Failure analysis of RAG in healthcare

Nathaniël Apawti

1. Main 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?

2. Motivation

- **Surface-Level Testing:** Past studies on medical RAG systems almost exclusively evaluate the final output text as a single entity.
- **Dutch guidelines**: there is limited ressearch done about RAG systems in dutch clinical settings
- **Patient Risk:** In a clinical setting, these hidden system failures can be critical for the patient. A single ungrounded assertion such as an AI model altering a medication dosage threshold can directly result in dangerous situations.

3. Data

Clinical/Factual Benchmark:

- Questions
- Ground truth answer
- Source text

RAG System:

- Retrieved context
- Generated answer

clinical question example: Mrs. Bakker, 45 years old, has the flu and a fever. She asks if she can keep exercising. What non-medication advice would you give her?

Factual question example: When is a referral to the internist-nephrologist indicated for albuminuria?

5. Methodology

1. Data Layer & System Execution
2. Failure Attribution In taxonamy design Using LLM-as-judge
3. Aggregation & Analysis
4. Improvement Recommendations

Configuration settings

- retrieved chunk (K) = 5
- LLM for answer generation = gpt-55
- LLM-as-judge = gemini-3.1-flash-lite

6. Factual results

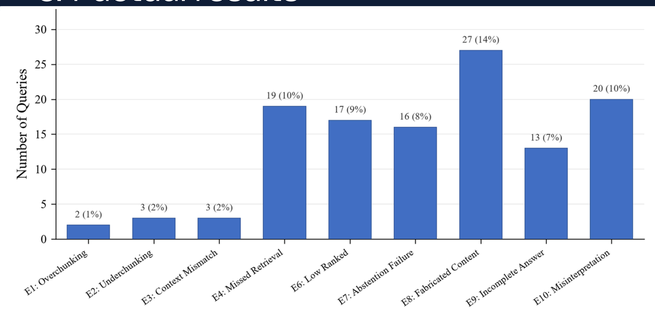

Figure 1: Total number of queries triggering each specific failure mode in the factual Figure 3: Distribution of the number of concurrent error types per query.benchmark.

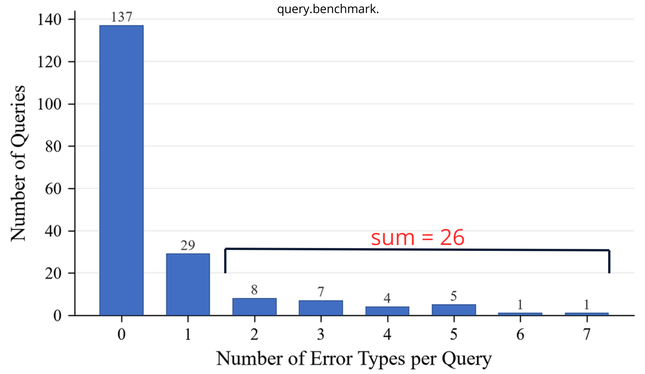

Figure 2: Distribution of the number of concurrent error types per query factual benchmark.

7. Clinical result

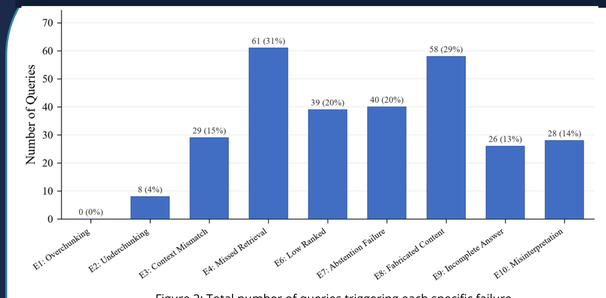

Figure 3: Total number of queries triggering each specific failure mode in the clinical benchmark.

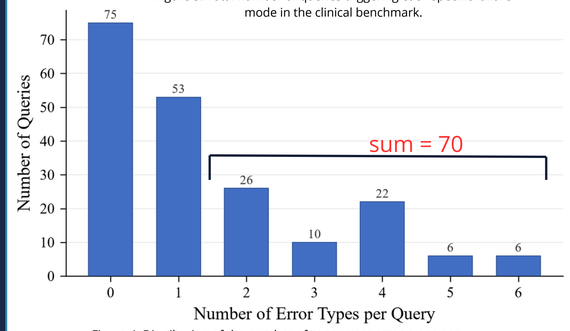

Figure 4: Distribution of the number of concurrent error types per query factual benchmark.

