Optimal LLM Size for Medical Document Classification Using Context Engineering

Data Sovereignty Procedures for Doctors (DSP4D)

Semesterarbeit

Studiengang:	CAS Generative KI
Autor*in:	Benjamin Haegler, Christian Sprecher
Betreuer*in:	[Betreuer einfügen]
Auftraggeber*in:	[Auftraggeber einfügen]
Expert*in:	[Experte einfügen]
Datum:	2025

Berner
Fachhochschule

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

Inhaltsverzeichnis

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?

2 Theory / State of Research

2.1 Context Engineering Strategies

2.1.1 Few-Shot Learning

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

2.1.2 Retrieval-Augmented Generation

Short description what it is, and why we did NOT use it.

2.2 Prior Work

3 Methodology

3.1 Procedure

Siehe nice Graphik von Beni

3.2 Data Source: GraSCCo

Instead of generic document types, this research utilizes the **Graz Synthetic Clinical text Corpus (GraSCCo)** (Lohr u. a. 2025; Modersohn u. a. 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 reports.

The task we give the models is to update a patients health record (HBA) based on supplied clinical report.

3.3 Golden Answer Generation

Due to lack of access to expert medical knowledge, we generate golden answers as ground truth for the models by asking a state of the art LLM to create those. We then validated at least a subset of those answers with a medical expert.

3.4 Experimental Setup

3.4.1 Architecture

Maschine von Beni, Chrigels notebook, Google cloud für Gemini, Evaluationsframeworks

3.4.2 Models Evaluated

TBD!!

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

3.4.3 Context Engineering Strategies

1. **Zero-Shot** - Instructions only (baseline)
2. **One/Few-Shot** - Multiple examples with Golden Answers
3. **Prompt Chaining**

TBD by Beni

3.5 Evaluation Metrics

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

4 Results

4.1 Impact of LLM Size

Compare the metrics including latency and inference cost.

4.2 Impact of Context Engineering

Compare the context engineering strategies for each model.

5 Discussion / Conclusion

5.1 Implications for Clinical Practice

5.2 Limitations

5.3 Future Work

List of Figures

List of Tables

Glossary

Context Engineering The practice of designing prompts and providing relevant information to improve LLM performance on specific tasks.

Edge Deployment Running machine learning models locally on devices rather than in the cloud.

Few-Shot Learning Providing a small number of examples in the prompt to guide model behavior.

GraSCCo Graz Synthetic Clinical text Corpus — a German clinical text corpus for NLP research.

RAG (Retrieval-Augmented Generation) A technique that combines information retrieval with text generation to improve accuracy.

References

[] Brown, Tom B., Benjamin Mann, Nick Ryder, u. a. 2020. «Language Models are Few-Shot Learners». *Advances in Neural Information Processing Systems* 33: 1877–901.

[] Lohr, Christina, Franz Matthies, Jakob Faller, u. a. 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, und 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.

Appendix

A. Prompt Templates

B. Detailed Results

Selbständigkeitserklärung

Ich bestätige, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der im Literaturverzeichnis angegebenen Quellen und Hilfsmittel angefertigt habe. Sämtliche Textstellen, die nicht von mir stammen, sind als Zitate gekennzeichnet und mit dem genauen Hinweis auf ihre Herkunft versehen.

Ich bestätige weiterhin, dass ich bei der Erstellung dieser Studienarbeit durchgehend steuernd gearbeitet habe und von einer KI erzeugte Texte bzw. Textfragmente nicht unreflektiert übernommen habe.

Ort, Datum: Unterschrift: