Benchmarking Open-Source vs. Closed-Source LLMs for Dutch Medical Guidelines

A Quantitative Evaluation of Retrieval-Augmented Generation using the NHG-Guidelines

Abdelrahman Tageldin[1]
Supervisor(s): 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: Abdelrahman Tageldin
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Yannick Ter Heerdt, Pradeep Murukannaiah

Abstract

Large Language Models (LLMs) integrated with Retrieval-Augmented Generation (RAG) offer promising clinical decision support capabilities. However, utilizing closed-source models for this purpose requires transmitting sensitive patient data to external servers, creating severe GDPR compliance and privacy risks. Conversely, open-source models can be securely hosted locally, but their clinical reasoning capabilities in non-English medical contexts remain unproven. This research quantitatively benchmarks the performance of three closed-source and three open-source LLMs operating over the Dutch NHG-guidelines. Using a standardized RAG pipeline and an automated "LLM-as-a-Judge" evaluation framework (RAGChecker), we analyze the exact trade-offs between clinical accuracy, computational cost, and inference speed. The results reveal a significant paradigm shift: top-tier open-source models (specifically DeepSeek V4 Pro) not only match but outperform closed-source models (GPT-5.5) in clinical accuracy and speed at a fraction of the cost, offering a highly viable, privacy-preserving alternative for Dutch healthcare institutions.

1 Introduction

Large Language Models (LLMs) offer highly promising capabilities for clinical decision support, but their reliance on static internal memory poses severe risks. Medical knowledge is highly context-dependent and constantly evolving; therefore, hallucinations or outdated advice can have direct, harmful consequences for patient safety. To mitigate these risks, Retrieval-Augmented Generation (RAG) grounds model responses in trusted, external medical documents before generating an answer, significantly improving factual accuracy and reliability [13, 1].

In the Netherlands, general practitioners rely on the NHG-guidelines as the absolute standard for primary care. While implementing RAG systems over these guidelines presents a massive opportunity, selecting the appropriate foundational LLM creates a critical tension between clinical intelligence and data privacy. Currently, closed-source models like GPT-5.5 and Claude Opus 4.7 demonstrate state-of-the-art clinical reasoning [8]. However, deploying these models requires transmitting highly sensitive patient data to external servers, introducing severe privacy and GDPR compliance risks, alongside high recurring API costs. Conversely, top-tier open-source models (such as Kimi K2.6 or DeepSeek V4) offer a compelling alternative. Because they can be hosted locally on a hospital's own secure servers, they effectively eliminate external data privacy risks.

Despite the clear privacy advantages of open-source models, a critical knowledge gap remains: it is currently unknown whether these open-source models possess the necessary clinical reasoning and Dutch language comprehension to safely replace closed-source models in a primary care setting. Therefore, the main research question of this work is: *How do open-source language models compare to closed-source models on automated NHG factual and clinical benchmarks?* We break this down into two specific sub-questions: (1) How does performance differ between simple factual retrieval versus complex clinical reasoning? and (2) What is the exact trade-off between a model's computational efficiency (speed and cost) and its clinical accuracy?

To contextualize these sub-questions, it is important to distinguish between the two types of evaluations. The factual benchmark tests direct knowledge retrieval using simple, localized queries (e.g., identifying a specific medication dosage). In contrast, the clinical benchmark

evaluates complex reasoning through realistic patient vignettes, requiring the model to synthesize multiple symptoms and patient history to determine a treatment plan. Prior to the experiments, our baseline expectation was that while open-source models might perform adequately on simple factual retrieval, they would likely struggle with the nuanced, multi-step logic required by the clinical vignettes compared to heavily trained proprietary models like GPT-5.5.

To answer these questions, we evaluate six state-of-the-art models (three closed-source and three open-source) using an end-to-end RAG pipeline. Relying on validated factual and clinical benchmark datasets derived from the NHG guidelines, we employ an automated "LLM-as-a-Judge" methodology via the RAGChecker framework [7] to score the models on metrics such as faithfulness and answer correctness. Our main contribution is a comprehensive, quantitative comparison that reveals the exact trade-offs between accuracy, latency, and cost, providing Dutch healthcare institutions with data-driven guidance on secure AI deployment.

