



**Automated Benchmark Construction for
Factual Question Answering over NHG Guidelines**

A Foundation for RAG Evaluation in Dutch Primary Care

Charlene Bakker¹

Supervisor(s): Dr.ir. J. Yang, Yannick ter Heerdt

¹EEMCS, 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: Charlene Bakker
Final project course: CSE3000 Research Project
Thesis committee: Dr.ir. J. Yang, Pradeep Murukannaiah

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

Abstract

Evaluating Retrieval-Augmented Generation systems in clinical domains requires reliable benchmarks, yet constructing these manually is costly and infeasible at a large scale. This paper presents an automated pipeline for constructing and evaluating a factual question answering benchmark over Dutch primary care guidelines. The pipeline uses large language model based question-answer generation with few-shot and chain-of-thought prompting, combined with automated filtering using BERTScore grounding and round-trip consistency to produce high quality question-answer pairs. Human validation confirmed that the final benchmark of 192 question-answer pairs across 10 Netherlands Huisartsen Genootschap guidelines achieves factual correctness, retraceability and clinical relevance. The benchmark was integrated into a Retrieval-Augmented Generation pipeline to evaluate whether RAGChecker, a claim-level automated evaluation framework, could serve as a reliable alternative to human evaluation. RAGChecker scores were consistent with human judgment though lower due to its strict claim-level checking. These results show that a reliable, automated benchmark can be constructed for Dutch primary care question answering and that RAGChecker serves as a reasonable but strict alternative for human evaluation of Retrieval-Augmented Generation systems in this domain.

1 Introduction

Artificial Intelligence (AI) and Large Language Models (LLMs) are becoming increasingly important and relevant in the medical field to support clinical tasks. However, standard models rely on static training data, which can lead to hallucinations, outdated information, and conflicting recommendations, which makes them unsafe for clinical decision-making [19]. Retrieval-Augmented Generation (RAG) mitigates these risks by dynamically retrieving information from verified external sources before generating a response, and has already shown promising results in clinical settings [19].

In the Netherlands, the *Nederlands Huisartsen Genootschap* (NHG) provides the official clinical guidelines for primary care. Initiatives such as AskAletta, explorations by the NHG itself and De Digitale Arts have successfully developed early RAG-based systems to assist practitioners [1, 4, 12]. However, a critical knowledge gap remains: evaluating the reliability and factuality of these systems is a complex challenge. Although frameworks have recently been introduced to evaluate RAG systems in Dutch hospital environments [7], there is currently no standardized, automated benchmarking framework specifically designed to assess factual question answering (QA) adherence to NHG guidelines in primary care. Human evaluation is the natural alternative, but it is costly, labor-intensive, and does not scale to the demand for continuous clinical deployment [11, 16]. Without a robust framework to automatically and reliably judge whether an LLM’s output aligns with NHG standards, safe clinical deployment remains limited.

To address this gap, this paper proposes an automated pipeline for constructing and validating a factual QA benchmark over NHG guidelines, defined here as a curated set of question-answer (Q/A) pairs with verified ground-truth answers used to evaluate system outputs. Here, *reliable* refers to benchmark quality: each Q/A pair of the benchmark must be factually correct, clinically relevant, and retraceable to the NHG source text. The pipeline combines LLM-based Q/A generation with automated filtering using two metrics:

BERTScore grounding, which measures semantic similarity between two texts using contextual embeddings, and round-trip consistency, which regenerates an answer to the original question using the source context and checks whether it matches the original answer, confirming that the question is answerable and unambiguous [18, 20]. This is followed by human validation to ensure clinical quality. The resulting benchmark of 192 Q/A pairs is then integrated into a RAG pipeline, providing a foundation for factual QA evaluation of RAG-based systems over NHG guidelines. This evaluation uses the RAGChecker framework, a claim-level method that scores generated answers by checking individual claims against the reference [15]. The main research question that this work aims to answer is: **"How can a reliable automated benchmark be constructed for general factual QA over the NHG guidelines?"**. This research intentionally focuses specifically on factual QA, defined here as questions with a single, directly retraceable answer in the NHG source text. This is different from clinical QA, which requires reasoning over NHG content in a realistic medical setting. This work is conducted in parallel with a peer who is developing the benchmark for clinical QA. Together, these efforts aim to facilitate systematic evaluation of RAG-based systems for primary care in the Netherlands.

