li>• <b>Context Precision (E6): 0.2000</b> (Only 1 out of 5 retrieved chunks was semantically relevant; Chunks 1–4 discussed completely unrelated polypharmacy, diet, and statin regimens).</li> <li>• <b>Context Recall (E4): 0.0000</b> (0 out of 2 Gold Claims retrieved). The actual guideline criteria governing RRT thresholds were omitted during retrieval due to severe information fragmentation.</li> </ul>
<b>Downstream Generation</b>	<ul style="list-style-type: none"> <li>• <b>Faithfulness (E8): 0.7273</b> (3 out of 11 generated assertions were ungrounded fabrications). The model injected external clinical assertions regarding specific eGFR cutoffs and complications.</li> <li>• <b>Answer Recall (E9): 0.5000</b> (Incomplete Answer). The generator omitted a critical negative constraint required by the reference standard (“Age is not a criterion for or against RRT”).</li> <li>• <b>Misinterpretation (E10): 0.2500</b> (Logical Distortion). The generator mapped general, non-specialist chronic kidney disease management metrics onto specialized RRT/dialysis triage frameworks.</li> <li>• <b>Abstention Failure (E7): 0.0000</b> (Verdict: INSUFFICIENT — Confidence: 1.00). The system explicitly acknowledged a lack of context internally, yet failed to trigger a safe clinical abstention.</li> </ul>

therapy, they are treated as well as possible, aimed at maximizing the preservation of kidney function and the management of symptoms. - Patients and family may be hesitant to forgo renal replacement therapy because they feel they are choosing no treatment. The general practitioner, supported consultatively by the internist-nephrologist, can play an important role in this treatment and later in the palliative phase (for further information, see the guideline Palliative care in end-stage renal failure (link is external)) (see Details). - collaboration between primary and secondary care, with low-threshold consultation in case of doubt, can take place to ensure sound decision-making, together with the patient, regarding whether or not to start renal replacement therapy. - age is not a criterion, but comorbidity is for proposing renal replacement therapy. - timely referral (eGFR < 30 ml/min/1.73 m<sup>2</sup>) to the 2nd line improves the outcomes of renal replacement therapy - patients with severe renal loss (eGFR < 30 ml/min/1.73 m<sup>2</sup>) receive information about the different forms of treatment for end-stage renal failure (renal replacement therapy by dialysis or kidney transplantation and conservative treatment), including the advantages and disadvantages of the interventions In the Netherlands, over one million people have chronic kidney damage. For over 16,000 people, kidney function is so poor that they are dependent on renal replacement therapy in the form of dialysis or a kidney transplant. With the increase in older patients with end-stage renal failure, an increasingly large group of patients is also emerging who consciously choose not to undergo dialysis. For the group that does not start dialysis, treatment consists of both therapy aimed at maximizing the preservation of kidney function and treatment of symptoms. Depending on age, the degree of decline in kidney function in the preceding period, and comorbidity, this phase of maximum conservative therapy can last from months to years. The mortality The mortality rate of dialysis patients is high: half of the patients who start dialysis between the ages of 45 and 65 die within five years. In addition, we also see an increase in the number of frail elderly patients with severe kidney failure. The need for palliative care is high in this group. Within one year of starting dialysis, 58% of the most frail elderly have died or are functioning worse than before starting dialysis. 5 to 10% of frail elderly decide themselves within six months of starting dialysis. Both conservative therapy and dialysis can be associated with many side effects and complications of treatment (medication and dialysis). This severely limits the quality of life. The perceived symptom burden of patients with renal failure (eGFR < 15 ml/min/1.73 m<sup>2</sup>) is high and comparable to that of many cancer patients. The guideline on Palliative care in end-stage renal failure provides recommendations for the organization of care (with attention to the psychosocial and spiritual needs of patients, questions regarding the discontinuation of dialysis, and support for informal caregivers) and the treatment of frequently occurring symptoms (sleep disorders, pain, dyspnea, itching, restless legs, delirium, and depression).

preventing the retrieval component from accessing the complete evidence required to answer the query. This retrieval deficiency subsequently triggered additional downstream errors, including missed retrieval, low-ranked retrieval, misinterpretation, fabricated content, and incomplete answers.

The example highlights that chunking errors are not isolated failures. Instead, they can create a cascading effect in which information loss at the document processing stage propagates through retrieval and generation, ultimately degrading the quality and reliability of the final response.

**Failure Analysis, Correlation Elaboration:** This case study with results in Table 8 demonstrates how a single chunking failure can propagate throughout the entire RAG pipeline. Due to underchunking, the relevant information was part of a large chunk which changes the semantic embedding,