The remainder of this paper is structured as follows. Section 2 discusses related work regarding medical RAG evaluation and the privacy versus performance debate. Section 3 details the methodology, including the decoupled two-phase RAG architecture and the automated evaluation paradigm. Section 4 presents the experimental setup, quantitative results, and performance trade-offs. Section 5 reflects on responsible research and ethical considerations. Finally, Section 6 concludes the paper with a discussion on the broader implications for clinical AI deployment.

2 Related Work

2.1 RAG in Healthcare

The integration of Large Language Models (LLMs) into clinical decision support systems offers transformative capabilities for parsing unstructured medical data, yet it introduces profound risks regarding factual hallucinations that can compromise patient safety. To mitigate this, Retrieval-Augmented Generation (RAG) frameworks have emerged as a fundamental architectural requirement, decoupling the generative reasoning engine from parametric memory by dynamically grounding outputs in verified external corpora [11, 6]. In the context of regional healthcare applications, such as querying Dutch medical guidelines, RAG restricts the model's inferential space to localized, evidence-based consensus rather than generalized pre-training distributions. Advanced implementations further reduce hallucination rates by explicitly cross-referencing generated clinical claims against retrieved medical texts [12]. Consequently, RAG acts as a critical safeguard, ensuring that AI-driven diagnostic and treatment recommendations remain strictly tethered to authoritative medical knowledge.

2.2 Evaluating RAG Systems

As clinical RAG architectures grow in complexity, traditional Natural Language Processing (NLP) metrics have proven fundamentally inadequate for assessing the nuanced interplay between document retrieval and long-form medical generation. This limitation has catalyzed a methodological shift toward “LLM-as-a-Judge” frameworks, which leverage the semantic reasoning of advanced frontier models to automate reference-free evaluation. The Retrieval

Augmented Generation Assessment (RAGAS) framework pioneered this approach by isolating system performance into dimensions such as context precision, answer relevancy, and generation faithfulness, minimizing reliance on expensive human annotations [5]. More recently, frameworks like RAGChecker [7] have introduced fine-grained, claim-level entailment diagnostics. By decomposing generated medical responses and ground-truth references into atomic claims, RAGChecker provides actionable, localized feedback on specific retriever failures or generator hallucinations. These automated evaluation pipelines are indispensable for benchmarking distinct model families over complex medical corpora.

2.3 The Privacy vs. Performance Debate

The deployment of these RAG pipelines necessitates navigating a critical trade-off between the out-of-the-box reasoning capabilities of proprietary models and the stringent data sovereignty requirements mandated by the General Data Protection Regulation (GDPR) [10]. While closed-source API models exhibit exceptional schema fidelity and zero-shot reasoning, transmitting Protected Health Information (PHI) to third-party cloud infrastructure introduces severe regulatory vulnerabilities and black-box opacity. Consequently, there is an accelerated transition toward locally hosted open-source and open-weight models. Recent evaluations of clinical information extraction over Dutch medical reports demonstrate that high-parameter open models can achieve competitive, cloud-grade performance in native-language processing when properly integrated into a local RAG pipeline [4]. By executing on-premises, these open-source architectures guarantee zero data leakage, fulfilling stringent European privacy laws while maintaining the transparency required for auditable and equitable clinical AI.

2.4 Synthesis: Informing Our Research Design

The aforementioned literature directly dictates the architectural and evaluative design of this study. First, recognizing the limitations of traditional NLP metrics in complex medical contexts [5], our experimental setup explicitly adopts the fine-grained, "LLM-as-a-Judge" methodology formalized by RAGChecker [7] to evaluate generator faithfulness and noise sensitivity. Second, the fundamental tension between cloud-based reasoning capabilities [8] and stringent European data sovereignty constraints [10] informs our exact model selection. We deliberately benchmark flagship closed-source APIs against leading open-weights models (e.g., DeepSeek V4 Pro, Kimi K2.6) to empirically quantify whether the privacy guarantees of local deployment necessitate a compromise in clinical accuracy. Finally, guided by recent findings on the importance of strict evidence grounding [11], our RAG pipeline is designed to decouple retrieval from generation, ensuring all models are evaluated on the exact same high-signal clinical context.