Human validation showed that the constructed, refined pipeline achieved 100% acceptance across three criteria (re-traceability, clinical relevance and factual correctness), compared to 64% clinical relevance acceptance before refinements, demonstrating the effectiveness of the proposed approach.

The main contributions of this paper are:

- An automated pipeline for constructing factual QA benchmarks over Dutch clinical guidelines, combining LLM-based generation with automated filtering.
- A validated benchmark of 192 Q/A pairs derived from 10 NHG guidelines.
- A human validation demonstrating the impact of prompt refinements and source text filtering on clinical relevance.
- An evaluation of RAGChecker as an automated evaluation framework for factual QA over Dutch NHG guidelines.

The remainder of the paper is structured as follows. Section 2 describes literature related to this work. Section 3 describes the methodology and pipeline design. Section 4 presents and discusses the results of experiments and validations. Section 5 reflects on the responsible research considerations of this work and Section 6 concludes the paper.

2 Related Work

2.1 Medical QA Datasets and Benchmark Construction

The construction of benchmarks for medical QA has gained increasing interest since LLMs are being increasingly applied in the medical field to support clinical tasks. Initial benchmarks such as MedQA [9] introduced large-scale datasets based on medical board exams to evaluate LLMs on solving medical problems across diverse specialties. However, these benchmarks are mainly based on multiple-choice questions, which do not focus on evaluating free-form clinical reasoning. So, these benchmarks do not necessarily show the model's

ability to produce clinically valid open-ended answers [2].

A recent effort to benchmark RAG specifically for the medical field is MIRAGE [14], which consists of questions from five different QA datasets. However, MIRAGE is limited to existing English-language multiple-choice datasets and does not consider free-form factual QA over domain-specific medical guidelines.

Automated Q/A generation using LLMs has appeared as a scalable alternative to manual annotation. Prompt engineering techniques such as few-shot [3] and chain-of-thought prompting [17] can improve the quality of generated Q/A pairs. However, in domain-specific settings the reliability of such benchmarks is highly dependent on the quality of the source text, prompt design and how output is validated, all of which remain challenges.

More recent work has explored automated benchmark construction specifically for RAG-based systems in the Dutch healthcare. Gonzalez Torres et al. [7] automatically generated Q/A pairs from over 1000 Dutch hospital leaflets and evaluated RAG chatbots through cross-evaluation over topically disjoint folds. Domain specialists confirmed over 83% of the generated pairs as relevant, demonstrating the feasibility of automated benchmark construction in a Dutch medical setting. However, their work focuses on a secondary-care hospital context and does not address primary care guidelines such as those of the NHG.

2.2 RAG Evaluation

Evaluating RAG-based systems is more complex than evaluating standard LLMs, since performance consists of both retriever and generation dimensions [6, 14]. Belmadani et al. [2] note that automated metrics such as BLEU [13], ROUGE [10] and BERTScore [20] are poorly suited for evaluating the clinical correctness of medical QA system outputs, since they capture surface-level text similarity rather than medical accuracy. This limitation, however, does not apply to the task for which BERTScore is used in this work, since it is only used to check whether Q/A pairs of the benchmark are grounded in the source text, which is exactly the kind of similarity it is built to measure. RAGAs [6] introduces a suite of metrics, e.g. faithfulness, answer relevance and context relevance, that can be used to evaluate the RAG-system dimensions without requiring human ground-truth references. However, the framework shows limited interpretability into where failures originate from.

