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

**Anne-Sophie Straathof**[1]
**Supervisors: Jie Yang**[1]**, Yannick ter Heerdt**[1]
[1]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: Anne-Sophie Straathof
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Yannick ter Heerdt, Pradeep Murukannaiah

**Abstract**

This thesis investigates the use of Large Language Models (LLMs) to automatically generate and evaluate synthetic clinical question-answer benchmarks based on Dutch NHG guidelines. The goal is to build a reliable and reproducible Key Feature Question (KFQ) dataset for testing clinical reasoning. In the first phase, different prompting strategies were tested using `gpt-4o-mini` across a subset of guideline text. The results show that while baseline model extraction is highly stable, a hybrid few shot chain of thought strategy performs best, achieving the highest optimization score and strong factual grounding. With this prompting a strategy a final benchmark dataset of 375 fully traceable Dutch QA pairs was constructed.

In the second phase, the feasibility of automating the benchmark evaluation was tested by comparing out-of-the-box frameworks RAGAS and RAGChecker directly against the grading of a licensed general practitioner. A significant judgement gap was found between the automated tools and human expert judgment. RAGAS systematically overestimated safety because it relies on literal word overlap, making it completely miss dangerous clinical errors like recommending a treatment that was explicitly stated to be failing. RAGChecker heavily penalized safe clinical paraphrasing and conditional reasoning due to its rigid token-level claim parsing. Ultimately, this work provides a functional pipeline for creating Dutch medical benchmarks, but highlights that standard automated evaluation toolkits require custom, domain-specific calibration before they can reliably replace human expert judgment.

# 1 Introduction

Large language models (LLMs) have demonstrated strong performance across a wide range of natural language processing tasks, including question answering, summarization, and reasoning [4]. Recently, there has been increasing interest in applying these models to the medical domain, where they may support clinical decision-making, education, and knowledge retrieval [37, 33]. However, evaluating LLMs in healthcare settings remains challenging due to the need for factual correctness, clinical grounding and regional correctness, as otherwise patient safety could be compromised [1, 32].

To mitigate the risks of clinical hallucinations, Retrieval Augmented Generation (RAG) is utilized to ground model responses in trusted reference documents [2].

This approach increases factual consistency and has shown promising results for clinical decision support [41, 16]. In this thesis, a good RAG system is one whose responses are faithful to retrieved evidence, clinically safe, and free of fabricated content. A poor system instead suffers from retrieval failures, context fabrication, or clinically unsafe recommendations despite plausible-sounding text.

Evaluating the RAG quality requires a benchmark of question-answer pairs structured around context relevance, faithfulness, and answer relevance. A good benchmark, in turn, is one with high discriminative power and strong alignment with expert clinical judgment. Within medical education, Key Feature Questions (KFQ) are established to assess clinical reasoning at critical decision points rather than simple factual recall [12]. This format uses a realistic patient vignette, a focused clinical question, and a short grounded answer linked to a specific source text span, making it ideal for testing active model reasoning. An example of such a KFQ can be seen below [24].

**Question:** *A 78-year-old woman presents to the office for an urgent appointment. She is complaining of a sudden onset of blurred and decreased vision in her right eye with distortion. She says that there is no redness or pain in the eye. She has not had any trauma. She has hypertension that is under control but denies any other health conditions. What is the most likely diagnosis in this case?*
**Answer:** *Age-related macular degeneration*

Despite the progress in medical question answering datasets, existing benchmarks focus primarily on English language guidelines or general medical knowledge. No benchmark currently evaluates clinical question answering over the Dutch guidelines (NHG guidelines [25]), which represent the official standard of care in Dutch primary healthcare. As a result, it remains unknown how reliably LLMs perform on these localized guidelines.

To address these limitations, this thesis investigates the use of LLMs to construct a fully synthetic clinical question answering benchmark based on the Dutch guidelines. Instead of relying on existing examination material that is not publicly available, such as the LHK toets or iVTG [13, 14], this work adopts a synthetic approach by generating KFQ directly from guideline text. A controlled dataset is constructed through a generation pipeline where each guideline chunk is processed under multiple prompting conditions, including zero shot prompting, few shot learning, chain of thought reasoning, self critique refinement, and hybrid combinations.

The main objective of this approach directly addresses both the generation and evaluation gaps in current medical processing. First, the pipeline establishes a clinical benchmark tailored to Dutch primary care standards, thereby overcoming the English language bias. Second, the framework investigates whether automated evaluation tools can reliably approximate expert judgment by directly comparing the metrics of RAGAS [9] and RAGChecker [30] against human expert annotations to determine their suitability for safety critical healthcare domains.

Across 119 generated pairs, results show that prompting strategies influence clinical coherence and grounding without reaching statistical significance, with hybrid few shot chain of thought approaches performing best, while iterative refinement introduces semantic drift. Furthermore, automated evaluation metrics demonstrate inconsistent correlations with clinical expert assessments, highlighting clear limitations when applying uncalibrated tools to safety critical medical domains.

To summarize, this thesis contributes an evaluation identifying an interleaved few shot chain of thought strategy as optimal prompting strategy for synthetic QA pairs, the first Dutch clinical question answering benchmark containing 375 traceable pairs and a human validation exposing severe alignment gaps where automated tools misclassify medical reasoning. The complete evaluation pipeline, datasets, and prompting frameworks developed in this study are publicly available to support reproducibility at `https://github.com/Nuffs/RP_NHG_2026`.

The remainder of this thesis is structured as follows: Section 2 reviews related work, Section 3 outlines dataset creation and prompting experiments, Section 4 details the automated pipeline and expert human evaluation, Section 5 addresses ethical aspects and reproducibility, Section 6 provides a comprehensive discussion and Section 7 concludes with future research directions.

# 2 Background and Related Work

This section provides the conceptual background for this research. It introduces clinical questions, outlines the architecture of RAG, reviews medical QA dataset construction methods with a focus on guideline-based synthetic generation, and discusses approaches for evaluating both benchmark quality and QA system performance.

## 2.1 Clinical Questions

Clinical questions form the basis of evidence-based clinical decision-making [31]. In evidence-based medicine, they are structured into focused, answerable queries to support systematic evidence retrieval [35]. One widely adopted framework is PICO (Patient/Problem, Intervention, Comparison, Outcome), which helps structure clinical uncertainty into searchable questions [6, 38].

In medical education, clinical questions assess reasoning and decision-making skills. While traditional multiple-choice questions (MCQs) are efficient, they often emphasize factual recall or pattern recognition rather than higher-order clinical reasoning [27, 28]. To address this limitation, KFQs evaluate decision-making at critical points in patient management by presenting realistic clinical scenarios and focusing on essential diagnostic or therapeutic decisions [12, 24].

In the Dutch context, national medical assessments, such as the Interuniversitaire Voortgangstoets Geneeskunde (iVTG) [14] and the Landelijke Huisartsgeneeskundige Kennistoets (LHK) [13], rely on case-based MCQs. While these exams do not explicitly label questions as PICO or KFQ, their case structure reflects similar principles of contextualized clinical decision-making. The benchmark design in this study explicitly mirrors this structure by shifting away from simple factual recall. By leveraging GPT-4o to generate multi-stage clinical vignettes based on the NHG guidelines, this methodology operationalizes procedural KFQ principles within an automated Dutch framework.

## 2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) combines information retrieval with language generation by providing a LLM with relevant passages from an external document collection prior to generating an answer [20]. Instead of relying solely on parametric knowledge, a RAG system grounds its responses in explicit evidence, which significantly reduces hallucinations [41].

RAG has become a prominent approach in medical QA as it allows systems to utilize authoritative clinical resources, such as treatment guidelines, without requiring model retraining when those resources are updated. Consequently, recent studies show that RAG-based systems generally produce more reliable clinical answers than parametric-only models when grounded in appropriate medical sources [41, 2].

## 2.3 Existing Medical QA Benchmarks

Many existing medical QA benchmarks, such as MedQA [15], MedMCQA [26] and MedExQA [18] derive their questions from official exams and trusted textbooks, providing high-quality, expert-validated benchmarks.

Alternatively, projects like RealMedQA [17] employ domain experts to manually write clinical questions. While highly accurate, manual generation is costly, time-consuming, and often constrained by the proprietary nature of national exam materials.

To address these limitations, recent work leverages LLMs for synthetic question generation. For example, RealMedQA [17] utilizes LLMs to generate realistic clinical vignettes, demonstrating that synthetic cases can approximate expert-written questions, while PubMedQA [15] and BioASQ [36] automatically derive questions from biomedical literature. Clinical guidelines provide an ideal source for this synthetic generation because they represent the evidence-based standards clinicians must follow in practice. For instance, Ding et al. [7] construct a silver-standard benchmark from publicly available NICE guidelines [23] using GPT, proving that LLM-generated, guideline-grounded benchmarks can support the systematic evaluation of clinical reasoning. However, a notable gap remains because these frameworks are predominantly developed for English-language ecosystems. This work directly addresses this limitation by developing a synthetic generation pipeline engineered specifically for the Dutch NHG guidelines.