3 Methodology

This section outlines the architecture of our evaluation framework. To ensure a fair, standardized comparison, the system is designed as a strict two-phase RAG pipeline, completely decoupling the retrieval phase from the generation phase.

3.1 Two-Phase RAG Architecture

The system operates by first transforming raw medical guidelines into an indexed vector database. When a user submits a query, the retrieval engine calculates similarity metrics to extract the top-k most relevant text blocks. Crucially, this retrieval step is executed only once per question prior to the generation loop. This guarantees that all evaluated language models receive the exact same retrieved context, eliminating retrieval variance from our benchmarking results.

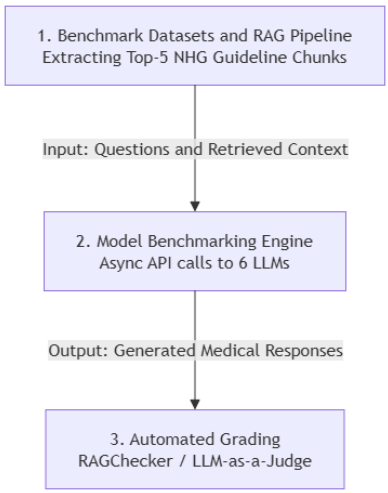

Figure 1: High-level overview of the decoupled evaluation pipeline, illustrating the data hand-off from retrieval to generation and automated grading.

3.2 Experimental Setup and Data

To accurately evaluate the systems, our experiments utilize two custom-built datasets derived from a curated corpus of 10 official Dutch NHG-guidelines, which were chunked into 2,516 distinct text blocks:

- **Factual QA Benchmark:** 192 questions testing direct knowledge retrieval. The raw question serves as both the LLM prompt and the retrieval query.
- **Clinical QA Benchmark:** A stratified subset of 200 patient vignettes (from a total pool of 375) testing complex clinical reasoning. Each vignette includes a separate, condensed `retrieval_query` optimized specifically for vector search.

The retrieval phase was executed using Google's `gemini-embedding-2` model, employing

an asymmetric embedding strategy stored within a local Qdrant vector database using the HNSW algorithm. To balance context richness with the context-window limits of the generators, the retriever was configured to return a maximum of top-5 chunks. Furthermore, a strict cosine similarity threshold of ≥ 0.7 was applied. This specific threshold was empirically determined during the initial retrieval pipeline optimization [3] to maximize the signal-to-noise ratio, ensuring that only highly relevant medical chunks were passed to the generation phase.

To assess the trade-off between intelligence, cost, and privacy, we evaluated six state-of-the-art LLMs. The closed-source models included GPT-5.5 (xhigh), GPT-5.4 (xhigh), and Claude Opus 4.7 (max). The open-source models, accessed via APIs to simulate local deployment capabilities, included Kimi K2.6, DeepSeek V4 Pro, and GLM-5.1. The benchmarking engine queried these models concurrently using `asyncio`, triggering maximum reasoning parameters where available.

3.3 Automated Evaluation Paradigm

Because manual grading of thousands of medical responses is infeasible and subjective, we employ an automated "LLM-as-a-Judge" methodology using the RAGChecker framework [7]. In this paradigm, GPT-5.5 acts as the adjudicator, mathematically comparing the generated chatbot answers against pre-validated ground-truth answers using a grading prompt mathematically validated via Cohen's Kappa. This allows for the extraction of fine-grained diagnostic metrics, such as clinical accuracy, faithfulness to the source text, and sensitivity to noisy retrieval.

4 Experimental Results

4.1 Overall Performance Summary

Table 1 summarizes the raw computational metrics and RAGChecker accuracy scores across the evaluated models.

