# Generating and Evaluating an Automated Dutch Clinical QA Benchmark Grounded in the NHG Guidelines

**Anne-Sophie Straathof**
TU Delft

Supervisors: Jie Yang, Yannick ter Heerdt

June 22, 2026

## 1. Introduction & Background

- Large Language Models (LLMs) increasingly support clinical decision-making, but ungrounded hallucinations can compromise patient safety.
- **Retrieval Augmented Generation (RAG)** grounds answers in trusted documents to reduce this risk.
- These systems require automated benchmarks to validate them, but existing benchmarks focus mostly on English-language guidelines.
- **NHG Guidelines** represent the official standard of care in Dutch general practice, however, no QA benchmark exists for them.
- **Key Feature Questions (KFQs)** test clinical reasoning, not factual lookup, and are used to build this benchmark. Below shows an example of a KFQ.

**Question:** *Mr. Jansen, 68, presents with a blood glucose of 18 mmol/L. He's been vomiting for four hours and feels short of breath. History of type 2 diabetes and hypertension, on metformin and lisinopril. On exam he's clearly dehydrated and appears confused. What action do you take now?*
**Answer:** *You call an ambulance with immediate dispatch (highest urgency level) due to suspected ketoacidosis.*

## 2. Research Questions

**RQ1: Synthetic QA Generation**
*How can we automatically and reliably generate clinical benchmarks from raw guideline text?*

- Due to the unavailability of open Dutch exam materials, a reproducible pipeline was built to generate KFQs from NHG text, testing diverse prompting strategies.
- **Output:** a Dutch clinical QA benchmark of 375 traceable pairs created with the optimal prompting strategy

**RQ2: Automated vs. Expert Evaluation**
*Can out-of-the-box automated RAG metrics reliably replace human expert clinical judgment?*

- Frameworks **RAGAS** and **RAGChecker** are trusted for RAG safety evaluation, but untested against real clinicians in non-English domains.
- **Output:** Correlated automated scores against a licensed Dutch GP's annotations, which exposed critical alignment and safety gaps.

## 3. Synthetic QA Dataset Generation

### Methodology

- **Dataset:** NHG Guidelines of 10 high-prevalence Dutch primary care conditions (e.g., Asthma, Dementia, Diabetes, Depression, COPD).
- **Experimental Conditions:** 7 different prompting configurations: zero-shot, few-shot, chain-of-thought, self-refinement and 3 interleaved hybrids.
- **Evaluation metrics:** (Natural Language Inference) NLI faithfulness (*mDeBERTa-v3*), semantic coverage (*BERTScore*), LLM-as-judge ratings.

### Results

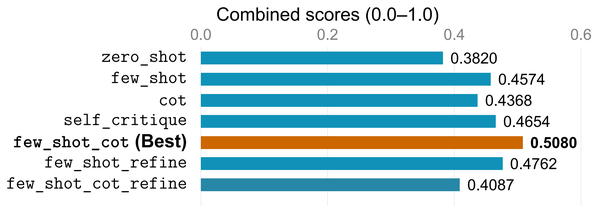


Figure 1. Automated evaluation scores across prompting strategies.

- **Optimal Configuration:** `few_shot_cot` achieved the highest overall performance and strongest local factual grounding.
- Self-refinement loops backfire due to semantic drift.
- No statistically significant differences ($p > 0.10$) across metrics, thus gains appear stylistic, not measurable.

**Limitations:** Strategy comparison used a single LLM (`gpt-4o-mini`) on 30 chunks, with no statistically significant differences across strategies ($p > 0.10$) `few_shot_cot` is the top numeric performer, not a proven-optimal one.

### Qualitative Analysis

To illustrate the underperformance of strategies, qualitative analysis revealed distinct failure modes:

Table 1. Shortened examples of QA generation failures

| Strategy | Guideline Source Text | Generated AI Error | Failure Mode |
|---|---|---|---|
| `zero_shot` | "Screen for diabetic nephropathy/retinopathy" | Paired with an acute foot-ulcer vignette | **Clinical Mismatch:** Acute emergency misaligned with chronic-care text. |
| `cot` | "Uncertain whether flu vaccination helps asthma (low-quality evidence)" | Concludes: "Recommend offering the flu vaccine." | **Hedge Misread:** Explicit clinical uncertainty converted into firm advice. |
| `few_shot_cot_refine` | "Call ambulance, urgent home visit" | Adds: sublingual nitroglycerin dose | **Unsafe Fabrication:** Contraindicated drug introduced, absent from source. |

## 4. Automating the Benchmark

### Methodology

- **Dataset:** Sampled $N = 30$ clinical QA pairs from the NHG Asthma guideline subset.
- A licensed Dutch doctor scored outputs on a 1–5 Likert scale across Context Adequacy, Faithfulness, and Clinical Safety.
- Evaluated identical pipelines zero-shot using **RAGAS** and **RAGChecker**.

### Results

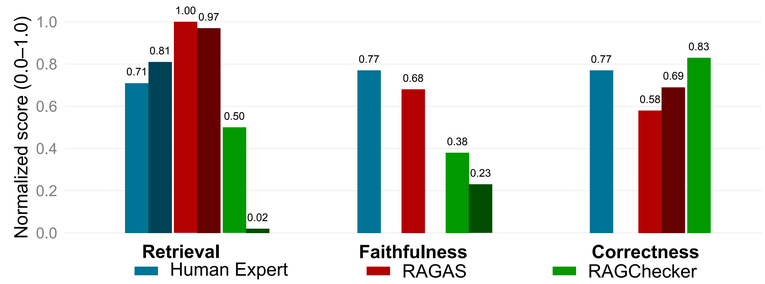


Figure 2. Visual comparison of cross-metric evaluation framework sub-scores.

- **RAGAS Overestimation:** Scored context recall at $1.00$ vs. expert adequacy at $0.71$, over-rewarding vocabulary overlap on clinically deficient text.
- **RAGChecker Over-Penalization:** Underestimated generator faithfulness ($0.38$ vs. expert $0.77$) due to rigid token-level claim splitting.
- **Domain Blindness:** Metrics awarded high scores to text containing fatal medical errors (e.g., misclassifying mild vs. severe protocols).

**Limitations:** Single expert annotator, single guideline (Asthma), $N = 30$, thus findings may not generalize across conditions or raters.

## 5. Conclusions & Future Work

- **Optimal Strategy:** Hybrid few-shot chain-of-thought performed best for synthetic KFQ generation.
- **Dataset:** Built a Dutch clinical QA benchmark of 375 traceable pairs across 10 primary care conditions.
- **Evaluation Gap:** Standard RAG toolkits (RAGAS, RAGChecker) align poorly with expert judgment as they miss real medical safety issues.

### Future Work

- Multi-expert setup for a stronger human baseline (inter-rater reliability).
- Extend the pipeline to the remaining NHG guidelines.
- Build custom Dutch clinical wrappers for RAG evaluation toolkits.