RAGChecker addresses this limitation with a fine-grained evaluation framework, which provides information on the origin of failures in the different dimensions. Using claim-level entailment checking the framework provides metrics such as context precision, claim recall, noise sensitivity, hallucination and faithfulness. A meta-evaluation showed that RAGChecker correlates more strongly with human judgements than alternative frameworks, making it well-suited for high-risk domains such as primary care where identifying the source of errors is crucial [15].

This work builds directly upon these frameworks. Following the approach of Gonzalez Torres et al. [7], automated LLM-based Q/A generation with prompt engineering and filtering is used to construct a benchmark tailored to the NHG guidelines. RAGChecker is adopted as the primary evaluation framework given its claim-level evaluation and a stronger

correlation with human judgment compared to alternatives such as RAGAs [15].

Since faithfulness and groundedness are central to RAG evaluation [6, 15], a benchmark’s ground-truth answers must themselves be retraceable to the source text to provide a reliable ground truth. Combined with the gaps identified above: free-form rather than multiple-choice questions, clinical rather than study-level relevance, and automated construction [7], this motivates three desiderata for a benchmark over NHG guidelines: answers must be retraceable to the source text, questions must reflect clinically relevant GP information needs, and construction must be automated. This work addresses all three through a prompt optimization experiment that identifies the most faithful prompt configuration, and human validation confirming that generated pairs are factually correct, retraceable, and clinically relevant.

3 Methodology

To answer the main research question, a three-stage pipeline was developed to construct and validate a reliable, automated factual QA benchmark over the NHG guidelines. An overview of the pipeline can be seen in Figure 1 and detailed below.

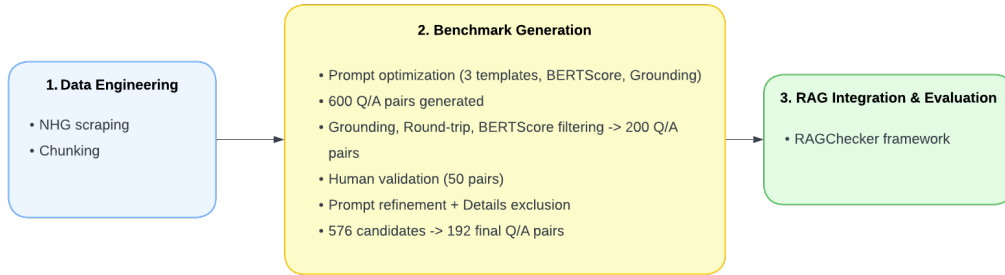


Figure 1: Overview of the automated benchmark construction and evaluation pipeline.

3.1 Data Engineering

The first phase involved collecting data from the NHG guidelines from the official website.

- **NHG Scraping:** For the scope of this project, data from a subset of 10 distinct NHG guidelines was scraped to form the foundational source text. The 10 NHG guidelines that were scraped: *Astma bij volwassenen*, *Diabetes*, *Dementie*, *Depressie*, *COPD*, *Angst*, *Chronische nierschade*, *Influenza*, *Hartfalen* en *Hand en polsklachten*.
- **Text Chunking:** The scraped raw text was segmented into distinct semantic chunks using the `spaCy` Dutch language NLP pipeline (`n1_core_news_sm`) for sentence boundary detection [8]. Chunks were compiled by grouping full sentences up to a maximum of 400 words per chunk, preserving textual integrity and structural context. This maximum was motivated by the use of BERTScore for evaluation, which relies on BERT-based encoders with a maximum limit of 512 tokens [5] (can be roughly translated to around 400 words).

3.2 Benchmark Generation

The second phase involved the full process of constructing the benchmark, from prompt optimization through candidate generation, filtering and iterative refinement.