Table 1: Model Benchmark Summary: Cost, Speed, and RAGChecker Performance

Model	Type	Avg Cost ($/query)	Avg Speed (tok/s)	Factual F1 (%)	Clinical F1 (%)	Combined F1 (%)	Avg Reason. (tokens)
`Claude Opus 4.7`	Closed-Source	$0.0369	56.8	56.2	61.5	58.9	0
`GPT-5.4`	Closed-Source	$0.0231	59.0	64.1	71.4	67.8	1087
`GPT-5.5`	Closed-Source	$0.0304	37.0	67.1	70.1	68.6	586
`DeepSeek V4 Pro`	Open-Source	$0.0022	60.5	68.8	70.5	69.7	1330
`GLM-5.1`	Open-Source	$0.0061	59.6	65.6	70.5	68.0	0
`Kimi K2.6`	Open-Source	$0.0075	68.7	63.6	66.7	65.2	0

As observed in the summary table, the performance landscape is highly varied. While `DeepSeek V4 Pro` and `GPT-5.5` dominate the accuracy metrics (achieving Combined F1 scores of 69.7% and 68.6% respectively), the variance in computational overhead is substantial. Inference speeds range from 37.0 tokens/second to 68.7 tokens/second, and per-query costs fluctuate by over an order of magnitude. These overarching results are decomposed and analyzed in detail in the following subsections.

4.2 Factual Retrieval vs. Clinical Reasoning

Figure 2 illustrates the performance disparities between simple factual retrieval and complex clinical reasoning. Counterintuitively, despite our initial expectations, all six models achieved higher Overall F1 scores on the complex clinical benchmarks than on the simple factual benchmarks.

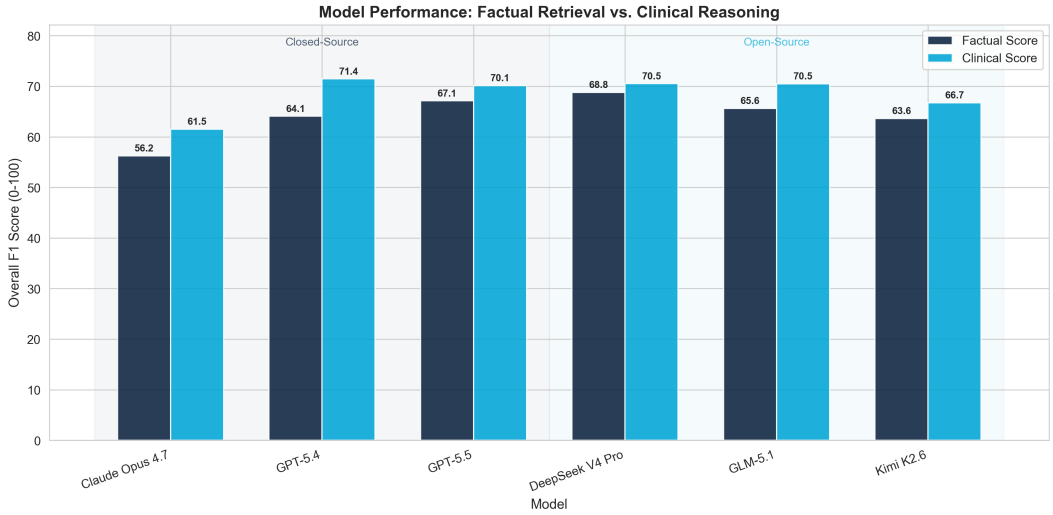

Figure 2: Model Performance: Factual Retrieval vs. Clinical Reasoning. Open-source models demonstrate highly competitive clinical reasoning capabilities.

While the questions for both datasets were generated using an identical `few_shot_cot` strategy [2, 9], this performance gap is highly attributable to architectural differences in the retrieval phase. The factual benchmark used raw, conversational questions for vector retrieval. In contrast, the clinical benchmark decoupled these steps: it utilized a condensed, keyword-optimized `retrieval_query` to search the database, while providing the LLMs with a rich, structured patient vignette as the reasoning prompt. Consequently, the clinical phase yielded higher-precision context chunks without conversational noise, anchoring the models' reasoning algorithms more effectively.