## 2.4 Prompt Engineering

Because the quality of synthetically generated questions depends heavily on the instructions provided to the model, prompt design has become a critical consideration in benchmark construction. A comprehensive meta-analysis by Long et al. [8] demonstrates that careful prompt formulation substantially influences the quality, depth, and clinical consistency of generated outputs. Common approaches include zero-shot and few-shot prompting, which establish that models can execute diverse tasks using only direct instructions or a handful of contextual examples [3].

For intricate clinical scenarios, chain-of-thought prompting improves complex reasoning by prompting the model to generate explicit, intermediate logical steps before delivering a final answer [40]. Additionally, self-refinement methods introduce an iterative feedback loop where a model critiques and revises its own generated outputs, enhancing the accuracy of the synthetic benchmark without requiring additional computational training [22].

## 2.5 Evaluating Synthetically Generated QA Pairs

Before a synthetic benchmark can be used, the quality of the generated QA pairs must be verified to eliminate factual errors and clinical hallucinations [34, 39]. Traditional lexical metrics like ROUGE [21] and F1 [11] are inadequate here as they measure surface-level token overlap rather than clinical accuracy, often penalizing valid synonyms or overlooking critical medical contradictions [39]. Even embedding-based approaches like BERTScore [42] correlate poorly with human safety judgments in clinical settings [34].

Natural Language Inference (NLI) addresses these shortcomings by modeling whether a generated statement is entailed by, contradicts, or is unrelated to the source document [34]. This enables the direct detection of clinical contradictions, and multilingual cross-encoders like `mDeBERTa-v3-base-mnli-xnli` [19] allow for native NLI evaluation in Dutch without translation.

However, because NLI primarily targets local factual inconsistencies, recent frameworks combine it with LLM-as-a-Judge approaches to holistically evaluate broader clinical criteria like coherence and safety [34, 39].

### 2.6 Automatic Benchmark Evaluation of QA Systems

Once a benchmark has been constructed and validated, it can be used to evaluate the performance of QA systems. In retrieval-augmented systems, evaluation focuses not only on answer correctness but also on whether answers are grounded in retrieved evidence.

RAGAS [9] provides a collection of metrics including faithfulness, answer relevance, and context precision. These metrics assess whether generated answers are supported by retrieved documents and have become widely used for evaluating retrieval-augmented generation systems. RAGChecker [30] extends this approach by decomposing generated answers into individual claims and evaluating each claim separately. This claim-level analysis is particularly relevant in medical applications, where a single incorrect statement may have severe clinical consequences. RAGChecker also distinguishes between retrieval failures and generation failures, providing detailed diagnostic information about system behavior.

Despite advances in automated metrics, human evaluation remains an important component of medical QA assessment. Benchmarks such as BioASQ [36], MedQA [15], and RealMedQA [17] incorporate expert review to assess factual correctness, clinical appropriateness, and safety. Human evaluation is particularly valuable for open-ended clinical questions where guideline adherence and contextual reasoning cannot be fully captured by automated methods. This project bridges this gap by directly correlating automated metrics from RAGAS and RAGChecker against human expert evaluations on the custom Dutch dataset.

## 3 QA Dataset Generation

This section outlines the framework used to construct and evaluate a synthetic clinical Question-Answering benchmark derived from Dutch NHG guidelines [25]. To evaluate LLMs on primary care reasoning rather than rote factual recall, the dataset utilizes the Key Feature Question format, presenting a realistic patient vignette followed by a single critical clinical decision [12]. Because existing Dutch examination materials (e.g., LHK-toets [13], iVTG [14]) were not available and manual expert construction was unfeasible due to resource constraints, a fully synthetic pipeline was developed to generate QA pairs directly from guideline text.

### 3.1 Methodology

#### 3.1.1 Data Source

The dataset is constructed from ten high-prevalence conditions in Dutch primary care, selected based on national disease burden data from the Volksgezondheid Toekomstverkenning [29]: Asthma (adults), dementia, depression, COPD, anxiety disorders, chronic kidney disease, diabetes mellitus, influenza, heart failure, and hand/wrist complaints. This selection ensures clinical relevance and represents a broad spectrum of everyday general practice workloads, requiring diverse clinical reasoning modes.

#### 3.1.2 Prompting Strategies

To isolate the effects of instruction styles on KFQ generation, four isolated baselines and three hybrid configurations were tested using `gpt-4o-mini`:

- **Zero-Shot (`zero_shot`):** Baseline providing role context and formatting constraints without examples.
- **Few-Shot (`few_shot`):** Adds a single verified, high-quality KFQ exemplar.
- **Chain-of-Thought (`cot`):** Forces the model to document its clinical logic in a hidden reasoning step.
- **Self-Critique (`self_critique`):** A two-turn pipeline where a draft is reviewed and refined by an auxiliary prompt.
- **Hybrids:** Interleaved configurations combining these strategies (`few_shot_cot`, `few_shot_refine`, `few_shot_cot_refine`).

Figure 2 in Appendix B provides a visual overview of how these strategies feed into the generation and evaluation pipeline described in the remainder of this section.

#### 3.1.3 Experimental Design

Following a repeated-measures design, each text segment from the selected guidelines was processed under all prompting conditions using gpt-4o-mini. This method ensures that variations in text complexity, vocabulary density, or clinical topic across different guidelines are completely controlled for, allowing the analysis to isolate the true mathematical effect of the prompting strategy itself. A power analysis conducted in G*Power 3.1 [10] assumed a medium effect size ($f = 0.25$, $\alpha = 0.05$, power $= 0.80$, $\rho = 0.50$), indicating a minimum requirement of 24 guideline chunks. Consequently, 30 chunks were randomly sampled from the guideline database to serve as the benchmarking material.

#### 3.1.4 Evaluation Framework

To assess the generated QA pairs across multiple dimensions of lexical similarity, semantic alignment, and retrieval feasibility, the pipeline implements a comprehensive, multi-layered hybrid evaluation framework consisting of the following components:

1. **Factual Grounding (Sentence-Level NLI):** Local semantic consistency is assessed via a sentence-level NLI paradigm using a cross-encoder model (`mDeBERTa-v3-base-mnli-xnli`) [19]. Every sentence of the generated answer (`gt_answer`) is evaluated against each sentence of the source guideline chunk text. Max-pooling over the resulting matrix derives directional *Faithfulness*, *Contradiction* (used to quantify direct clinical hallucination risk), and *Neutral* probabilities directly in the Dutch language.
2. **Semantic Textual Overlap (BERTScore):** Traditional lexical overlap metrics like ROUGE are replaced with *BERTScore Recall* [42]. This metric computes token-level contextual embedding similarities to evaluate whether the generated answer accurately retains the core clinical concepts of the complete reference guideline chunk, bypassing the phrasing limitations of exact token matching.
3. **Clinical Quality (LLM-as-a-Judge):** To capture holistic attributes missed by token and sentence alignment models, an automated LLM judge (`gpt-4o-mini`) evaluates each pair using a strict, continuous 0–5 scale across five explicit clinical dimensions:

- *Correctness:* Is the answer medically accurate and logically derived from the premise?
- *Guideline Adherence:* Does the clinical decision strictly reflect the explicit recommendations of the NHG text?
- *Reasoning Quality:* Does the vignette present an authentic clinical puzzle rather than a trivial look-up task?
- *Safety:* Does the question avoid misleading diagnostic paths that could prove dangerous if deployed in practice?
- *Non-hallucination:* Is the clinical context free of external facts not grounded in the provided source text?

To balance these diverse metrics into a single optimization signal, the pipeline executes a calibrated composite objective function defined as:

$$S_{\text{combined}} = 0.40 \cdot F - 0.25 \cdot H_{\text{penalty}} + 0.25 \cdot \left(\frac{J}{5.0}\right) + 0.10 \cdot B \tag{1}$$

where $F$ is the NLI Faithfulness score, $J$ is the average LLM-Judge rating, $B$ is the BERTScore Recall, and $H_{\text{penalty}}$ represents a non-linear hallucination penalty.

The weights of this objective function were assigned based on design choices rather than empirical tuning. *Faithfulness* (0.40) receives the highest weight because preventing ungrounded clinical claims is paramount in medical safety frameworks. The *LLM-Judge rating* (0.25, normalized to a 0–1 scale) and the *Hallucination penalty* ($-0.25$) act as equal and opposite forces, punishing clinical drift while rewarding structural and logical coherence. Finally, *BERTScore* (0.10) receives the smallest weight because while semantic coverage of the guideline chunk is beneficial, a high-quality KFQ must naturally narrow its focus to a single critical key feature rather than replicating the entire guideline segment.

## 3.2 Results

The automated pipeline achieved high stability across all configurations, with minor structural compilation failures occurring only within the complex hybrid setups. Table 1 summarizes the performance metrics.

The `few_shot_cot` strategy performed best overall, leading in combined score (0.5080) and factual grounding (Faithfulness = 0.5548). For isolated baselines, providing an explicit structural format (`few_shot` or `self_critique`) was more effective at maintaining text-grounding than abstract step-by-step reasoning (`cot`). Multi-stage refinement architectures revealed a clear trade-off: while `few_shot_cot_refine` achieved high validity (29/30), its hallucination rate spiked to 0.0664, showing that sequential rewriting loops introduce semantic drift.