3.2.1 Prompt Optimization

Before generating the full benchmark, a prompt experiment was conducted to find the prompt that produces the most faithful Q/A pairs. This experiment was conducted on a sample of 30 chunks extracted from 10 guidelines (3 chunks per guideline). This sample was chosen to cover all ten guidelines while remaining computationally feasible. Three chunks per guideline was deemed sufficient to capture variation within each guideline without running the full scraped data for each prompt candidate. Three prompting techniques were evaluated using BERTScore to measure grounding against the source text. BERTScore was chosen over surface-level metrics such as BLEU and ROUGE for its use of contextual embeddings to capture semantic similarity rather than exact word overlap [20]. The techniques evaluated were:

1. **Zero-Shot:** Model receives only task instruction without examples.
2. **Few-Shot:** Model receives task instruction with output examples [3].
3. **Chain-Of-Thought:** Model is prompted to output its reasoning steps before formulating a response [17].

The best performing technique was chosen as a baseline for the prompt. After that, a combination between two techniques was evaluated to see if that would improve the prompt. The final prompt strategy was chosen based on highest BERTScore grounding.

Using the optimized prompt design, an initial benchmark was created. For each of the 10 NHG guidelines, 20 source chunks were selected, which yielded 200 chunks in total. The 20 chunks per guideline was chosen to ensure broad coverage across the guidelines while remaining feasible both in terms of downstream human validation and model API calls, since generating pairs for all chunks would exceed the scope of the project.

Candidate Pair Generation For each chunk, the model generated three pairs, resulting in 600 candidate pairs. Generating multiple candidates per chunk increases the chance of obtaining a high-quality pair, since prompt-based generation can create variable results even with the same input.

Filtering To filter out high-quality pairs per chunk, each candidate pair was scored on two metrics:

1. **BERTScore Grounding:** Selects candidate pair whose generated answer has the highest semantic similarity to the chunk. [20].
2. **Round-trip Consistency:** The generated question of the candidate pair, along with the source chunk, is given to the model, and the model is asked to generate an answer to this question using the source chunk. A high similarity between the reproduced answer and the original answer of the pair shows that the question is unambiguous and can be answered from the chunk alone [18].

These two scores were combined into a single score and the highest-scoring candidate was kept for each chunk. After the filtering step, the pipeline yielded an initial benchmark of **200 factual Q/A pairs**.

3.2.2 Refinement & Regeneration

Human Validation To assess the quality of the initial benchmark, a random subset of 50 Q/A pairs (25% of the total benchmark) was manually evaluated on three criteria inspired by prior work in medical QA evaluation [19]: **factual correctness**, **re-traceability of the answer to the source text**, and **clinical relevance for a general practitioner (GP)**. Each criterion was scored on a binary scale (Yes/No). This provided a representative sample for quality assessment while remaining feasible for single-annotator review. Table 1 shows an example of a rejected and an accepted pair from this validation.

Table 1: Example Q/A pairs from the initial human validation round (translated from Dutch).

	Pair #1 (Accepted)	Pair #2 (Rejected)
Chunk ID	astma_bij_volwassenen_0008_0022	chronische_nierschade_0010_0010
Guideline	Asthma in adults	Chronic kidney disease
Question	What alternative is mentioned for adding a LABA to a low-dose ICS in asthma maintenance treatment?	What clinical outcomes was intensive treatment associated with in the Steno-2 study in diabetic patients with moderately elevated albuminuria?
Answer	An alternative is an intermediate dose of ICS without LABA.	Intensive treatment was associated with a reduced risk of cardiovascular disease, nephropathy, retinopathy, and autonomic neuropathy.
Retraceable	Yes	Yes
Clinically relevant	Yes	No
Factually correct	Yes	Yes
Remarks	—	Refers to a study (by name).

Refinements Based on the findings of the first human evaluation, two changes were applied. First, the prompt was adjusted to include negative constraints and an example to instruct the model to avoid generating questions based on study-level evidence (e.g. research names, sample sizes, effect sizes or diagnostic accuracy statistics) or content that lacks direct applicability in clinical practice. The prompt is provided in Appendix A. Additionally, *Details* sections were excluded from the scraped text during the Data Engineering phase, since these were found to mostly contain study-related information rather than clinical recommendations. The pipeline was rerun with the refined prompt and updated source text. A second round of human validation was then done on the regenerated benchmark. The results of these human validations are shown in Section 4.