To qualitatively illustrate this, we compared the retrieval behavior between the factual and clinical pipelines. While the factual benchmark performed adequately when medical terms naturally dominated the sentence, it proved highly susceptible to semantic dilution. For example, when testing the factual approach with a full conversational query (e.g., *"Wanneer moet spalkbehandeling bij carpaletunnelsyndroom worden gestaakt?"*), common phrasing such as "Wanneer moet" and "worden gestaakt" skewed the embedding vector. This caused the database to match on general semantic similarity rather than clinical intent. As a result, it returned irrelevant chunks discussing evidence uncertainty rather than the actual stopping criteria, causing `context_precision` to drop to 0.15.

In stark contrast, the clinical benchmark bypassed this noise via its decoupled architecture. When resolving a conversational vignette about an overweight asthma patient, the pipeline exclusively used the condensed `retrieval_query`: *"gewichtsverlies astma overgewicht cont-*

role verbetering". By stripping away narrative verbs and pronouns, the retriever successfully isolated the exact, laser-focused NHG protocol. This direct contrast confirms that optimized retrieval queries provide the generative models with high-signal context, driving the superior performance seen in the clinical benchmark.

As shown in Figure 2, grounded in this superior clinical context, the open-source `DeepSeek V4 Pro` achieved the highest Factual F1 score (68.8%), outperforming all closed-source models in this domain. In the clinical benchmark, the closed-source `GPT-5.4` achieved the highest overall score (71.4%). However, `DeepSeek V4 Pro` and `GLM-5.1` followed closely, tying for the highest open-source Clinical F1 score (70.5%) and successfully outperforming OpenAI's flagship `GPT-5.5` (70.1%). Claude Opus 4.7 struggled significantly across both domains.

4.3 Computational Efficiency Trade-offs

To answer Sub-question 2, Figure 3 maps the models' Combined F1 scores against their API costs and inference speeds. The cost disparity is extreme: closed-source models command a massive premium without delivering proportionate accuracy. GPT-5.5 costs approximately $0.030 per query, whereas DeepSeek V4 Pro achieves a superior combined accuracy at less than 1/13th of the cost ($0.0022 per query).

Regarding latency, the open-source `Kimi K2.6` delivered the fastest inference speed (68.7 tokens/sec), establishing it as a highly efficient lightweight option. Conversely, GPT-5.5 was the slowest model tested (37.0 tokens/sec). The marker sizes in Figure 3 illustrate the "Reasoning Tax"—models like DeepSeek and GPT-5.4 generated over 1,000 hidden reasoning tokens per query, utilizing computational time to "think" before outputting their highly accurate clinical answers.

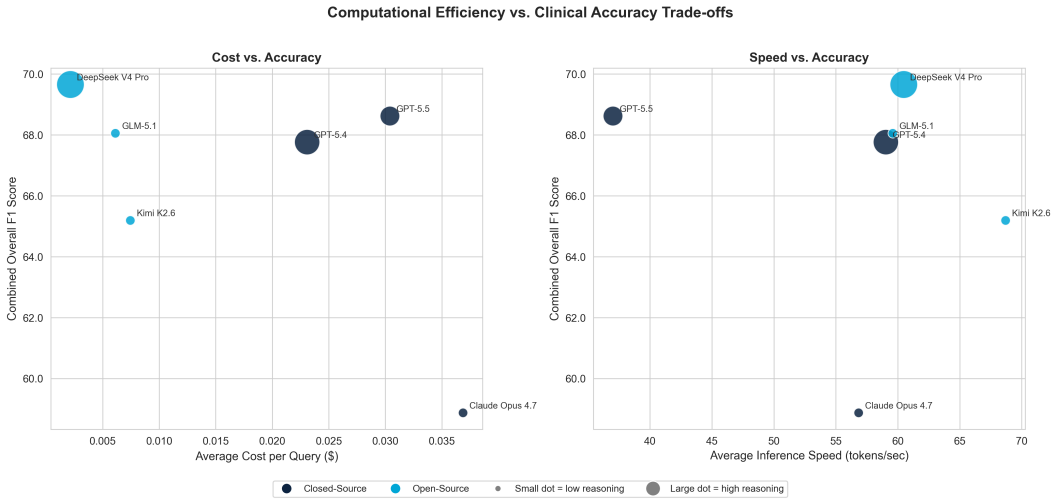

Figure 3: Computational Efficiency vs. Clinical Accuracy Trade-offs. Marker size indicates the volume of reasoning tokens utilized during inference.