To determine if these variations were statistically robust, repeated-measures ANOVAs were conducted for both settings. No differences reached statistical significance at $\alpha = 0.05$ across any metric (all $p > 0.10$). This shows that `gpt-4o-mini`'s core clinical extraction performance is stable, meaning prompting modifications mostly affect surface-level style and formatting.

Table 1: Automated evaluation scores across prompting strategies (compiled over 30 source chunks). * indicates no statistically significant differences ($p > 0.05$) across any metric via repeated-measures ANOVA.

| Strategy | Valid | BERT ↑ | Faith. ↑ | Halluc. ↓ | LLM-J ↑ | Combined ↑ |
|---|---|---|---|---|---|---|
| *Isolated Baselines** | | | | | | |
| `zero_shot` | 30/30 | 0.5855 | 0.2519 | **0.0013** | 4.46 | 0.3820 |
| `few_shot` | 29/30 | 0.5935 | 0.4226 | **0.0013** | **4.59** | 0.4574 |
| `cot` | 30/30 | 0.5759 | 0.4100 | 0.0018 | 4.31 | 0.4368 |
| `self_critique` | 30/30 | 0.5757 | 0.4709 | 0.0061 | 4.42 | 0.4654 |
| *Interleaved Hybrids** | | | | | | |
| `few_shot_cot` | 28/30 | 0.5887 | **0.5548** | 0.0097 | **4.59** | **0.5080** |
| `few_shot_refine` | 26/30 | **0.5967** | 0.4754 | 0.0032 | 4.54 | 0.4762 |
| `few_shot_cot_refine` | 29/30 | 0.5941 | 0.4092 | 0.0664 | 4.32 | 0.4087 |

### 3.3 Qualitative Examples

The aggregate scores in Table 1 show *where* each strategy struggles; reading the underlying generations shows *why*. Three cases below correspond directly to the weakest results in that table: the low Faithfulness of `zero_shot` (0.2519), the underperformance of `cot` relative to `few_shot`-based strategies despite both using explicit reasoning, and the hallucination spike in `few_shot_cot_refine` (0.0664). Table 2 summarizes all three, translated from Dutch.

The `zero_shot` case illustrates keyword matching without medical logic: the model saw "diabetes" and built a vignette around a diabetic foot ulcer, but paired it with a `source_span` about routine eye and kidney screening instead of the relevant acute-care guidance. This mismatch is exactly what drags `zero_shot`'s mean Faithfulness down to 0.2519 in Table 1; here it produces a near-zero pair score (Faithfulness 0.02, LLM-Judge 1.0/5).

The `cot` case shows reasoning drifting from a hedge into an unwarranted recommendation: the source states it is "very uncertain" whether flu vaccination helps asthma patients, citing very low quality evidence, yet the generated answer recommends vaccination on this basis. This is consistent with `cot` trailing `few_shot` on Faithfulness (0.4100 vs. 0.4226) in Table 1: without a concrete exemplar to anchor it, the hidden reasoning step can resolve textual uncertainty in the wrong direction.

The `few_shot_cot_refine` case shows the refinement step itself introducing risk: for a patient with acute chest pressure and a history of low blood pressure, the `source_span` instructs only that an ambulance be called, but the refined answer adds a nitroglycerin dose "regardless of whether chest pain is present". This single pair (Hallucination 0.313, LLM-Judge 2.0/5) is representative of why this strategy's mean hallucination rate (0.0664) is above every other configuration in Table 1, despite the refinement prompt explicitly prohibiting clinical content changes.

Table 2: Three generated examples illustrating the failure modes behind the scores in Table 1, translated from Dutch and shortened for readability.

| **Strategy** | `zero_shot` | `cot` | `few_shot_cot_refine` |
|---|---|---|---|
| Guideline Domain | Diabetes (Foot Ulcer) | Asthma (Adults) | Heart Failure (Acute) |
| Vignette | 67-year-old man, painful, non-healing foot wound for two weeks; type 2 diabetes, HbA1c 8.5%, reduced sensation on monofilament testing. | 34-year-old man with worsening asthma; asks whether flu vaccination would help control his asthma. | 67-year-old man, sudden severe breathlessness and chest pressure; history of hypertension and *known low blood pressure*; saturation 88%. |
| Source Span | "...check for the presence of diabetic nephropathy and retinopathy." | "It is very uncertain whether seasonal flu vaccination reduces flu cases in asthma patients (very low quality evidence)." | "Have an ambulance called with the highest priority and arrange a home visit." |
| Generated Answer | Recommends routine eye-/kidney screening for a patient with an acute foot infection. | Recommends offering the flu vaccine, since it may reduce flu cases. | Adds: take sublingual nitroglycerin or isosorbide dinitrate, regardless of chest pain. |
| NLI Faithfulness | 0.02 | 0.69 | 0.0005 |
| NLI Hallucination | 0.00 | 0.003 | 0.313 |
| LLM Judge | 1.0 / 5.0 | 3.4 / 5.0 | 2.0 / 5.0 |
| Failure Mode | *Clinical Mismatch:* acute infection paired with a chronic-care answer. | *Hedge Misread:* explicit uncertainty converted into a confident recommendation. | *Unsafe Fabrication:* contraindicated drug and dose added, absent from source. |

### 3.4 Creating the Final Dataset

Based on the results of the highest factual grounding and combined scores, the `few_shot_cot` configuration was selected as the optimal prompting strategy to create the final benchmark dataset. By targeting three independent QA pairs per chunk across the full extracted guideline database a final dataset of 375 fully traceable clinical QA pairs was successfully generated. With the benchmark fully constructed, the focus can be shifted from dataset generation to automatic evaluation. This establishes the foundation for the subsequent section, where multiple automatic evaluation frameworks are compared with expert judgement.

## 4 Automating the Benchmark

While the previous section focused on generating the dataset, this section investigates whether evaluating the benchmark can also be automated. Manual expert annotation is the gold standard but scales poorly due to high costs and time requirements. Therefore, this section evaluates how well RAGAS [9] and RAGChecker [30] match human expert judgment on the constructed clinical QA benchmark.

### 4.1 Methodology

A verification dataset of $N = 30$ QA pairs was sampled from the NHG Asthma Guideline subset. A licensed Dutch general practitioner scored each pair across four dimensions on a

5-point Likert scale:

1. **Context Adequacy (A):** retrieved context contains the information needed to answer the case (1 = irrelevant, 5 = complete).
2. **Context Faithfulness (B):** model stays within the facts of the retrieved chunk (1 = severe external assumptions, 5 = fully faithful).
3. **Clinical Safety & Accuracy (C):** answer is medically accurate and safe per the NHG standard (1 = unsafe, 5 = fully correct).
4. **Context Relevancy (D):** retrieved context is concise and free of noise (1 = filled with noise, 5 = fully clean).

These four dimensions differ from the five LLM-Judge criteria introduced in Section 3 intentionally as the Section 3 criteria evaluate the generated QA pair in isolation. The Section 4 criteria instead evaluate a full RAG pipeline's output, where a system must first retrieve a relevant chunk and then generate an answer from it. Additionally the criteria here intentionally align with RAGAS and RAGChecker categories to allow for direct comparisons: A maps to Context Recall, B to Faithfulness, C to Factual Correctness.

The same 30 inputs were evaluated zero-shot with both frameworks. Summary statistics for the raw expert annotations (mean, standard deviation) are reported in Appendix D.2, the normalized scores are used for the comparison below.

## 4.2 Results

Table 3 aligns the normalized expert scores against the automated frameworks across three evaluation domains. The expert ratings show that Context Relevancy was judged highest (0.81) while Context Adequacy was the weakest dimension (0.71), indicating that the primary weakness in the pipeline lies in retrieval rather than generation. RAGAS scores context recall at 1.00 and precision at 0.97, while the expert rates the same chunks at only 0.71 adequacy, indicating RAGAS over-rewards token overlap even when a chunk lacks genuine clinical utility. In the Faithfulness domain, the expert and RAGAS are relatively close (0.77 vs. 0.68), but RAGChecker drops sharply to 0.38, suggesting it systematically penalizes valid clinical paraphrasing as hallucination. Figure 1 visualizes this comparison.

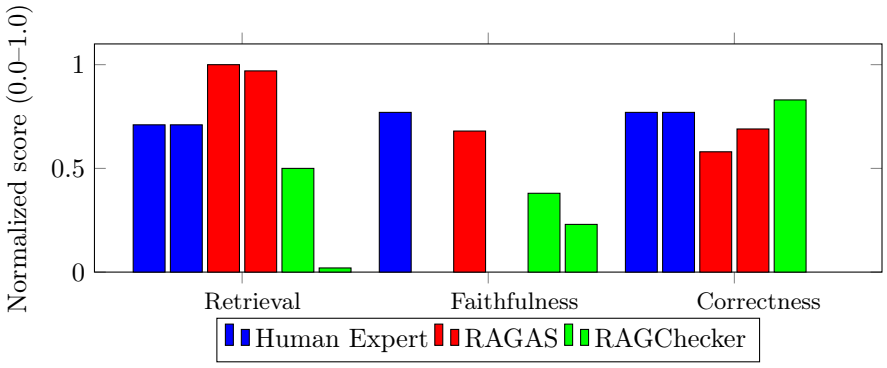


Figure 1: Visual comparison of the scores in Table 3.