3.3 RAG Integration & Evaluation

The final phase consisted of integrating the constructed benchmark into the RAG pipeline constructed by another peer and evaluating an automated evaluation framework to determine its effectiveness and fit for this scenario. As discussed in Section 2, two frameworks were considered for this: RAGAs [6] and RAGChecker [15]. RAGAs was not chosen for this evaluation since it operates at response-level and does not decompose answers into individual claims. In a clinical setting, a partial incorrect answer can be misleading or harmful, making claim-level evaluation preferable. RAGChecker was therefore selected as the primary evaluation framework, since it decomposes answers into claims and checks each claim individually, which makes it easier to pinpoint where the failures are coming from.

For the purpose of evaluating factual QA performance, four metrics of RAGChecker were most relevant: **overall precision**, **overall recall**, which measures factual correctness and completeness of the generated answers against the reference, **faithfulness**, which measures whether the generated answers are grounded in the reference and **hallucination**, which measures the amount of incorrect claims that are not grounded in the source text.

Human Validation of RAGChecker Output To validate the applicability of RAGChecker for this specific scenario, a random subset of generated answers by the RAG pipeline were evaluated manually on two criteria: factual correctness and traceability of answer to the source text. These correspond to RAGChecker’s claim precision and faithfulness metrics, which allows a direct comparison between human and automated judgment. The evaluated pairs consisted of the integrated RAG pipeline’s generated answers, contexts, benchmark questions, and ground-truth answers. The results of this comparison are shown in Section 4.

4 Results

This section reports the outcomes of the benchmark construction and validation pipeline. The results are organized according to the order described in the Methodology section: the prompt optimization experiment, the human validation of the initial benchmark, the impact of refinements and the RAGChecker validation.

4.1 Prompt Optimization

Table 2 shows the BERTScore grounding performance across four prompting configurations evaluated on a sample of 30 chunks to establish a prompt configuration.

Table 2: Overall BERTScore grounding performance across prompt configurations.

Prompt Strategy	Precision	Recall	F1-Score
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

Few-Shot prompting scored highest among the individual strategies (F1: 0.648), and adding Chain-of-Thought to it increased the score to 0.654. The improved performance of Few-Shot + CoT is likely because the model is prompted to first reason through the medical content before formulating a question, which can reduce generating ambiguous pairs. Both outperformed Zero-Shot and Chain-of-Thought alone, so Few-Shot + CoT was selected as the final prompt configuration.

4.2 Benchmark Generation

4.2.1 Human Validation

Table 3 compares the human validation results across a sample of 50 pairs before and after the prompt and pipeline refinements.

Table 3: Human validation results before and after refinements.

Round	Pairs evaluated	Factual correct	Retraceable	Clinically relevant
Round 1 (before)	50	100%	100%	64%
Round 2 (after)	50	100%	100%	100%

All 50 pairs passed the factual correctness and retraceability criteria, indicating that the pipeline effectively generated and filtered highly-factual and well-grounded pairs. However, 36% of the pairs in round 1 were deemed not to be clinically relevant, often targeting study-related information or additional background information instead of actionable clinical content. A representative example of a rejected pair is shown in Table 1 (Pair #2), where the generated question refers to the Steno-2 study and asks about trial outcomes rather than a direct clinical recommendation. This finding led to the prompt changes and exclusion of the *Details* sections as described in Section 3.2.2.

As a result of removing the *Details* sections, the *Influenza* guideline contained only 12 eligible chunks after filtering, reducing the final benchmark to **192 factual Q/A pairs** generated from 576 candidates.

Following the refinements, a second round of human validation was done on a random subset of 50 pairs from the final benchmark of 192 pairs using the same criteria. All Q/A pairs passed all three criteria, which showed that the applied changes substantially increased the clinical relevancy of the generated benchmark.

4.3 RAGChecker Evaluation

Table 4 shows the comparison between human validation and RAGChecker’s automated scores. Human validation yielded a score of 72% for both factual correctness and faithfulness, while RAGChecker’s evaluation yielded a precision score of 57.7% and a faithfulness score of 67.5%.

Table 4: Comparison between Human Evaluation and RAGChecker

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

The two evaluation methods tended to agree on unambiguous cases: answers that closely matched the ground-truth answer and were traceable to the source chunk were accepted by both, while answers containing information absent from the reference or source were rejected by both. Table 5 shows one example of each. Disagreement happened in two scenarios: paraphrased but correct answers, and correct answers containing additional faithful information.

Table 5: Examples of agreement between human evaluation and RAGChecker (translated from Dutch).

	Accepted by both	Rejected by both
Question	What characteristics indicate ABCDE instability in an acute asthma attack?	Which medications should the GP ask about as a possible cause of respiratory complaints?
Generated answer	Cyanosis, decreased consciousness and/or exhaustion.	No answer found in the context.
Human / RAGChecker	Yes, Yes – Precision: 1.0, Faithfulness: 1.0	No, No – Precision: 0.0, Faithfulness: 0.0

First, disagreement occurred when generated answers were *correct but paraphrased*. Table 6 shows an example where the generated answer states the same clinical fact as the ground truth, just differently phrased. Human validation rated it fully correct, while RAGChecker scored it 0.5 on precision despite a faithfulness score of 1.0, which shows that its claim matching can penalize correct answers that are phrased differently.

Table 6: Example of a paraphrased answer accepted by human validation but penalised by RAGChecker (translated from Dutch).

Question	Which RUDAS score is indicative of dementia in the context of dementia diagnostics?
Ground truth	A RUDAS score < 22 is indicative of dementia.
Generated answer	A score of less than 22 on the RUDAS is indicative of dementia.
Human / RAGChecker	Yes, Yes – Precision: 0.5, Faithfulness: 1.0

Second, disagreement happened when generated answers included *additional but faithful* clinical information from the retrieved context but absent in the ground-truth answer. Table 7 shows such a case, where the answer correctly addresses the question and adds a contextually grounded recommendation on medication. Human validation accepted this as clinically correct, whereas RAGChecker assigned a precision score of 0.27 despite a faithfulness score of 1.0, since the additional claim could not be matched against the reference answer.

Table 7: Example of an answer with additional faithful information, accepted by human validation but penalised by RAGChecker on precision (translated from Dutch).

Question	What should the GP do when starting an antidepressant in patients older than 70 years with a depressive disorder?
Ground truth	The GP should pay close attention to evaluating the effect and side effects, due to the likely lower effectiveness of antidepressants and greater sensitivity to side effects compared to younger adults.
Generated answer	When starting an antidepressant in patients older than 70 years with a depressive disorder, pay close attention to evaluating the effect and side effects, due to the likely lower effectiveness of antidepressants and greater sensitivity to side effects compared to younger adults. Citalopram or sertraline are first-choice medications due to experience, a favorable side-effect profile, limited interactions and risks in case of overdose, and ease of use.
Human / RAGChecker	Yes, Yes – Precision: 0.27, Faithfulness: 1.0

4.4 Discussion

The results show that a reliable, automated factual QA benchmark over Dutch primary care guidelines can be constructed without requiring human experts to manually write or verify each Q/A pair. The prompt optimization experiment showed that combining Few-Shot with Chain-of-Thought produces the most grounded Q/A pairs. This builds on the individual strengths of both techniques, where Few-Shot prompting improves response quality through examples [3] and Chain-of-Thought encourages the model to reason through content before generating a response [17]. Furthermore, the two rounds of human validation confirm that the pipeline, after changes, generates pairs that are retraceable, factually correct, and clinically relevant, which addresses the reliable part of the research question.