4.4 Generator Safety and RAG Diagnostics

Clinical accuracy must be contextualized with model safety. Figure 4 presents RAGChecker's diagnostic metrics for generator reliability. All models exhibited extremely high Faithfulness (>85%), indicating a strong adherence to the retrieved Dutch NHG-guidelines without defaulting to internal hallucinations.

Crucially, DeepSeek V4 Pro and GPT-5.5 demonstrated exceptional robustness against poor retrieval, scoring the lowest in Irrelevant Noise Sensitivity (2.6 and 2.5, respectively). This proves that when the RAG pipeline retrieves off-topic text chunks, these models are intelligent enough to ignore the noise rather than incorporating it into their medical advice, which is paramount for patient safety.

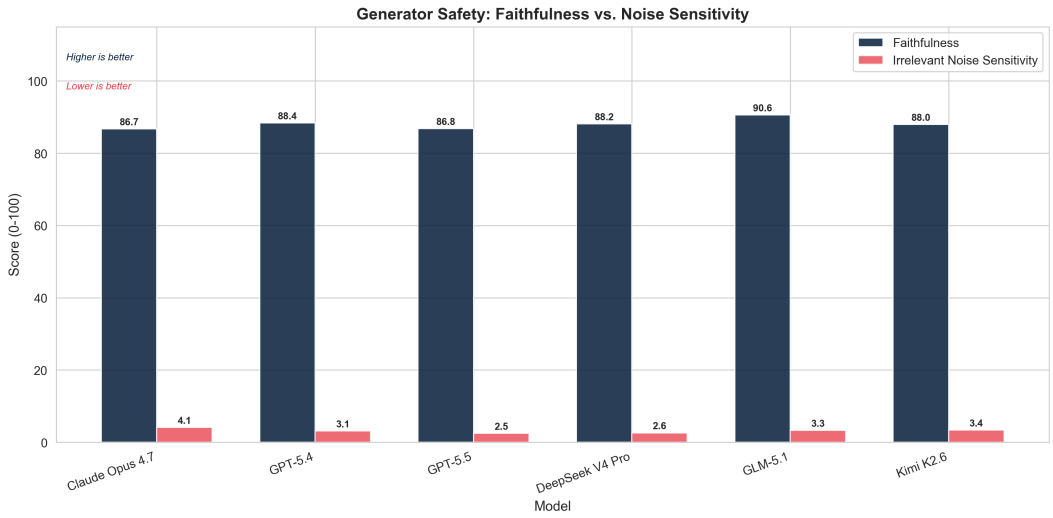

Figure 4: Generator Safety Diagnostics. Higher faithfulness indicates strict adherence to guidelines, while lower noise sensitivity indicates robustness against poor retrieval.

4.5 Qualitative Analysis: Divergent Model Behaviors

To provide concrete evidence of when open-source and closed-source models diverge in quality, we isolated two specific edge cases within the clinical benchmark where model behaviors did not align.

Case A: The Verbosity Penalty (Open-Source Superiority). In a vignette asking who the main practitioner is for a stable heart failure patient returning from the cardiologist, the pipeline retrieved the correct context. The open-source `DeepSeek V4 Pro` generated a highly precise answer matching the guideline (F1 = 0.909). Conversely, `GPT-5.5` suffered a severe drop in accuracy (F1 = 0.364). Rather than answering concisely, GPT-5.5 over-generated, hallucinating unsolicited clinical scenarios (e.g., kidney function, hypertension) not present in the prompt. This demonstrates that open-source reasoning models can outperform proprietary models by strictly adhering to the prompt without verbose over-generation.