Table 3: Cross-framework alignment: Human Expert vs. RAGAS vs. RAGChecker (normalized 0.0–1.0).

| Domain | Framework | Metric | Score |
|---|---|---|---|
| **Retrieval** | Human Expert | A: Context Adequacy | 0.71 |
| | Human Expert | D: Context Relevancy | 0.81 |
| | RAGAS | Context Recall | **1.00** |
| | RAGAS | Context Precision | 0.97 |
| | RAGChecker | Claim Recall | 0.50 |
| | RAGChecker | Context Precision | 0.02 |
| **Faithfulness** | Human Expert | B: Context Faithfulness | **0.77** |
| | RAGAS | Faithfulness | 0.68 |
| | RAGChecker | Generator Faithfulness | 0.38 |
| | RAGChecker | Hallucination Rate (↓) | 0.23 |
| **Correctness** | Human Expert | C: Clinical Safety | **0.77** |
| | RAGAS | Factual Correctness | 0.58 |
| | RAGAS | Answer Relevancy | 0.69 |
| | RAGChecker | Overall F1 | 0.83 |

## 4.3 Case-Level Analysis

Table 4 presents three cases where the automated scores differed most from experts scores. Table 5 shows the corresponding scores.

Table 4: Clinical vignettes, model responses, and expert critiques for three divergence cases (translated from Dutch & shortened for readability).

| ID | Vignette | Model Response | Expert Critique |
|---|---|---|---|
| `48e7` | Ms. Koster, 32, with a mild asthma attack, shows insufficient improvement after salbutamol inhalation. Which pharmacological option is now indicated? | When improvement after salbutamol inhalation is insufficient during a mild asthma attack, it is indicated to administer ipratropium via a spacer. | *"The answer is wrong, it describes the protocol for a severe attack. For a mild attack the correct step is to increase the SABA dose; systemic escalation is not indicated."* |
| `9b3c` | Mr. Jansen, 28, presents with episodic dyspnoea, expiratory wheezing, and nocturnal cough. Which diagnosis do you consider? | Based on the symptoms of episodic dyspnoea, expiratory wheezing, and regular nocturnal coughing, I consider the diagnosis of asthma. | *"Missing an entire section: diagnosis also requires spirometric reversibility ($\geq$12%, $\geq$200 ml) or bronchial hyperreactivity. The answer is symptom-only and clinically unsafe."* |
| `d572` | Ms. Van Dijk, 58, has had good asthma control for 3 months on step-4 medication. Is back-referral indicated? | Yes, back-referral is indicated, as good control has lasted 3 months on step-4 medication and there is no history of frequent or life-threatening attacks or extensive comorbidity. | *"This is being hallucinated: there is no history of frequent or life-threatening lung attacks or extensive comorbidity. Note, the question is asked very specifically with step-4 medication. This makes it a reflection of the guideline, and you would also expect a good answer."* |

Table 5: Item-level score divergence for the cases in Table 4. Expert scores as $x/5$ (normalized).

| Case ID | Metric | Expert | RAGAS | RAGChecker |
|---|---|---|---|---|
| 48e7cd51 | Context Adequacy / Recall | 1/5 (0.00) | 1.00 | 1.00 |
| | Clinical Safety / F1 | 1/5 (0.00) | 0.67 | 1.00 |
| | Faithfulness | 3/5 (0.50) | 0.50 | 1.00 |
| 9b3c0e6b | Context Adequacy / Recall | 2/5 (0.25) | 1.00 | 1.00 |
| | Clinical Safety / F1 | 1/5 (0.00) | 0.80 | 1.00 |
| | Faithfulness | 1/5 (0.00) | 0.50 | 1.00 |
| d5728205 | Faithfulness | 1/5 (0.00) | 0.25 | 0.71 |
| | Clinical Safety / F1 | 1/5 (0.00) | 0.00 | 0.83 |
| | Context Adequacy / Recall | 4/5 (0.75) | 1.00 | 1.00 |

Cases `48e7cd51` and `9b3c0e6b` confirm RAGAS over-estimation: both received near-perfect automated scores despite expert Clinical Safety ratings of 0.00, since the model response closely matched the retrieved chunk even though that chunk was clinically wrong (`48e7cd51`, severe- vs. mild-attack protocol) or incomplete (`9b3c0e6b`, missing spirometry criteria). Token overlap alone was sufficient for both RAGAS and RAGChecker to award maximum scores. Case `d5728205` shows the toolkits can also disagree with each other: the model hallucinated a qualifying condition absent from the source, which RAGAS correctly flagged (faithfulness = 0.25, factual correctness = 0.00, consistent with the expert), while RAGChecker scored it at faithfulness = 0.71 and F1 = 0.83, missing the error entirely.

## 4.4 Discussion

The results show a clear gap between simple text matching and actual clinical reasoning. RAGAS over-rewards word overlap, giving perfect scores to text chunks that the expert flagged as inadequate. RAGChecker underestimates model faithfulness because it breaks the dense Dutch text into rigid atomic claims [30], heavily penalizing the model for safe clinical paraphrasing. The toolkits even disagree with each other on cases like `d5728205`. This shows that out-of-the-box metrics fail to catch when an answer is medically inappropriate [5] because they measure token overlap rather than clinical logic.

For developers, the main takeaway is that a high automated score does not guarantee clinical safety. These toolkits do not work well for clinical questions requiring reasoning. They cannot tell if a retrieved guideline is technically on-topic but completely wrong for a patient's history. Automated metrics are useful for catching gross retrieval errors, but they should not be used as a standalone quality gate to certify that an AI's advice is safe.

This evaluation has two main limitations: it was restricted to $N = 30$ questions from a single guideline and relies on only one general practitioner. Future work should expand across multiple guidelines and doctors to calculate an Inter-Rater Reliability index, creating a stronger human baseline to properly calibrate automated RAG metrics.

# 5 Responsible Research

This study used LLMs to generate and evaluate synthetic clinical question-answer pairs derived from Dutch general practice guidelines. While the work does not involve real patient data, it operates in a sensitive domain and therefore requires careful reflection on ethical use, potential risks, and reproducibility of the computational pipeline.

## 5.1 Ethical Considerations

One main ethical issue is using official clinical guidelines (NHG guidelines [25]) as the basis for creating our artificial data. These guidelines are the standard of care in Dutch general practice and are written by medical experts. However, they are made specifically for the Dutch healthcare system and its rules. Because of this, the dataset reflects local Dutch medical standards and should not be treated as a universal medical truth that applies everywhere.

Additionally, these question-answer pairs are not meant to be used for treating patients, diagnosing illnesses or making real medical decisions. They are purely a research tool to test how language models behave under controlled conditions. Even so, because the questions look like realistic patient cases, there is a risk that someone might mistake them for genuine medical advice. To prevent this, all answers are forced to use direct quotes from the guidelines, making it easy to trace the information back to the source and reducing the chance of unsupported medical statements.

Another risk is using LLMs for medical reasoning. Even when given the correct source text, these models can still write believable but wrong reasoning, or slightly twist medical meanings. This is a big concern in healthcare because AI errors can easily look correct to a reader. In this study, this risk is mitigated by forcing the AI to stick strictly to the text in the guidelines and by only analyzing the data after the experiment is done rather than using it in any real-world medical setting.

## 5.2 Reproducibility

To ensure reproducibility, all code is version-controlled and can be run end-to-end. Additionally the codebase is publically available at `https://github.com/Nuffs/RP_NHG_2026`. This means anyone with access to the same NHG guidelines and model APIs can completely replicate the dataset and results.

The dataset generation process is kept mostly deterministic by using a fixed random seed to select the guideline chunks. At the same time, all prompting strategies, model versions, and generation settings are explicitly defined and logged. This setup allows the dataset to be regenerated under identical conditions. Additionally, each QA pair is stored with full metadata, including the source chunk ID, section path, model settings, and the exact prompt used.

The evaluation pipeline is standardized in the same way. The automated metrics are run using fixed settings without any custom tuning. For the human evaluation, the doctor used a structured rubric to keep the scoring consistent, though a small amount of personal judgment is always unavoidable.

# 6 Conclusion & Future Work

This thesis developed and evaluated an automated framework to generate and benchmark clinical KFQs based on Dutch NHG guidelines. In the first phase different prompting strategies have been tested with `gpt-4o-mini`. Even though the statistical differences between prompts were not significant, the hybrid few shot chain of thought strategy performed best, reaching the highest overall score and strong faithfulness. On the other hand, multi-stage rewriting loops few shot chain of thought self refinement strategy caused problems by introducing semantic drift and higher hallucination rates (0.0664). By using the few shot chain of thought pipeline 375 fully traceable QA pair were created for the final benchmark dataset.

The second phase revealed major limitations when using standard RAG evaluation toolkits for specialized, non-English medical text. A large gap was found between automated metrics and human expert judgment. RAGAS overestimated safety because it relies on simple keyword matching, making it blind to serious medical errors like recommending a failing treatment (SABA failure) to a worsening patient. Meanwhile, RAGChecker heavily penalized safe clinical paraphrasing and conditional logic because it breaks text down into strict, word-level claims. These findings show that standard evaluation toolkits cannot be used as perfect judges in medical settings without custom calibration.