Regarding the evaluation of the chosen automated metrics, RAGChecker scored lower than human judgment on both precision and faithfulness. As shown in Section 4.3, this gap comes from RAGChecker’s strict, claim-level matching: it penalizes additional, factual correct information not present in the ground-truth answer and also flags answers that are paraphrased but correct. Human annotators, on the contrary, judge answers as a whole and are more tolerant of such deviations. Overall, RAGChecker appears to be more conservative than human evaluation but follows the same overall trends, suggesting it can serve as

a reliable automated alternative in this domain as long as its strictness is taken into account.

There are some limitations to consider in this work. First, the benchmark only covers 10 NHG guidelines, which means that the results may not generalize fully to the full range of Dutch primary care topics. Second, the differences in the BERTscore results between prompt strategies in the prompt experiment are moderate, so it is hard to say whether these actually reflect meaningful differences in quality in practice. Third, all human evaluations were done by a single annotator (the author), which introduces a risk of subjective bias and limits inter-rater reliability. Therefore, the human scores should be interpreted as indicative rather than fully robust ground truth.

5 Responsible Research

Conducting responsible research requires carefully reflecting on the ethical aspects of the methods used and ensuring transparency and reproducibility of the work. This section describes the ethical considerations related to the use of medical guidelines, steps taken to make the work reproducible and use of generative AI.

5.1 Ethical Considerations and Limitations

Because this study focuses on Dutch primary care, the clinical accuracy of the generated data is very important. However, it should be noted that this research only used publicly available, aggregated medical guidelines from the Nederlands Huisartsen Genootschap (NHG). No real patient data, electronic health records, or any personal identifiable information (PII) was used during this research. As a result, formal medical ethical approval was not necessary. This also limits privacy risks, since all data that was used is publicly available and non-sensitive.

Furthermore, given the limitations that were mentioned in Section 4.4, the framework developed in this thesis is only intended as an exploration of automated evaluation workflows and should not be used for direct clinical decision-making without expert verification.

5.2 Reproducibility

To ensure that the experiments and evaluation methods are fully reproducible, transparency was maintained throughout the research. The source text consists of official, publicly available NHG guidelines. To ensure consistency, the same language model (GPT-5.5) was used for the prompt experiment, benchmark generation, and evaluation steps in the pipeline. The only exception was the RAG pipeline and the RAGChecker framework, which both used GPT-4o. Also, to reduce randomness in results, all LLM calls were run with fixed temperature settings. Finally, all prompt configurations, prompt experiment results, evaluation configurations, the subset used for the last evaluation and the full implementation can be found in the codebase.

5.3 Use of Generative AI

Generative AI tools, specifically Gemini, Github Copilot, and Claude, were used in a supporting role during this research, in compliance with the TU Delft guidelines on the use of generative AI in projects. The tools were used to assist with academic writing (grammar, style improvements), to support coding tasks, and to brainstorm and refine ideas during the research process. The author holds full responsibility for all ideas, arguments, interpretations, and written content. All AI-assisted output was reviewed and verified before included in the work.

6 Conclusion and Future Work

This paper addressed the following research question: **How can a reliable automated benchmark be constructed for general factual Q/A over the NHG guidelines?**. To answer this, a three-stage pipeline was developed that automatically constructs and evaluates a factual QA benchmark over Dutch primary care guidelines without requiring human experts to manually write or verify every Q/A pair.

The pipeline combines LLM-based Q/A generation with automated filtering using BERTScore grounding and round-trip consistency, followed by refinements based on human validation. The final benchmark consists of 192 factual QA pairs across 10 NHG guidelines, of which a subset passed human validation on factual correctness, retraceability and clinical relevance, after refinements. This demonstrates that the proposed pipeline can reliably construct a clinically relevant benchmark in a domain where manual annotation is costly.

