Optimal LLM Size for Medical Document Classification Using Context Engineering

Data Sovereignty Procedures for Doctors (DSP4D)

true

2025

Abstract

This paper investigates the minimum viable Large Language Model (LLM) size required for reliable medical document classification and clinical action generation. We evaluate multiple context engineering strategies—including few-shot learning, retrieval-augmented generation (RAG), and long-context approaches—to determine optimal trade-offs between model size, inference cost, and clinical accuracy. Our experiments focus on edge deployment scenarios where data sovereignty requirements mandate local processing.

Keywords: Large Language Models, Few-Shot Learning, Medical Document Classification, Edge Deployment, Data Sovereignty

Contents

1 Introduction

Doctors face an increasing volume of medical documents requiring timely review and action. After office hours, the challenge of efficiently processing X-ray results, lab reports, and specialist referrals becomes critical for patient care.

This research addresses a fundamental question: *What is the smallest LLM that can reliably classify medical documents and generate appropriate clinical actions?*

1.1 Motivation

1.2 Research Questions

1. What is the minimum model size for reliable document classification (>95% accuracy)?
2. How do different context engineering strategies affect the size-accuracy trade-off?
3. Can sub-3B parameter models achieve clinical safety standards with appropriate context?

1.3 Contributions

- A systematic evaluation framework for medical document classification with LLMs
- Comparative analysis of context engineering strategies (few-shot, RAG, long-context)
- Practical deployment recommendations for edge devices

2 Background

2.1 Large Language Models and Scaling Laws

2.2 Context Engineering Strategies

2.2.1 Few-Shot Learning

In-context learning enables models to perform tasks by conditioning on examples provided in the prompt (Brown et al. 2020).

2.2.2 Retrieval-Augmented Generation

2.2.3 Long-Context Approaches

2.3 Medical Document Processing

3 Methodology

3.1 Data Source: GraSCCo

Instead of generic document types, this research utilizes the **Graz Synthetic Clinical text Corpus (GraSCCo)** (Lohr et al. 2025; Modersohn et al. 2022). GraSCCo is the first publicly shareable, multiply-alienated German clinical text corpus, designed specifically for clinical NLP tasks without compromising patient privacy.

The corpus provides a diverse set of clinical scenarios, which we use to evaluate the models' ability to classify document intent and generate appropriate clinical actions based on German-language clinical narratives.

3.2 Experimental Setup

3.2.1 Models Evaluated

Model	Parameters	Deployment
Llama 3.2	1B	Edge/WebLLM
Llama 3.2	3B	Edge
Phi-3 Mini	3.8B	Edge/WebLLM
Llama 3.1	7B	Hosted

3.2.2 Context Strategies

1. **Zero-Shot** — Instructions only (baseline)
2. **One-Shot** — Single example
3. **Few-Shot** — 3-5 curated examples
4. **RAG** — Retrieved guidelines/similar cases

3.3 Evaluation Metrics

- **Classification Accuracy** — Correct document type identification
- **Action Appropriateness** — Clinical validity of suggested actions
- **Latency** — Inference time on target hardware

4 Experiments

4.1 Classification Task

4.2 Action Generation Task

4.3 Breakpoint Analysis

5 Results

5.1 Size vs. Accuracy Trade-offs

5.2 Impact of Context Engineering

5.3 Deployment Viability

6 Discussion

6.1 Implications for Clinical Practice

6.2 Limitations

6.3 Future Work

7 Conclusion

8 Data Availability

The datasets generated and/or analyzed during the current study are available in the Zenodo repository: **Graz Synthetic Clinical text Corpus (GraSCCo) v2** (Lohr et al. 2025).

9 References

Brown, Tom B., Benjamin Mann, Nick Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–901.

Lohr, Christina, Franz Matthies, Jakob Faller, et al. 2025. *GraSCCo_PII_V2 - Graz Synthetic Clinical text Corpus with PII Annotations*. Version v2. https://doi.org/10.5281/zenodo.15747389.

Modersohn, Luise, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. "GRASCCO—the First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus." In *German Medical Data Sciences 2022–Future Medicine: More Precise, More Integrative, More Sustainable!* IOS Press.