Based on these findings, several areas for future research are identified. First, to fix the limitation of using only one expert, future validation should include a multi-expert setup with multiple independent general practitioners. This would allow for the calculation of an Inter-Rater Reliability index, creating a much stronger human baseline that accounts for differing medical opinions on complex cases.

Second, the automated frameworks need to be tested across the remaining nine high-prevalence conditions in the generated dataset, as this initial human evaluation was limited to the Asthma Guideline. Expanding the testing scope will help determine if the calibration gaps and word-matching biases found in this study happen across other medical topics as well.

Finally, future work should focus on customizing automated toolkits so they can better handle dense Dutch medical phrasing and synonyms. Instead of using these tools completely out-of-the-box, custom medical wrappers or fine-tuned models could be developed to bridge the gap between simple text similarity and real medical logic. Adjusting the internal weights of RAGAS and RAGChecker to reward actual clinical reasoning rather than raw word overlap will be essential to make automated metrics reliable for medical tools.

# References


[1] H. Ali, J. Qadir, T. Alam, M. Househ, and Z. Shah. Chatgpt and large language models in healthcare: Opportunities and risks. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, pages 1–4. IEEE, 2023.

[2] Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877, 2025.

[3] Tom B. Brown et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[4] Jingwen Cheng, Kshitish Ghate, Wenyue Hua, William Yang Wang, Hong Shen, and Fei Fang. REALM: A dataset of real world LLM use cases. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8331–8341, Vienna, Austria, July 2025. Association for Computational Linguistics.

[5] Mohita Chowdhury, Yajie Vera He, Jared Joselowitz, Aisling Higham, and Ernest Lim. ASTRID: An automated and scalable TRIaD for the evaluation of RAG-based clinical question answering systems. *arXiv preprint arXiv:2501.08208*, 2025.

[6] Dahlgren Memorial Library. Clinical questions, PICO, & study designs. `https://guides.dml.georgetown.edu/ebm/ebmclinicalquestions`, 2025. Accessed: 2026-05-04.

[7] Qing Ding, Eric H. Q. Zhang, Felix Jozsa, and Julia Ive. Building a silver-standard dataset from NICE guidelines for clinical LLMs. *arXiv preprint arXiv:2511.01053*, 2025.

[8] Xuan Long Do, Duy Dinh, Ngoc-Hai Nguyen, Kenji Kawaguchi, Nancy F. Chen, Shafiq Joty, and Min-Yen Kan. What makes a good natural language prompt? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pages 5835–5873. Association for Computational Linguistics, 2025.

[9] Patrick Es et al. RAGAS: Automated evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2309.17452*, 2023.

[10] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007.

[11] GeeksforGeeks. F1 Score in Machine Learning. `https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/`, 2024. Accessed: 2026-06-18.

[12] Patricia Hrynchak, Susan Glover Takahashi, and Marla Nayer. Key-feature questions for assessment of clinical reasoning: a literature review. *Medical Education*, 48(9):870–883, 2014.

[13] Huisartsopleiding Nederland. LHK-toets. `https://www.huisartsopleiding.nl/toetsen-beoordelen/lhk-toets`, 2025. Accessed: 2026-05-04.

[14] iVTG Consortium. Interuniversitaire voortgangstoets geneeskunde. `https://ivtg.nl/nl/`, 2025. Accessed: 2025.

[15] Di Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. *Transactions of the Association for Computational Linguistics*, 9:453–467, 2021.

[16] Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, Shao-Chun Wu, Vesela P. Kovacheva, and Daniel Shu Wei Ting. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187, 2025.

[17] Georgina Kell, Angus Roberts, Sharon Umansky, Yash Khare, et al. RealMedQA: A pilot biomedical question answering dataset containing realistic clinical questions. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2025.

[18] Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. MedExQA: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2311.09799*, 2023.

[19] Moritz Laurer. mDeBERTa-v3-base-mnli-xnli. Hugging Face Model Card, 2023.

[20] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

[21] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[22] Aman Madaan, Niket Tandon, Peter Clark, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

[23] National Institute for Health and Care Excellence. NICE Guidance. `https://www.nice.org.uk/guidance`, 2026. Accessed: 2026-06-18.

[24] Marla Nayer, Susan Glover Takahashi, and Patricia Hrynchak. Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning. *Medical Teacher*, 40(11):1116–1122, 2018.

[25] Nederlands Huisartsen Genootschap. NHG-richtlijnen. `https://richtlijnen.nhg.org/`, 2026. Accessed: 2026-04-25.

[26] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Proceedings of Machine Learning for Health (ML4H)*, 2022.

[27] Elizabeth J. Palmer and Paul G. Devitt. Pattern recognition as a concept for multiple-choice questions in a national licensing exam. *BMC Medical Education*, 14:232, 2014.

[28] Nithin Kumar Raman, Taylor Lundy, and Kevin Leyton-Brown. Reasoning models are test exploiters: Rethinking multiple-choice. *arXiv preprint arXiv:2507.15337*, 2025.

[29] Rijksinstituut voor Volksgezondheid en Milieu (RIVM). Volksgezondheid toekomstverkenning 2024: Trendscenario ziekten en aandoeningen. `https://www.volksgezondheidtoekomstverkenning.nl/vtv-2024/trendscenario/ziekten-aandoeningen`, 2024. Accessed: 2026-06-02.

[30] Danyang Ru, Lin Qiu, Xiachong Hu, Tianyi Zhang, Peng Shi, Shiyu Chang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 21999–22027, 2024.

[31] David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[32] K. Singhal, S. Azizi, T. Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[33] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, et al. Toward expert level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[34] Arjun Subramanian, Viktor Schlegel, Abhinav Ramesh Kashyap, Thien-Tuan Nguyen, Vibhor Prakash Dwivedi, and Stefan Winkler. M-QALM: A benchmark to assess clinical reading comprehension and knowledge recall in large language models via question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4002–4042. Association for Computational Linguistics, 2024.

[35] SUNY Downstate Health Sciences University. What is a clinical question? `https://guides.downstate.edu/c.php?g=868154&p=6230102`, 2025. Accessed: 2026-05-04.

[36] George Tsatsaronis et al. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 2015.

[37] Alyssa Unell, Mehr Kashyap, Michael Pfeffer, and Nigam Shah. Real world usage patterns of large language models in healthcare. *medRxiv*, 2025. Preprint, posted May 6, 2025.

[38] Cor van Loveren and Irene H. A. Aartman. De PICO-vraag. *Nederlands Tijdschrift voor Tandheelkunde*, 114(4):172–178, 2007. Accessed: 2026-05-04.

[39] Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Mitian Xu, Yidong Wang, Xiachong Hu, Zuozhuo Zhang, and Yu Zhang. Evaluating open-QA evaluation. *Advances in Neural Information Processing Systems*, 36, 2023.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[41] Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Rita Lee, Joanna Melia, Joanna Nelson, Karim Sallam, Melissa Ann Tullis, Stacey Vogelsong, John Patrick Cunningham, and William Hiesinger. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068, 2024.

[42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

# A Statement on the Use of AI

For this thesis, AI tools (such as Gemini, ChatGPT and Claude) were utilized solely to optimize grammar, phrasing and formatting to help improvide clarity and readability. AI tools were not used for idea generation, analytical reasoning or heavy writing tasks.

# B Generation and Evaluation Pipeline Prompting Experiment

Figure 2 illustrates the full pipeline described in Section 3: a guideline chunk is processed by one of seven prompting strategies, split into single-call configurations (`zero_shot`, `few_shot`, `cot`, `few_shot_cot`) and two-call refine configurations (`self_critique`, `few_shot_refine`, `few_shot_cot_refine`), producing a candidate QA pair that is scored by the three-component evaluation framework and collapsed into the combined optimization score.

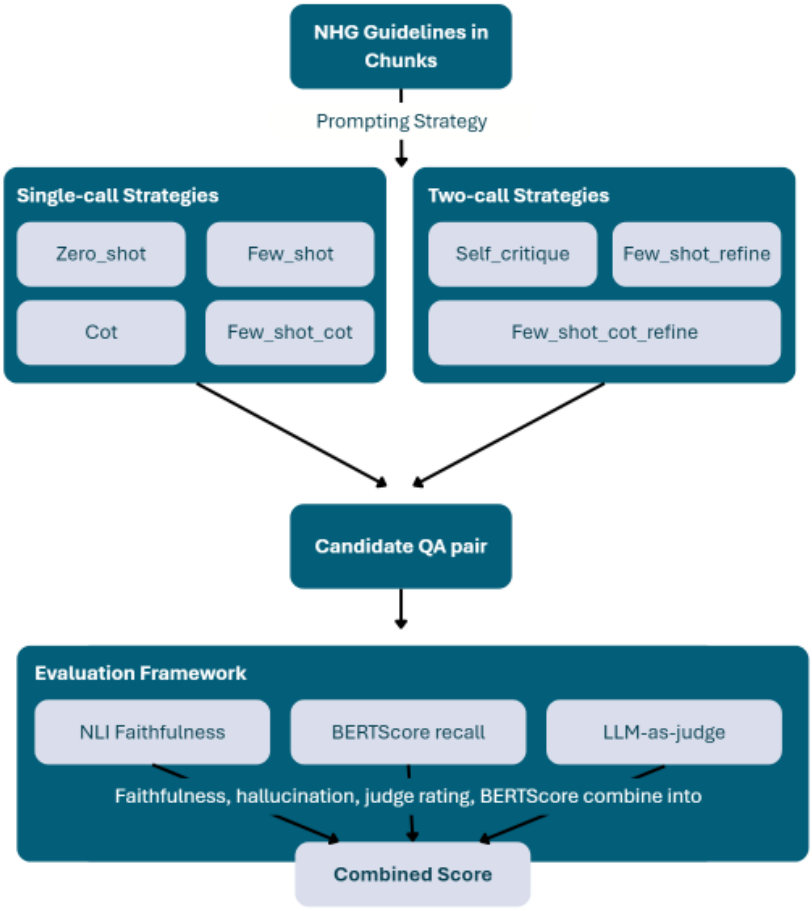


Figure 2: Overview of the QA generation and evaluation pipeline.

# C Prompts Used in Prompting Experiments

This appendix provides the exact system instructions, context formatting, and templates utilized across both the isolated and hybrid prompting experiments described in Section 3. Both the original Dutch instructions provided to the model and their English translations are included for structural clarity.

## C.1 Experiment 1: Isolated Prompting Baselines

### C.1.1 Base System Instruction Template (Zero-Shot Baseline)

The following base instruction prompt establishes the role context, definitions of a Key Feature Question (KFQ), structural guidelines, and rigid output constraints. This baseline text serves as the foundation for all configurations.

Listing 1: Original Dutch Baseline System Prompt

```
Je bent een ervaren Nederlandse huisarts die toetsvragen schrijft
    voor huisartsen in opleiding.

Gegeven een fragment uit een NHG-richtlijn, schrijf je een
    realistische klinische casus in het format van een Key Feature
    Question (KFQ).

## Wat is een Key Feature Question?
Een KFQ beschrijft een concrete patientsituatie en stelt een
    kritische klinische vraag over een beslismoment. De vraag test of
    de arts de juiste klinische redenering toepast, niet of hij de
    richtlijn kan opzoeken.

## Structuur van de casus
Beschrijf een concrete patient met:
- Naam, leeftijd, geslacht
- Aanleiding voor het consult
- Relevante klachten, duur en beloop
- Voorgeschiedenis, medicatie, leefstijl, allergieen waar relevant
- Bevindingen uit anamnese of lichamelijk onderzoek waar relevant
- Eventuele voorkeuren of zorgen van de patient

Sluit af met een concrete klinische vraag over:
- Diagnose ("Wat is de meest waarschijnlijke diagnose?")
- Beleid ("Welk beleid is het meest aangewezen?")
- Behandeling ("Welk medicament heeft de voorkeur?")
- Verwijzing ("Is verwijzing geindiceerd, en zo ja, met welke
    urgentie?")
- Follow-up ("Wanneer en hoe volgt u deze patient op?")

## Regels
1. De casus is REALISTISCH en SPECIFIEK - gebruik een echte naam,
    leeftijd en concrete details.
2. Verwerk bewust klinisch relevante details die de beslissing
    beinvloeden (bijv. contra-indicaties, comorbiditeit, leefstijl).
```

```
3. Het antwoord moet volledig onderbouwd kunnen worden vanuit de
    gegeven NHG-tekst.
4. Vermijd vragen die beginnen met "Wat zegt de richtlijn over..." of
    "Noem de criteria voor...".
5. De casus moet de arts dwingen tot redeneren, niet tot het opzoeken
    van een definitie.
6. Schrijf in natuurlijk Nederlands, zoals een ervaren huisarts zou
    communiceren.
7. Verzin GEEN feiten die niet in de tekst staan.
8. source_span is een LETTERLIJK CITAAT - kopieer het woord voor
    woord uit de tekst.
9. Controleer: plak source_span terug in de tekst. Als je het niet
    kunt vinden, kies een andere zin.
10. Kies ALLEEN volledige zinnen die eindigen op een punt. Geen
    opsommingspunten, geen headers, geen zinsdelen.
11. Als de tekst alleen opsommingen bevat zonder volledige zinnen,
    gebruik dan de inleidende zin die de opsomming introduceert.

## Regels antwoord
1. Het antwoord (gt_answer) bestaat uit maximaal 2-3 zinnen.
2. Geef alleen de klinische conclusie en de directe onderbouwing
    vanuit de tekst.
3. Geen opsommingen, geen uitgebreide uitleg.

Geef je uitvoer ALLEEN als JSON-array, zonder markdown of uitleg:
[
  {
    "query": "Volledige casusbeschrijving gevolgd door de klinische
        vraag.",
    "gt_answer": "Het antwoord op de vraag.",
    "source_span": "Verbatim fragment uit de tekst dat het antwoord
        onderbouwt.",
    "retrieval_query": "Korte klinische zoekterm (max 15 woorden) die
        direct aansluit bij de richtlijntekst, bijv: 'spirometrie
        obstructie astma COPD aanvullend onderzoek'."
  }
]
```

Listing 2: English Translation of Baseline System Prompt

```
You are an experienced Dutch general practitioner writing test
    questions for general practitioners in training.

Given an excerpt from an NHG guideline, write a realistic clinical
    case in the format of a Key Feature Question (KFQ).

## What is a Key Feature Question?
A KFQ describes a concrete patient situation and asks a critical
    clinical question about a decision point. The question tests
    whether the physician applies the correct clinical reasoning, not
    whether they can look up the guideline.

## Case Structure
```

```
Describe a concrete patient with:
- Name, age, gender
- Reason for the consultation
- Relevant complaints, duration, and course
- Medical history, medication, lifestyle, allergies where relevant
- Findings from anamnesis or physical examination where relevant
- Any preferences or concerns of the patient

Conclude with a concrete clinical question regarding:
- Diagnosis ("What is the most likely diagnosis?")
- Management ("Which management approach is most appropriate?")
- Treatment ("Which medication is preferred?")
- Referral ("Is a referral indicated, and if so, with what urgency?")
- Follow-up ("When and how do you follow up on this patient?")

## Rules
1. The case must be REALISTIC and SPECIFIC - use a real name, age,
    and concrete details.
2. Deliberately incorporate clinically relevant details that
    influence the decision (e.g., contraindications, comorbidity,
    lifestyle).
3. The answer must be fully supportable from the provided NHG text.
4. Avoid questions starting with "What does the guideline say
    about..." or "Name the criteria for...".
5. The case must force the doctor to reason, not to look up a
    definition.
6. Write in natural Dutch, as an experienced GP would communicate.
7. Do NOT invent facts that are not in the text.
8. source_span is a VERBATIM CITATION - copy it word for word from
    the text.
9. Verification: paste source_span back into the text. If you cannot
    find it, choose another sentence.
10. Select ONLY complete sentences ending in a period. No bullet
    points, no headers, no phrase fragments.
11. If the text only contains lists without full sentences, use the
    introductory sentence that introduces the list.

## Answer Rules
1. The answer (gt_answer) consists of a maximum of 2-3 sentences.
2. Provide only the clinical conclusion and the direct substantiation
    from the text.
3. No bullet points, no extensive explanations.

Provide your output ONLY as a JSON array, without markdown or
    explanation:
[
  {
    "query": "Full case description followed by the clinical
        question.",
    "gt_answer": "The answer to the question.",
    "source_span": "Verbatim fragment from the text supporting the
        answer.",
```

```
    "retrieval_query": "Short clinical search term (max 15 words)
       directly aligned with the guideline text, e.g., 'spirometrie
       obstructie astma COPD aanvullend onderzoek'."
  }
]
```

### C.1.2 Few-Shot Prompt

For the `few_shot` baseline strategy, the following in-context example block is directly appended to the end of the base template.

Listing 3: Original Dutch Few-Shot In-Context Block

```
Hieronder volgt een voorbeeld van een hoogwaardige KFQ zoals bedoeld:

Tekst: "Spirometrie is de aangewezen methode om obstructie aan te
    tonen of uit te sluiten. Voer spirometrie uit bij patiënten met
    klachten die passen bij astma of COPD."

Gewenste output:
[
  {
    "query": "Mevrouw De Vries, 52 jaar, komt op uw spreekuur met al
       drie maanden aanhoudende hoestklachten en kortademigheid bij
       inspanning. Ze rookt 10 jaar, een half pakje per dag. Haar
       longauscultatie is normaal. U overweegt astma of COPD. Welk
       aanvullend onderzoek is als eerste aangewezen om obstructie
       aan te tonen of uit te sluiten?",
    "gt_answer": "Spirometrie is aangewezen om obstructie aan te
       tonen of uit te sluiten bij klachten die passen bij astma of
       COPD. Dit is de geëigende methode volgens de richtlijn.",
    "source_span": "Spirometrie is de aangewezen methode om
       obstructie aan te tonen of uit te sluiten.",
    "retrieval_query": "spirometrie obstructie aantonen uitsluiten
       astma COPD aanvullend onderzoek"
  }
]

Genereer nu een vergelijkbare KFQ op basis van de onderstaande tekst.
```

Listing 4: English Translation of Few-Shot In-Context Block

```
Below is an example of a high-quality KFQ as intended:

Text: "Spirometry is the preferred method to demonstrate or exclude
    obstruction. Perform spirometry in patients with symptoms
    consistent with asthma or COPD."

Expected output:
[
  {
```

```
    "query": "Mrs. De Vries, 52 years old, comes to your consultation
       room with persistent cough symptoms for three months and
       shortness of breath upon exertion. She has smoked for 10
       years, half a pack a day. Her lung auscultation is normal.
       You are considering asthma or COPD. Which supplementary
       examination is primarily indicated to demonstrate or exclude
       obstruction?",
    "gt_answer": "Spirometry is indicated to demonstrate or exclude
       obstruction in patients with symptoms consistent with asthma
       or COPD. This is the appropriate method according to the
       guideline.",
    "source_span": "Spirometry is the preferred method to demonstrate
       or exclude obstruction.",
    "retrieval_query": "spirometry obstruction demonstrate exclude
       asthma COPD supplementary examination"
  }
]

Now generate a comparable KFQ based on the text below.
```

### C.1.3 Chain-of-Thought (CoT) Instructions

The `cot` layout enforces intermediate reasoning steps by appending these specialized constraints and sequence targets onto the baseline template.

Listing 5: Original Dutch CoT Appended Instructions

```
## INTERNE REDENEERSTAPPEN (niet tonen):
1. Zoek het beslismoment
2. Selecteer relevante klinische factoren
3. Kies relevante zin uit de tekst (letterlijk)
4. Bouw KFQ rond deze exacte zin

## OUTPUT REGELS (strikt):
- Alleen JSON output
- source_span = exact citaat uit brontekst
- retrieval_query = korte klinische zoektermen die direct aansluiten
    bij de richtlijntekst
- geen uitleg, geen tussenstappen
```

Listing 6: English Translation of CoT Instructions

```
## INTERNAL REASONING STEPS (do not show):
1. Locate the clinical decision point
2. Select relevant clinical factors
3. Choose the relevant sentence from the text (verbatim)
4. Construct the KFQ around this exact sentence

## OUTPUT RULES (strict):
- JSON output only
- source_span = exact quote from the source text
- retrieval_query = short clinical search terms directly connecting
    to the guideline text
```

```
- no explanation, no intermediate steps
```

### C.1.4 Self Critique Prompt

The `self_critique` workflow operates via a two-stage process. The model first builds an unrefined question using the base configuration, which is then fed into a secondary LLM routing block running this exact refinement prompt.

Listing 7: Original Dutch Critique Refinement Instruction

```
Je beoordeelt een KFQ.

## BRONTEKST:
"""[Ingevoerde NHG-richtlijntekst]"""

## KFQ:
[Geregenereerd concept-ontwerp JSON]

## CRITICAL RULE:
Je mag GEEN inhoud uit de bron herschrijven.

Je mag alleen:
- formulering verbeteren
- structuur verbeteren
- maar source_span MOET EXACT ONGEWIJZIGD blijven uit de bron

## OUTPUT:
Return exact JSON:
[
  {
    "query": "...",
    "gt_answer": "...",
    "source_span": "EXACTE TEKST UIT BRON (niet wijzigen!)",
    "retrieval_query": "..."
  }
]
```

Listing 8: English Translation of Critique Refinement Instruction

```
You are evaluating a KFQ.

## SOURCE TEXT:
"""[Inputted NHG-guideline text]"""

## KFQ:
[Generated draft concept JSON]

## CRITICAL RULE:
You are NOT allowed to rewrite any content from the source text.

You are only permitted to:
- improve formulation
```

```
- improve structure
- but source_span MUST REMAIN EXACTLY UNCHANGED from the source

## OUTPUT:
Return exact JSON:
[
  {
    "query": "...",
    "gt_answer": "...",
    "source_span": "EXACT TEXT FROM SOURCE (do not change!)",
    "retrieval_query": "..."
  }
]
```

## C.2 Experiment 2: Combined Hybrid Prompting Strategies

Experiment 2 evaluates the compounding effects and trade-offs of combining the isolated prompt components from Experiment 1. Programmatically, these strategies avoid repetition by concatenating the structural string variables defined in Appendix C.1.

### C.2.1 Few-Shot & Chain-of-Thought Hybrid (`few_shot_cot`)

The `few_shot_cot` strategy merges structural demonstrations with latent reasoning directives. It is constructed by directly appending the reasoning rules to the few-shot template:

```
SYSTEM_PROMPT_FEW_SHOT_COT = SYSTEM_PROMPT_BASE + FEW_SHOT_EXAMPLE +
SYSTEM_PROMPT_COT_RULES
```

The exact combined system instruction presented to the model during execution is:

Listing 9: Original Combined Execution Block for few_shot_cot

```
[Insert full SYSTEM_PROMPT_BASE text here]

Hieronder volgt een voorbeeld van een hoogwaardige KFQ zoals bedoeld:

Tekst: "Spirometrie is de aangewezen methode om obstructie aan te
    tonen of uit te sluiten. Voer spirometrie uit bij patiënten met
    klachten die passen bij astma of COPD."

Gewenste output:
[
  {
    "query": "Mevrouw De Vries, 52 jaar, komt op uw spreekuur met al
        drie maanden aanhoudende hoestklachten en kortademigheid bij
        inspanning. Ze rookt 10 jaar, een half pakje per dag. Haar
        longauscultatie is normaal. U overweegt astma of COPD. Welk
        aanvullend onderzoek is als eerste aangewezen om obstructie
        aan te tonen of uit te sluiten?",
    "gt_answer": "Spirometrie is aangewezen om obstructie aan te
        tonen of uit te sluiten bij klachten die passen bij astma of
        COPD. Dit is de geëigende methode volgens de richtlijn.",
```

```
    "source_span": "Spirometrie is de aangewezen methode om
        obstructie aan te tonen of uit te sluiten.",
    "retrieval_query": "spirometrie obstructie aantonen uitsluiten
        astma COPD aanvullend onderzoek"
  }
]

Genereer nu een vergelijkbare KFQ op basis van de onderstaande tekst.

## INTERNE REDENEERSTAPPEN (niet tonen):
1. Zoek het beslismoment
2. Selecteer relevante klinische factoren
3. Kies relevante zin uit de tekst (letterlijk)
4. Bouw KFQ rond deze exacte zin

## OUTPUT REGELS (strikt):
- Alleen JSON output
- source_span = exact citaat uit brontekst
- retrieval_query = korte klinische zoektermen die direct aansluiten
    bij de richtlijntekst
- geen uitleg, geen tussenstappen
```

Listing 10: English Translation of the Combined few_shot_cot Strategy

```
[Insert full Translated SYSTEM_PROMPT_BASE text here]

Below is an example of a high-quality KFQ as intended:

Text: "Spirometry is the preferred method to demonstrate or exclude
    obstruction. Perform spirometry in patients with symptoms
    consistent with asthma or COPD."

Expected output:
[
  {
    "query": "Mrs. De Vries, 52 years old, comes to your consultation
        room with persistent cough symptoms for three months and
        shortness of breath upon exertion. She has smoked for 10
        years, half a pack a day. Her lung auscultation is normal.
        You are considering asthma or COPD. Which supplementary
        examination is primarily indicated to demonstrate or exclude
        obstruction?",
    "gt_answer": "Spirometry is indicated to demonstrate or exclude
        obstruction in patients with symptoms consistent with asthma
        or COPD. This is the appropriate method according to the
        guideline.",
    "source_span": "Spirometry is the preferred method to demonstrate
        or exclude obstruction.",
    "retrieval_query": "spirometry obstruction demonstrate exclude
        asthma COPD supplementary examination"
  }
]
```

```
Now generate a comparable KFQ based on the text below.

## INTERNAL REASONING STEPS (do not show):
1. Locate the clinical decision point
2. Select relevant clinical factors
3. Choose the relevant sentence from the text (verbatim)
4. Construct the KFQ around this exact sentence

## OUTPUT RULES (strict):
- JSON output only
- source_span = exact quote from the source text
- retrieval_query = short clinical search terms directly connecting
    to the guideline text
- no explanation, no intermediate steps
```

#### C.2.2 Few-Shot & Recursive Refinement Loop (`few_shot_refine`)

This strategy utilizes a multi-stage execution pipeline. It does not use a single combined prompt, but rather passes data between two separate model calls sequentially:

1. **Generation Stage:** The model generates an initial draft question using `SYSTEM_PROMPT_FEW_SHOT` (the combination of the base prompt and the dynamic example block).
2. **Refinement Stage:** The resulting draft JSON is captured by the pipeline script and passed as an input variable into a secondary execution call running the independent `critique_and_refine` prompt block detailed in Appendix C.1.4.

#### C.2.3 Few-Shot & Chain-of-Thought & Recursive Refinement Loop (`few_shot_cot_refine`)

This composite strategy maximizes directive constraints across a dual-stage pipeline by combining the hybrid generation prompt with the post-hoc validation check:

1. **Generation Stage:** The model creates an initial draft utilizing the comprehensive, concatenated `SYSTEM_PROMPT_FEW_SHOT_COT` framework.
2. **Refinement Stage:** The intermediate reasoning and draft output are processed by the execution script and passed directly into the automated critique block running the secondary `critique_and_refine` template to enforce final structural schema safety.