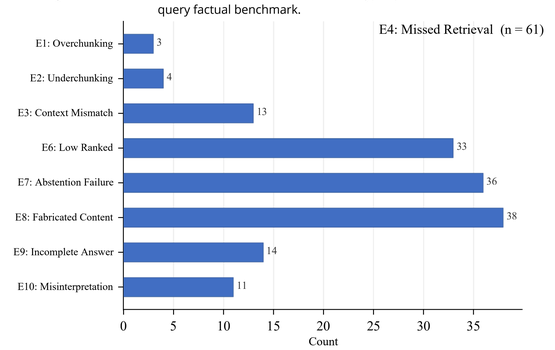

Figure 5: Co-occurence for Missed retrieval (E4) (sign of cascading failure)

4. Taxonamy design

Chunking Errors

E1: Overchunking
Semantically coherent information is fragmented across multiple chunks, preventing complete reconstruction.

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

E2: Underchunking
Large chunks contain excessive amounts of irrelevant information.

$$\text{IoU} = \frac{|\text{Groundtruth claims} \cap \text{Retrieved claims}|}{|\text{Groundtruth claims} \cup \text{Retrieved claims}|}$$

E3: Context Mismatch
Chunk boundaries break logical semantic relationships between entities.

$$\text{RelationRecall} = \frac{\text{preserved relations}}{\text{gold relations}}$$

Retrieval Errors

E4: Missed Retrieval
Required ground-truth information is absent from the retrieved context.

$$\text{ContextRecall} = \frac{\text{gold claims in retrieved context}}{\text{total gold claims}}$$

E5: Low Relevance
Retrieved chunks are not relevant for answering the user's question.

$$\text{ContextPrecision} = \frac{\text{relevant retrieved chunks}}{\text{retrieved chunks}}$$

E6: Low Ranked
Relevant chunks are retrieved but appear too low in the ranking order.

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

Generation Errors

E7: Abstention Failure
The system generates an ungrounded response instead of admitting missing evidence.

Evaluated via LLM-as-a-judge (No formula)

E8: Fabricated Content
The response introduces hallucinated claims unsupported by the retrieved evidence.

$$\text{Faithfulness} = \frac{\text{supported claims}}{\text{total claims in response}}$$

E9: Incomplete Answer
The response only partially addresses the query despite having the necessary evidence.

$$\text{Answer Recall} = \frac{\text{gold claims in answer}}{\text{total gold claims in context}}$$

E10: Misinterpretation
The generated claims misunderstand or incorrectly reformulate the retrieved evidence.

$$\text{Misinterpretation Rate} = \frac{\text{Misinterpreted Claims}}{\text{Supported Claims}}$$

- One Instance can be assigned to mulitple failures
- Metrics are used to reduce evaluation bias
- mostly claim level metrics , for example:
  - **Answer**: The most likely diagnosis is osteoarthritis of the CMC1 joint. The symptoms of pain, swelling, and stiffness fit with this condition.
  - **extracted claims: 1.** The most likely diagnosis is osteoarthritis of the CMC1 joint. **2.** The symptoms of pain, swelling, and stiffness fit this condition.

8. Finetuning strategies

Retrieval Fine-Tuning

- **The Strategy:** Deploy domain-specific retriever fine-tuning via contrastive learning on Dutch primary care datasets.
- **Frameworks**: Implement specialized frameworks like REMED or CERA to realign the clinical embedding space.

Generation Fine-Tuning

- **The Strategy:** Integrate retrieval-aware generation training to teach the generator to recognize context deficits, ignore noisy distractors.
- **Frameworks**: Implement RAFT or Self-RAG strategies to improve grounding and prevent hallucination

9. Conclusion

- **Most dominant failures**
  - Clinical benchmark: E4: Missed Retrieval (31%)
  - Factual benchmark: E8: Fabricated content (14%)
- **Cascading Failure**
  - from E4 (Missed Retrieval) to (E8 Fabricated) content and E9 (Abstention Failure)
- **Finetuning strategies**
  - realign clinical embedding space with REMED/Cera
  - improve grounding with RAFT or Self-RAG