Creating a Retrieval-Augmented Generation Pipeline for the Guidelines of the Dutch College of General Practitioners

Leander Bindt

1 Introduction

- **Unsustainable workloads:** General Practitioners face extreme daily schedules (>9 hours even inside delegated teams) [6].
- **An AI opportunity:** Large Language Models (LLMs) speed up medical searches but risk dangerous clinical hallucinations [8].
- **The fix:** Retrieval-Augmented Generation (RAG) can safely ground AI to the guidelines from the Dutch College of General Practitioners [10].
- **The research gap:** However, off-the-shelf RAG systems fail on the highly dense medical terminology, deeply nested sub-headers, and the required information being often fragmented across multiple guidelines [3].

Research Question

How can a RAG pipeline be constructed for Dutch NHG-guidelines?

While evaluating the overall goodness of a system involves many possible quality dimensions, this research strictly isolates two:

- **Clinical factuality:** Preventing the generation of false clinical claims (hallucination).
- **Storage scalability:** Making sure the system is useable and scalable for real-world clinical settings.

2 Related Work

- **Fine-tuning vs RAG:** Creating a new model like Med-PaLM 2 requires massive computing power and domain-specific datasets [1]. While adding a system like Almanac bypasses these requirements by grounding off-the-shelf models with external clinical repositories [10].
- **BM25 keyword search:** The industry-standard algorithm utilizing term saturation, length normalization, and header weighting for structured text [7].
- **Hybrid retrieval:** BM25 and an AI meaning based vector search complement each other [2].
- **Model grounding:** Using instructions and examples improves trustworthiness, improving grounded refusal capabilities by 45%.
- **Trust-score evaluation:** Using a two-dimensional metric based on grounded refusal and claim recall, it is possible to grade the grounding of the system [9].

TU Delft

The system implements an Approximate Nearest Neighbor (ANN) search to not loop over the entire database every time.

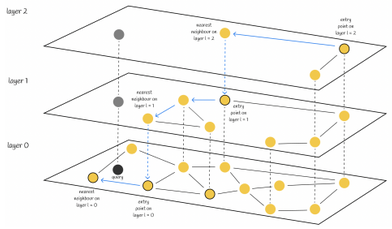

Figure 1: Approximate Nearest Neighbor with a Hierarchical Navigable Small Worlds (HNSW) graph [4].

- **HNSW:** A graph structure that splits the embeddings into multiple layers of increasing density.
- **ANN:** Figure 1 shows the algorithm. A random starting point is picked, which moves through the graph until the distance to the prompt point is below a certain threshold [5].

3 Methodology

- **Retrieval optimization:** When setting a strict amount of context blocks limit, what is the recall accuracy for the BM25, AI meaning, and hybrid search algorithms?
- **Instruction grounding:** For the baseline, instruction, and example prompts, what are the claim recall and enforced grounded refusal metrics?

4 Results

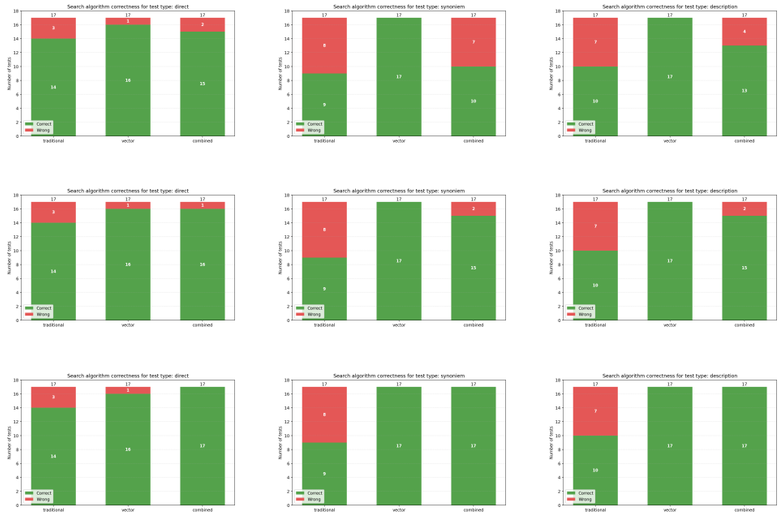

Figure 2: The amount of correctly returned context blocks for three different limits after reranking. 20, 32 and 44 from top to bottom. The columns represent the three searching algorithms which are the traditional, AI meaning, and hybrid search from left to right.