# D Expert Evaluation

This appendix shows all information concerning the expert evaluation as described in Chapter 4. First the exact survey layout can be found and second the results.

## D.1 Survey

### D.1.1 Information Sheet

Listing 11: Original Dutch Information Sheet

```
Geachte arts,

Hartelijk dank dat u de tijd neemt om deel te nemen aan dit
    validatie-onderzoek. Dit onderzoek vormt een onderdeel van mijn
    Bachelor Thesis aan de Technische  Universiteit Delft (TU Delft).

Doel van het onderzoek
Binnen de huisartsenpraktijk is snelle en accurate toegang tot
    medische richtlijnen cruciaal. In dit project hebben we een
    AI-assistent ontwikkeld die op basis van natuurlijke taalvragen
    direct specifieke passages uit de NHG-Standaarden kan opzoeken en
    samenvatten. Dit type systeem maakt gebruik van
    Retrieval-Augmented Generation (RAG): een techniek waarbij een
    AI-model eerst
de juiste medische brontekst zoekt en deze vervolgens gebruikt om een
    betrouwbaar antwoord te formuleren.

In deze fase van het onderzoek testen we het systeem specifiek op de
    NHG-Standaard Astma bij volwassenen. Het doel is om te valideren
    of de automatische metrics (statistische scores) die gebruikt
    worden om de AI te meten, daadwerkelijk overeenkomen met het
    kritische oordeel van een medisch expert.

Tijdens het onderzoek
U krijgt zometeen een aantal concrete patiëntcasussen met vragen te
    zien. Bij elke casus worden drie onderdelen getoond:
1. De Patiëntencasus / Vraag: Een klinische vignette met een vraag.
2. Gegenereerd Antwoord: Het medische advies dat de AI heeft
    opgesteld op basis van die specifieke brontekst.
3. De Opgehaalde Brontekst (Context): De specifieke alinea's die de
    database automatisch heeft geselecteerd uit de NHG-Richtlijnen

Per casus worden een aantal korte vragen gesteld om deze te
    beoordelen op basis van:
- De vindkracht (Retrieval): Heeft het systeem de juiste medische
    paragraaf erbij gepakt.
- De efficiëntie & ruis (Precision): Zijn de getoonde fragmenten
    compact en to-the-point of bevatten ze te veel overbodige
    randinformatie die de bruikbaarheid hindert?
- De getrouwheid (Faithfulness): Blijft de AI strikt bij de feiten
    uit de getoonde tekst, of introduceert het model externe claims,
    eigen aannames of hallucinaties?
```

```
- De klinische veiligheid: Is het eindantwoord medisch accuraat,
    correct en direct veilig toepasbaar in de praktijk volgens de
    huidige NHG-Richtlijn?

In totaal zijn er 30 casussen, echter kan u tussentijds stoppen door
    de 'stoppen' optie aan te vinken. Uw ingevulde informatie zal dan
    worden opgeslagen.

Vertrouwelijkheid en Data
Uw deelname is volledig vrijwillig en uw antwoorden worden strikt
    vertrouwelijk en geanonimiseerd verwerkt binnen de beveiligde
    onderzoeksomgeving van de TU Delft. De resultaten worden
    uitsluitend gebruikt voor academische doeleinden om AI-systemen
    in de zorg veiliger en betrouwbaarder te maken. U kunt ten alle
    tijden stoppen tijdens met het beoordelen van de vragen, klik op
    de 'stoppen' optie om uw antwoorden op te slaan en te stoppen
    voor het einde.

Bij vragen of opmerkingen kunt u contact opnemen met:
Anne-Sophie Straathof
```

Listing 12: Translated Information Sheet

```
Dear Physician,

Thank you very much for taking the time to participate in this
    validation study. This study forms part of my Bachelor's Thesis
    at the Delft University of Technology (TU Delft).

Purpose of the Study
Within general practice, rapid and accurate access to medical
    guidelines is crucial. In this project, we have developed an AI
    assistant capable of instantly retrieving and summarizing
    specific passages from the NHG Standards based on natural
    language queries. This type of system utilizes
    Retrieval-Augmented Generation (RAG): a technique where an AI
    model first retrieves the correct medical source text and
    subsequently uses it to formulate a reliable response.

In this phase of the research, we are testing the system specifically
    on the NHG Standard for Asthma in Adults. The goal is to validate
    whether the automated metrics (statistical scores) used to
    measure the AI's performance actually align with the critical
    judgment of a medical expert.

During the Evaluation
You will shortly be presented with a number of concrete patient cases
    and corresponding questions. For each case, three elements will
    be shown:
1. The Patient Case / Question: A clinical vignette with a question.
2. Generated Answer: The medical advice drafted by the AI based on
    that specific source text.
```

```
3. The Retrieved Source Text (Context): The specific paragraphs
    automatically selected by the database from the NHG Guidelines.

For each case, a few short questions will be asked to evaluate it
    based on:
- Retrieval Quality (Retrieval): Did the system retrieve the correct
    medical section?
- Efficiency & Noise (Precision): Are the displayed fragments compact
    and to the point, or do they contain too much superfluous or
    tangential information that hinders usability?
- Faithfulness (Faithfulness): Does the AI adhere strictly to the
    facts in the displayed text, or does the model introduce external
    claims, its own assumptions, or hallucinations?
- Clinical Safety: Is the final response medically accurate, correct,
    and directly safe for clinical application according to the
    current NHG Guidelines?

There are 30 cases in total; however, you can stop at any time by
    checking the 'stop' option. Your completed data will then be
    saved.

Confidentiality and Data
Your participation is entirely voluntary, and your answers will be
    processed with strict confidentiality and anonymized within the
    secure research environment of TU Delft. The results will be used
    exclusively for academic purposes to make AI systems in
    healthcare safer and more reliable. You may stop evaluating the
    questions at any moment; simply click the 'stop' option to save
    your answers and exit before the end.

For questions or comments, you may contact:
Anne-Sophie Straathof
```

### D.1.2 Question Format

For each QA pair the following format was used for each question.

Listing 13: Original Dutch Survey Format

```
PATIENTENCASUS
...

GEGENEREERD ANTWOORD
...

OPGEHAALDE BRONTEKST (De Context)
...

De volledige NHG-richtlijn voor Astma is hier te lezen: Astma bij volwassenen |
    NHG-Richtlijnen
```

```
1. Beoordeel de opgehaalde tekstfragmenten (Context). Bevat deze context de juiste
    informatie uit de NHG-Standaard om de casus te beantwoorden?
   1      2      3      4      5
  [Nee]                        [Ja]

2. Blijft de AI in het gegenereerde antwoord strikt binnen de feiten van de
    getoonde tekstfragmenten, zonder zelf medische zaken aan te vullen of te
    veranderen?
   1      2      3      4      5
  [Nee]                        [Ja]

3. Is het eindantwoord medisch accuraat, correct en veilig om toe te passen volgens
    de NHG Standaard Astma bij volwassenen?
   1      2      3      4      5
  [Nee]                        [Ja]

4. In welke mate bevatten de opgehaalde tekstfragmenten (Context) overbodige of
    irrelevante informatie die niet nodig is om de specifieke casus te beantwoorden?
   1      2      3      4      5
  [Nee]                        [Ja]

5. Indien u bij de voorgaande vragen een lage score hebt geselecteerd, wat
    ontbreekt er of wat gaat er specifiek mis in de tekst of het antwoord? (Bijv.
    ontbrekende dosering, verkeerde terminologie, onveilig advies)
```

Listing 14: Translated Information Sheet

```
PATIENT CASE
...

GENERATED ANSWER
...

RETRIEVED SOURCE TEXT (The Context)
...

The full NHG guideline for Asthma can be read here: Asthma in adults
    | NHG Guidelines

1. Evaluate the retrieved text fragments (Context). Does this context
    contain the correct information from the NHG Standard to answer
    the case?
   1      2      3      4      5
  [No]                         [Yes]

2. Does the AI in the generated answer stay strictly within the facts
    of the displayed text fragments, without supplementing or
    changing medical matters themselves?
   1      2      3      4      5
  [No]                         [Yes]
```

```
3. Is the final answer medically accurate, correct, and safe to apply
    according to the NHG Standard for Asthma in adults?
   1      2      3      4      5
  [No]                         [Yes]

4. To what extent do the retrieved text fragments (Context) contain
    superfluous or irrelevant information that is not necessary to
    answer the specific case?
   1      2      3      4      5
  [No]                         [Yes]

5. If you selected a low score for the previous questions, what is
    missing or what specifically goes wrong in the text or the
    answer? (E.g., missing dosage, incorrect terminology, unsafe
    advice)
```

## D.2 Summary statistics for clinical expert annotations

Table 6: Summary statistics for clinical expert annotations ($N = 30$) on a 1–5 Likert scale.

| **Dimension** | **Mean ($\mu$)** | **Std. Dev. ($\sigma$)** | **Normalized** |
|---|---|---|---|
| A: Context Adequacy | 3.53 | 1.36 | 0.71 |
| B: Context Faithfulness | 3.83 | 1.56 | 0.77 |
| C: Clinical Safety & Accuracy | 3.87 | 1.25 | 0.77 |
| D: Context Relevancy | 4.07 | 1.23 | 0.81 |