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

Abdelrahman Tageldin [1] Supervisors: Jie Yang & Yannick Ter Heerdt Committee: Pradeep Murukannaiah

[1]BSc Computer Science and Engineering, Delft University of Technology, The Netherlands

Background & Motivation

Retrieval-Augmented Generation (RAG) grounds LLM responses in trusted external documents, reducing hallucination risk – critical for clinical decision support. Implementing RAG over the Dutch NHG-guidelines presents a massive opportunity, but selecting the foundational model creates a tension between intelligence and privacy.

- **The Privacy Dilemma:** Closed-source models (GPT-5.5, Claude Opus 4.7) offer state-of-the-art reasoning, but require sending sensitive patient data to external servers – a severe GDPR risk with high API costs.
- **The Local Alternative:** Top-tier open-source models (DeepSeek V4 Pro, Kimi K2.6) can be hosted on a hospital's own secure servers, guaranteeing 100% data privacy.
- **The Knowledge Gap:** Do open-source models possess the clinical reasoning and Dutch language comprehension needed to safely replace Big Tech models?

Research Question

Main RQ: *"How do open-source language models compare to closed-source models on automated NHG factual and clinical benchmarks?"*

Sub-Questions:

- **SQ1:** How do models compare on simple factual retrieval versus complex clinical reasoning?
- **SQ2:** What is the exact trade-off between computational efficiency (speed and cost) and clinical accuracy?

Methodology

We benchmark 6 state-of-the-art LLMs as the generation engine of a strict, decoupled two-phase RAG pipeline, guaranteeing every model sees the exact same context.

Corpus & Data:

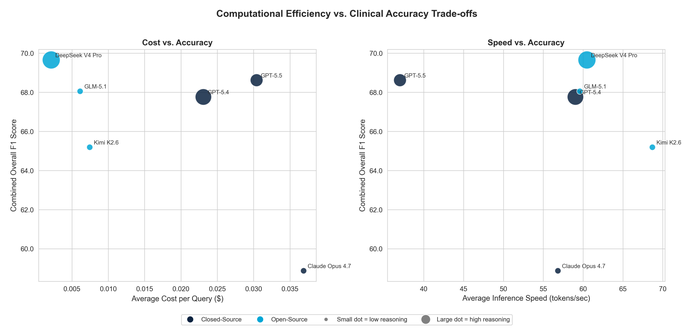
- 10 official NHG-guidelines, chunked into **2,516** text blocks; indexed with Google's `gemini-embedding-2` in a local Qdrant (HNSW) vector store.
- **Factual QA:** 192 direct-retrieval questions.
- **Clinical QA:** 200 patient vignettes, each utilizing a separate, condensed retrieval query to optimize vector search.

Models Tested:

- **Closed:** GPT-5.5 (xhigh), GPT-5.4 (xhigh), Claude Opus 4.7 (max).
- **Open:** DeepSeek V4 Pro, Kimi K2.6, GLM-5.1.

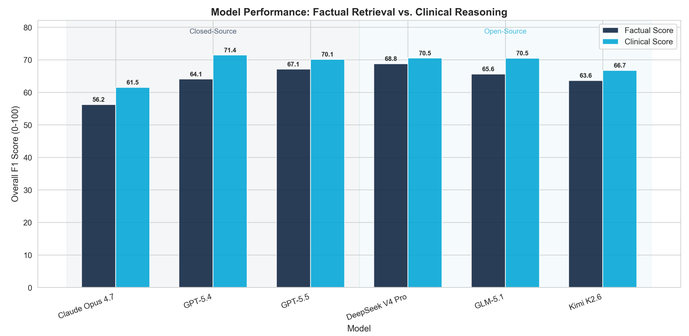
Evaluation: Automated "LLM-as-a-Judge" grading via **RAGChecker**, scoring Accuracy (F1), Faithfulness, and Noise Sensitivity, alongside Cost & Speed.