Case B: The Parametric Safety Net (Closed-Source Superiority). Closed-source models maintain a distinct advantage when the retrieval pipeline fails entirely. In a vignette re-

garding gestational diabetes checkup frequencies, the RAG pipeline failed to meet the 0.7 cosine similarity threshold, returning zero chunks. Forced to answer without context, the open-source `GLM-5.1` failed completely (F1 = 0.000), hallucinating an incorrect timeframe and explicitly asking the user to provide the guideline text. In stark contrast, `GPT-5.5` leveraged its massive internal parametric memory to accurately recall the exact Dutch guideline ("annually for the first 5 years"), achieving an F1 of 0.889. This highlights a critical trade-off: while open-source models are highly cost-effective and secure, flagship closed-source models provide a vastly superior safety net for edge cases where the retrieval database fails.

5 Responsible Research

5.1 Ethical Considerations

The deployment of AI in healthcare involves critical ethical and legal challenges, primarily governed by the GDPR and the EU AI Act. Using closed-source APIs for clinical tasks inherently requires transmitting sensitive patient data to third-party servers, posing a severe breach of data sovereignty. This research directly addresses this ethical dilemma by investigating whether open-source models—which can be hosted locally "on-premise" to ensure 100% patient data privacy—are viable alternatives to closed-source models.

5.2 Reproducibility

To ensure full reproducibility, the retrieval phase of this experiment is entirely decoupled from the generation phase. All models are fed the exact same top-5 retrieved context chunks, eliminating retrieval variance. Furthermore, standard generation parameters (such as 'xhigh' reasoning effort and 'adaptive' thinking) are explicitly defined for the respective API calls, and the automated grading is conducted using the open-source RAGChecker framework with a validated prompt.

To further ensure transparency and facilitate future research, the complete Python benchmarking engine, along with the automated RAGChecker evaluation scripts and data pipelines, has been open-sourced and is publicly available.[1]

5.3 Use of Generative AI Tools

In compliance with the TU Delft guidelines on generative AI in end projects, the use of AI tools in this research is explicitly disclosed. As the core focus of this thesis, various generative AI models (including GPT-5.5, Claude Opus 4.7, and DeepSeek V4 Pro) were utilized as the primary subjects of experimental benchmarking. Additionally, GPT-5.5 was employed as an automated evaluation tool ("LLM-as-a-Judge") within the RAGChecker methodology.

During the research and writing process, generative AI assistants (including ChatGPT and Google Gemini) were utilized strictly in a supporting capacity. Their scope of application was limited to: summarizing relevant academic literature for the related work section, assisting with Python scripting for data visualization (e.g., `matplotlib` charting), and refining the grammatical clarity and LaTeX formatting of the final manuscript. All AI-assisted code and text were rigorously reviewed, verified, and critically evaluated by the author, who retains full intellectual and creative responsibility for the contents of this thesis.

[1] `https://github.com/Nuffs/RP_NHG_2026`

6 Conclusions and Future Work

The primary objective of this research was to determine how open-source LLMs compare to closed-source models for querying Dutch primary care guidelines, specifically balancing clinical accuracy with computational efficiency and patient data privacy.

The empirical results from our automated RAGChecker framework fundamentally challenge the prevailing assumption that expensive, proprietary models are required for complex medical reasoning. In our evaluation, the open-source `DeepSeek V4 Pro` emerged as the undisputed leader. It achieved the highest Combined F1 score (69.7%), outperforming OpenAI's flagship GPT-5.5 (68.6%) in both factual retrieval and complex clinical scenarios. Furthermore, DeepSeek exhibited identical safety metrics to GPT-5.5 regarding noise sensitivity, proving it is highly robust against retrieval errors. Most significantly, it achieved this state-of-the-art performance at a staggering fraction of the cost—averaging $0.0022 per query compared to GPT-5.5's $0.0304. However, qualitative analysis did reveal one distinct advantage for proprietary models: in edge cases where the retrieval pipeline failed entirely (0 chunks), closed-source models acted as a superior "parametric safety net," utilizing internal memory to answer correctly where open-source models hallucinated or failed.

