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

Evaluating the performance of language models requires quantifiable metrics that capture both accuracy and semantic quality. While subjective assessment remains valuable, reproducible benchmarks enable systematic comparison across models and configurations. This section reviews established evaluation frameworks — from classical NLP metrics through modern LLM-based assessment methods — and situates them within the medical domain where accuracy requirements are particularly stringent.

2.1 Evaluations in Classical Text Analysis

In classical natural language processing (NLP) and information retrieval, evaluation relies heavily on comparing system output against a “gold standard” or ground truth. These metrics are particularly relevant for classification tasks, such as identifying clinical intent or extracting specific medical entities.

2.1.1 String Similarity & Edit Distance

When exact matches are too strict, string similarity metrics quantify the difference between two sequences.

- **Levenshtein Distance** (or Edit Distance) counts the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word or text string into the other (Levenshtein 1966). This is valuable for correcting typos or measuring near-matches in entity extraction.

2.1.2 Classification Metrics

For tasks involving categorization, the confusion matrix serves as the foundation for most metrics, tracking true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Manning u. a. 2008).

- **Accuracy** measures the overall correctness of the model but can be misleading in unbalanced datasets, which are common in medical contexts (e.g., rare diseases).
- **Precision** (Positive Predictive Value) measures the proportion of identified positive cases that were actually correct. In a clinical setting, high precision minimizes false alarms.
- **Recall** (Sensitivity) measures the proportion of actual positive cases that were identified. High recall is critical in medicine to ensure no pathology is overlooked.
- **F1-Score** provides the harmonic mean of precision and recall, offering a balanced view when finding a compromise is necessary (Sokolova und Lapalme 2009).

2.1.3 Generation Metrics

For tasks involving text generation, such as summarizing findings or suggesting actions, classical n-gram based metrics are often employed:

- **BLEU (Bilingual Evaluation Understudy)** measures the precision of n-grams in the generated text compared to reference texts. While popular, it is often criticized for focusing only on exact matches and ignoring semantic meaning (Papineni u. a. 2002).
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** improves upon BLEU by incorporating stemming and synonym matching, resulting in better correlation with human judgment (Banerjee und Lavie 2005).

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** focuses on recall, measuring how much of the reference text appears in the generated output, widely used for summarization (Lin 2004).

While these metrics provide objective, reproducible scores, they often correlate poorly with human judgment for complex reasoning tasks, necessitating more advanced evaluation paradigms.

2.1.4 Semantic & Embedding-based Metrics

To overcome the limitations of exact n-gram matching, semantic metrics utilize word or sentence embeddings to measure similarity in meaning rather than just surface form.

- **BERTScore** computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embeddings (e.g., from BERT). This allows for a more robust evaluation of paraphrases and synonyms (Zhang u. a. 2020).
- **Word Mover's Distance (WMD)** and its variants (like MoverScore) measure the minimum "distance" required to move the embedded words of one document to the other. This approach captures semantic distance effectively, even when no words overlap (Kusner u. a. 2015; Zhao u. a. 2019).

2.1.5 LLM-Based Evaluation (LLM-as-a-Judge)

Recent advances have shifted towards using Large Language Models themselves as evaluators, a paradigm known as "LLM-as-a-Judge". This approach uses the reasoning capabilities of capable models (such as GPT-5) to assess the quality of generated text based on complex criteria such as helpfulness, safety, and coherence, often achieving higher correlation with human judgment than traditional metrics.

- **G-Eval** is a framework that uses LLMs with Chain-of-Thought (CoT) reasoning to evaluate generated text. By decomposing the evaluation task into a series of steps, it provides fine-grained scores that align closely with human preference (Liu u. a. 2023).
- **GPTScore** evaluates texts by calculating the probability of the generated text given a specific instruction or context, using the model's own likelihood scores as a proxy for quality (Fu u. a. 2024).
- **Prometheus** is an open-source LLM specifically fine-tuned for evaluation purposes. It allows for custom evaluation criteria and feedback generation, offering a cost-effective alternative to using proprietary models like GPT-4 as judges (Kim u. a. 2024).

- **Ragas** (Retrieval Augmented Generation Assessment) is a framework specifically designed for evaluating RAG pipelines. It defines metrics such as *context precision*, *faithfulness*, and *answer relevancy*, using an LLM to verify if the generated answer is grounded in the retrieved documents and if it actually answers the user's question (Es u. a. 2024).