Computational Efficiency vs. Clinical Accuracy



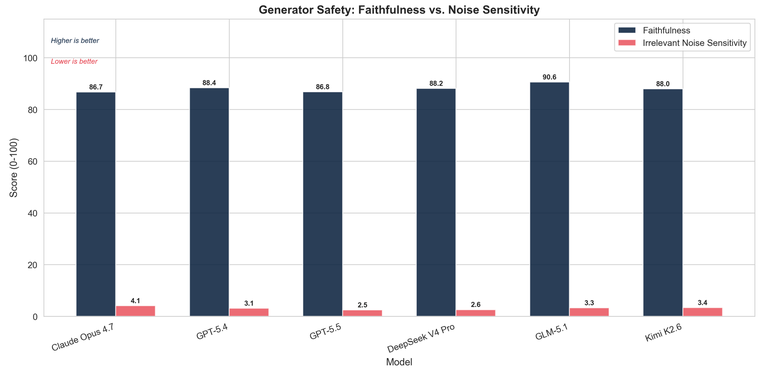
- **The Open-Source Win:** DeepSeek V4 Pro achieves the highest Combined F1 score (69.7%) at $0.0022/query – outperforming OpenAI's flagship GPT-5.5 (68.6%) while being over **13× cheaper**.
- **The "Reasoning Tax":** DeepSeek and GPT-5.4 exhibit slower inference speeds because they utilize $> 1,000$ hidden "thinking" tokens per query. Kimi K2.6 remains the fastest lightweight option (68.7 tok/s).

Factual Retrieval vs. Clinical Reasoning



- **Surprising Trend:** All six models scored *higher* on complex clinical vignettes than on simple factual questions.
- **The Cause (Decoupled Retrieval):** Clinical queries used condensed, keyword-optimized retrieval. Factual queries used raw conversational text, which semantically diluted the embeddings and retrieved noisier context.
- **Leaderboard:** GPT-5.4 led on Clinical F1 (71.4%); DeepSeek V4 Pro tied for the top open-source score (70.5%), beating GPT-5.5.

Generator Safety Diagnostics



- **Faithfulness:** All models adhered strictly to the retrieved Dutch guidelines ($> 85\%$ Faithfulness), mitigating severe internal hallucinations.
- **Noise Robustness:** DeepSeek V4 Pro & GPT-5.5 are the most robust to irrelevant context (lowest Noise Sensitivity: 2.6 & 2.5 respectively) – a critical safeguard for patient safety when the retriever fails.

Qualitative Insights: Divergent Behaviors

Case A – Verbosity Penalty (Open-Source Win): On a heart-failure vignette, DeepSeek V4 Pro answered concisely matching the guideline (F1 = 0.909). GPT-5.5 hallucinated unsolicited extra clinical scenarios, diluting its accuracy (F1 = 0.364).

Case B – Parametric Safety Net (Closed-Source Win): When retrieval failed entirely (returning 0 chunks), GLM-5.1 crashed and hallucinated (F1 = 0.000). GPT-5.5 safely recalled the correct Dutch guideline from its internal memory (F1 = 0.889).

Conclusions & Future Work

- Open-source **DeepSeek V4 Pro** matches or outperforms flagship closed-source GPT-5.5 on clinical accuracy and noise robustness – at a fraction of the cost.
- Choosing an open-source, on-premise architecture is no longer a compromise on intelligence; it is an accurate, highly cost-effective, and GDPR-compliant path for clinical AI.
- **Future Work:** Implement hybrid sparse-dense retrieval (BM25 + vectors) to fix factual retrieval bottlenecks, and scale testing to unstructured Electronic Health Records.

References

[1] Ru et al. (2024). RAGChecker. *NeurIPS*. **[2]** Es et al. (2024). RAGAS. *EACL*. **[3]** Singhal et al. (2025). Expert-level medical QA with LLMs. *Nature Medicine*. **[4]** Zakka et al. (2024). Almanac – RAG for clinical medicine. *NEJM AI*. **[5]** Builtjes et al. (2025). Open-source LLMs for clinical NLP. *JAMIA Open*. **[6]** Wolk (2025). RAG for clinical decision support. *Electronics*. **[7]** Amugongo et al. (2025). RAG for LLMs in healthcare. *PLOS Digital Health*.

Code & Data: `github.com/Nuffs/RP_NHG_2026`