RAGChecker was evaluated as an automated alternative to human evaluation for scoring the RAG pipeline on the benchmark. Its scores were consistently lower than human judgment but followed the same overall trend, suggesting that it can serve as a reasonable, though stricter, alternative for human evaluation in this domain. This strictness mainly stems from its claim-level matching, which penalizes generated answers containing additional, factually correct information beyond the ground-truth answer, a notable mismatch for medical QA, where such completeness is often valuable rather than problematic.

Future work could focus on multiple directions. First, the benchmark currently only covers ten NHG guidelines, which limits generalizability. Expanding to the full range of NHG guidelines and increasing the number of Q/A pairs per guideline would improve coverage and allow more robust evaluation of RAG systems across the full range of primary care topics. Second, human validation in this work was performed by a single annotator, introducing a risk of subjective bias. Future work should involve multiple annotators, ideally clinical experts such as Dutch general practitioners, to have inter-rater reliability and strengthen the validity of the benchmark. Third, RAGChecker could possibly be extended for this scenario by adjusting the precision metric to not penalize additional factually correct claims that go beyond the reference answer, which better addresses the needs of open-ended medical QA.

A Prompt Template

Few-Shot + CoT Prompt

Note: The prompt was originally written in Dutch to align with the language of the NHG guidelines. The following is a translated version for readability.

Given the text below from an NHG guideline, generate 3 specific question-answer pairs.

The question must:

- Be answerable solely based on the given text
- Be factual and verifiable
- Focus on a concrete medical fact
- Be clinically relevant for a general practitioner in practice

The question must NOT:

- Refer to 'the text', 'the study', or a study by name
- Be about study details such as research names, study duration, or sample sizes

The answer must:

- Be direct and factual
- Stay as close as possible to the exact wording of the source text

Text:

{chunk}

Example 1:

[question-answer pair about a single fact with reasoning steps...]

Example 2:

[question-answer pair about multiple facts with reasoning steps...]

Negative example:

[example of a question that must NOT be generated...]

// Full prompt available in prompts folder of codebase

Return your final answer as:

```
{
  "qa_pairs": [
    {"question": "...", "answer": "..."},
    ...
  ]
}
```

References

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

April 22, 2026.

- [2] Ikram Belmadani, Oumaima El Khettari, Pacôme Constant dit Beaufils, Richard Dufour, and Benoit Favre. Who judges the judge? evaluating llm-as-a-judge for french medical open-ended qa, 2026.
- [3] 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.
- [4] De Digitale Arts. Over de ai tool - de digitale arts. <https://dedigitalearts.nl/about-tool>, 2026. Accessed: April 22, 2026.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [6] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2025.
- [7] Juan José González Torres, Mihai Bogdan Bindilă, Sebastiaan Hofstee, Daniel Szondy, Quang-Hung Nguyen, Shenghui Wang, and Gwenn Englebienne. Automated question-answer generation for evaluating RAG-based chatbots. In Dina Demner-Fushman, Sophia Ananiadou, Paul Thompson, and Brian Ondov, editors, *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 204–214, Torino, Italia, May 2024. ELRA and ICCL.
- [8] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in python, 2020.
- [9] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020.
- [10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [11] Neil Mallinar, A. Ali Heydari, Xin Liu, Anthony Z. Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, Javier L. Prieto, Daniel McDuff, and Ahmed A. Metwally. A scalable framework for evaluating health language models, 2026.
- [12] NHG. Ai in de huisartsenpraktijk. <https://www.nhg.org/thema/digitale-zorg/kunstmatige-intelligentie-in-de-huisartsenzorg/ai-in-de-huisartsenpraktijk/>, 2026. Accessed: April 22, 2026.

- [13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [14] Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuseok Lim. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation, 2025.
- [15] 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.
- [16] Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation, 2024.
- [17] 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.
- [18] Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Pauline Lucas, H el ene Sauz eon, and Pierre-Yves Oudeyer. Selecting better samples from pre-trained llms: A case study on question generation, 2022.
- [19] 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.
- [20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.