Despite this edge case, these findings provide compelling, data-driven evidence for Dutch healthcare institutions. Because high-parameter open-source models can be hosted securely on-premises, they completely eliminate the GDPR compliance risks associated with cloud-based Big Tech APIs. As demonstrated by this study, opting for an open-source architecture is no longer a compromise on intelligence; rather, it is a highly accurate, exceptionally cost-effective, and legally secure solution for the future of clinical AI deployment.

While this study provides a robust comparative baseline, it is subject to certain limitations. Notably, the benchmarking was conducted using a fixed retrieval configuration, including a static cosine similarity threshold of ≥ 0.7. Subsequent experiments within the broader project group have since indicated that dynamic thresholds and hybrid sparse-dense retrieval (e.g., combining BM25 with vector search) may yield superior context retrieval. Furthermore, the performance disparity between our factual and clinical benchmarks highlighted that RAG systems are highly sensitive to query phrasing, demonstrating that decoupled, keyword-based retrieval is vastly superior to raw conversational queries.

Future work should investigate how updating these retrieval parameters impacts the downstream clinical accuracy and noise sensitivity of these specific LLMs. Additionally, subsequent research should investigate whether these open-source models maintain this high level of clinical reasoning when scaled beyond structured guidelines to unstructured, real-world Electronic Health Records (EHRs) or live doctor-patient transcripts.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Yannick Ter Heerdt, and my responsible professor, Dr. Jie Yang, for their invaluable guidance, constructive feedback, and for facilitating the API access necessary to conduct this research.

This project would not have been possible without the exceptional collaboration of my peer group. I extend my thanks to Leander for engineering the core RAG retrieval pipeline, to Charlene and Anne-Sophie for meticulously constructing the factual and clinical benchmark datasets, and to Nathaniel for his dedicated work on the qualitative failure mode analysis.

Finally, I would like to extend a special thanks to Willem-Rein Kooiman for sharing his valuable thoughts on this research, and for connecting me with Dr. Koos. Furthermore, I am deeply grateful to Dr. Koos for taking the time to share his clinical expertise. His real-world insights into the deployment of AI in Dutch primary care, particularly regarding medical ethical perspectives, were instrumental in shaping the practical context and motivation of this work.

References

[1] L M Amugongo, P Mascheroni, S Brooks, S Doering, and J Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877, 2025.

[2] Charlene Bakker. Automated benchmark construction for factual question answering over nhg guidelines. Bachelor's Thesis, Delft University of Technology, 2026.

[3] Leander Bindt. Creating a retrieval-augmented generation pipeline for the guidelines of the dutch college of general practitioners. Bachelor's Thesis, Delft University of Technology, 2026.

[4] Luc Builtjes, Joeran Bosma, Mathias Prokop, Bram van Ginneken, and Alessa Hering. Leveraging open-source large language models for clinical information extraction in resource-constrained settings. *JAMIA Open*, 8(5):ooaf109, 2025.

[5] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.

[6] Omid Kohandel Gargari and Gholamreza Habibi. Enhancing medical ai with retrieval-augmented generation: A mini narrative review. *Digital Health*, 11, 2025.

[7] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

[8] Karan Singhal, Tao Tu, Juraj Gottweis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2025.

[9] Anne-Sophie Straathof. Generating and evaluating an automated dutch clinical qa benchmark grounded in the nhg guidelines. Bachelor's Thesis, Delft University of Technology, 2026.

[10] Yutaka Sugihara, Aleksandar Milosavljevic, S Skaidre Jankovskaja, and M Magnus Falk. A review for navigating the trade-offs: evaluating open-source and proprietary large language models for clinical and biomedical information extraction. *Frontiers in Digital Health*, 8, 2026.

[11] Krzysztof Wołk. Evaluating retrieval-augmented generation variants for clinical decision support: Hallucination mitigation and secure on-premises deployment. *Electronics*, 14(21):4227, 2025.

[12] Shan Xu, Zhaokun Yan, Chengxiao Dai, and Fan Wu. Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health*, 13, 2025.

[13] Cyril Zakka, Rohan Shad, Akshay Chaurasia, et al. Almanac-retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068, 2024.