- **Meaning search superiority:** AI meaning search consistently outperforms traditional BM25 retrieval across all prompt categories as seen in figure 2.
- **Hybrid recall accuracy:** Combining the searches achieves 100% recall accuracy, allowing users to scale block limits to fit budgetary requirements.

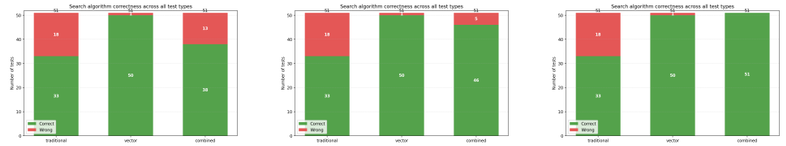

Figure 3: The overall test results when limiting the amount of context blocks to 20, 32 and 44 blocks from left to right.

	Baseline	Instruction	Example
Grounded refusal	0/4	4/4	3/4
Claim recall	6/6	6/6	6/6

Grounded refusal drops when adding examples because there was usable information in one of the example context blocks.

5 Conclusions and Future Work

This research successfully results in a fully functional, reliable pipeline. In future work, key points can be:

- **Expert validation:** Have certified medical experts develop and clinically validate the prompt instructions and examples, alongside testing on a more extensive grounding dataset.
- **New models:** Explore newer embedding models, as they become more intelligent and cheaper.
- **Data maintenance:** Developing a system that can handle guideline updates without having to re-embed the whole database.

References

[1] R. Anil, A. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. Clark, L. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, and Y. Wu. *PaLM 2 Technical Report*. May 2023. DOI: 10.48550/arXiv.2305.10403.

[2] N. Arabzadeh, X. Yan, and C. Clarke. *Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection*. Sept. 2021. DOI: 10.48550/arXiv.2109.10739.

[3] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek. *Seven Failure Points When Engineering a Retrieval Augmented Generation System*. 2024. arXiv: 2401.05856 [cs.SE].

[4] V. Efimov. *Similarity Search, Part 4: Hierarchical Navigable Small World (HNSW)*. Towards Data Science. Image retrieved from https://towardsdatascience.com/wp-content/uploads/2023/06/1ziU6_KIDqfmaDXKAlcMa8w.png; part of the article series "Similarity Search, Part 4". June 2023.

[5] Y. Malkov and D. Yashunin. "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (Mar. 2016). DOI: 10.1109/TPAMI.2018.2889473.

[6] S. Porter Boyd and Laiteerapong. *Revisiting the Time Needed to Provide Adult Primary Care*. 2023. DOI: 10.1007/s11606-022-07707-x.

[7] S. Robertson and H. Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3 (Sept. 2009), pp. 333–389. DOI: 10.1561/1500000019.

[8] K. Singhal, T. Tu, and J. e. a. Gottweis. "Toward expert-level medical question answering with large language models". In: *Nature Medicine* 31.3 (2025), pp. 943–950.

[9] M. Song, S. Sim, R. Bhardwaj, H. Chieu, N. Majumder, and S. Poria. *Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse*. Sept. 2024. DOI: 10.48550/arXiv.2409.11242.

[10] C. Zakka, A. Chaurasia, R. Shad, A. R. Dalal, J. L. Kim, M. Moor, K. Alexander, E. Ashley, J. Boyd, K. Boyd, K. Hirsch, C. Langlotz, J. Nelson, and W. Hiesinger. *Almanac: Retrieval-Augmented Language Models for Clinical Medicine*. 2023. arXiv: 2303.01229 [cs.CL].