Charlene Bakker
Delft University of Technology
Supervisors: Dr.ir. J. Yang, Yannick ter Heerdt
Thesis committee: Pradeep Murukannaiah

Evaluating RAG in Dutch Primary Care: A Benchmark over NHG Guidelines

Background

- **Why this matters:** Standard LLMs (e.g. GPT-4) rely on static training data, risking hallucinations and outdated clinical advice. RAG mitigates this by retrieving information from verified sources, and Dutch systems like AskAletta and De Digitale Arts already apply this to NHG guidelines for primary care.
- **The critical gap:** No standardized, automated benchmark exists to evaluate factual QA adherence to NHG guidelines. Human evaluation is accurate but does not scale to continuous clinical deployment.
- **This work:** An automated pipeline that constructs and validates a factual QA benchmark over NHG guidelines, designed so that each pair is factually correct, clinically relevant, and retraceable to the source text.

Research Question

How can a reliable automated benchmark be constructed for general factual Q/A over the NHG-guidelines?

Methodology

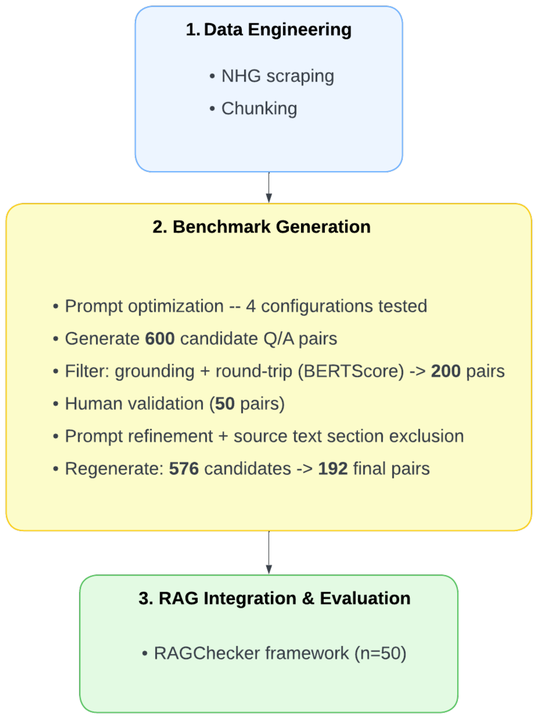

Figure 1. Overview of the methodology pipeline.

Results

Before generating the final dataset, a prompt experiment was conducted on 30 chunks across the 10 guidelines to determine most faithful prompt strategy.

Prompt Strategy	Precision	Recall	F1
Zero-Shot	0.739	0.512	0.604
Chain-of-Thought (CoT)	0.723	0.510	0.597
Few-Shot	0.761	0.565	0.648
Few-Shot + CoT (Final)	**0.764**	**0.572**	**0.654**

Table 1. BERTScore grounding across prompt strategies.

- **Key Finding:** Combining Few-Shot with Chain-of-Thought achieved the highest grounding ($F1 = 0.654$) and was selected as the final prompt configuration.

Benchmark Generation & Refinement:

Generated 576 candidate pairs (3 per chunk); retained the best pair per chunk using BERTScore grounding and round-trip consistency, yielding **192 final QA pairs**.

Human Validation - Before & After Refinements:

100%
Factual Correctness
both rounds

100%
Retraceability
both rounds

64% → 100%
Clinical Relevance
after refinements

RAGChecker vs. Human Evaluation:
RAGChecker scores more strictly than humans, but follows the same overall trend – making it a viable automated alternative.

Metric	Human	RAGChecker
Factual Correctness / Precision	72.0%	57.7%
Faithfulness	72.0%	67.5%

Table 2. Human evaluation vs. RAGChecker scores (n=50).

Where they disagree:

- Correct answers that are *paraphrased*.
- Correct answers with *additional but faithful* clinical details not in the ground truth.

Conclusion

- **Reliable benchmark construction is feasible:** The pipeline automatically generates clinically relevant, factually correct QA pairs with no manual expert annotation, yielding **192 QA pairs** across 10 NHG guidelines.
- **Refinements are impactful:** Prompt adjustments and source filtering raised clinical relevance from **64% to 100%**.
- **RAGChecker is a viable but strict evaluator:** RAGChecker follows the same trends as human judgment but consistently scores lower (precision: 57.7% vs. 72%), due to strict claim-level checking that penalizes extra correct information not present in the ground-truth answer.

Future Work & Limitations

Limitations

- **Limited coverage:** The benchmark covers only 10 NHG guidelines, limiting generalizability across the full range of Dutch primary care topics.
- **BERTScore constraints:** Differences between prompt strategies are moderate and BERTScore measures semantic similarity only, not clinical correctness.
- **Single annotator:** All human validation was performed by one annotator, introducing subjective bias and limiting inter-rater reliability.

Future Work

- **Expand coverage:** Scale to the full range of NHG guidelines and increase QA pairs per guideline for more robust evaluation.
- **Multiple annotators:** Involve multiple clinical experts (e.g. Dutch GPs) to establish inter-rater reliability and strengthen benchmark validity.
- **Improve RAGChecker:** Adjust the precision metric to not penalize additional correct claims beyond the reference answer, better suiting open-ended clinical QA.

References

[1] AskAletta.
About us - ask aletta.
https://askaletta.com/en/about, 2026.
Accessed: April 22, 2026.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.
Language models are few-shot learners, 2020.

[3] De Digitale Arts.
Over de ai tool - de digitale arts.
https://dedigitalearts.nl/about-tool, 2026.
Accessed: April 22, 2026.

[4] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang.
Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation, 2024.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou.
Chain-of-thought prompting elicits reasoning in large language models, 2023.

[6] Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer.
Selecting better samples from pre-trained llms: A case study on question generation, 2022.

[7] Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, and William Hiesinger.
Almanac: Retrieval-augmented language models for clinical medicine, 2023.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi.
Bertscore: Evaluating text generation with bert, 2020.