2.2 LLM in the Context of Medical Science

The application of Large Language Models (LLMs) in medicine is an evolution of clinical Natural Language Processing (NLP), which gained significant momentum with the release of specialized models like ClinicalBERT (Alsentzer u. a. 2019). While early models focused on entity recognition and extraction, modern LLMs offer the potential to summarize charts and suggest clinical actions. However, their integration into clinical workflows is constrained by critical requirements for accuracy, data privacy, and data sovereignty.

2.2.1 Privacy, Security, and Data Sovereignty

The use of cloud-based LLMs in healthcare introduces significant risks that have been documented since the early days of transformer models.

- **Data Leakage and Memorization:** Foundational research has shown that LLMs can memorize and inadvertently "regurgitate" sensitive training data, including personally identifiable information (PII) (Carlini u. a. 2021). In a medical context, this poses a risk of exposing protected health information (PHI) through model outputs.
- **Adversarial Vulnerabilities:** Modern aligned models are susceptible to adversarial attacks, such as prompt injection, which can bypass safety filters and potentially lead to the disclosure of sensitive context or the generation of incorrect medical advice (Zou u. a. 2023).
- **Ethical and Regulatory Gaps:** A 2025 scoping review identifies a persistent lack of ethical oversight and informed consent in many LLM-based medical studies, highlighting an urgent need for privacy-preserving architectures (**Zhong2025Considerations?**).

To mitigate these risks, researchers are exploring **Data Sovereignty**—the principle that health data should remain under the control of the originating institution or the patient. This has led to two main research directions:

1. **On-Device Deployment:** Operating models entirely on local hardware (e.g., Jetson Nano) to ensure no sensitive data ever leaves the clinical environment (Wu u. a. 2025).

2. **Privacy-Preserving Training:** Techniques like “Whispered Tuning” and differential privacy are being developed to prevent PII memorization during model adaptation (**Singh2024WhisperedTuni**

2.2.2 Specialized Medical Applications

Dual-stage and Lightweight Patient Chart Summarization
Wu et al. (2025) proposed a dual-stage system specifically for emergency departments. By using a Small Language Model (SLM) on embedded devices, they demonstrate that it is possible to provide actionable clinical summaries without cloud dependencies, thereby fulfilling the highest standards of data sovereignty (Wu u. a. 2025).

ELMTEX: Structured Clinical Information Extraction
Guluzade et al. (2024) showed that fine-tuned smaller models can outperform larger, general-purpose counterparts in extracting structured data from unstructured German clinical reports. Their work demonstrates that for specialized medical tasks, increased parameter count does not guarantee improved performance — a finding that supports the feasibility of local deployment (Guluzade u. a. 2024).

GraSCCo: A Foundation for Privacy-Preserving Research
The Graz Synthetic Clinical text Corpus (GraSCCo) remains a cornerstone for this research area. As a multiply-alienated German clinical corpus, it allows researchers to benchmark models on realistic medical narratives without the legal and ethical risks associated with real patient data (Modersohn u. a. 2022; Lohr u. a. 2025).

2.3 Scaling Laws and Model Efficiency

A central question for deploying LLMs in privacy-sensitive environments is: how small can a model be while maintaining acceptable performance? Early scaling laws suggested a straightforward trade-off, but recent developments in Small Language Models (SLMs) have significantly shifted expectations.

2.3.1 Historical Context

Early work by Kaplan et al. (2020) and Hoffmann et al. (2022) established that language model performance follows predictable power-law relationships with model size and training data (**kaplan2020scaling?**; **hoffmann2022training?**). While foundational, these findings predate the current generation of highly optimized small models and do not fully capture the capabilities of modern SLMs.

2.3.2 The Rise of Small Language Models

A comprehensive survey by Lu et al. (2024) benchmarked 59 SLMs (100M–5B parameters) across commonsense reasoning, mathematics, and in-context learning tasks. Their findings reveal substantial performance improvements: SLMs improved by 10–13% between 2022 and 2024, outpacing larger models which improved by only 7.5% over the same period (**lu2024slmsurvey?**). Notably, the Phi-3 model (3.8B parameters) achieves 69% on MMLU — performance comparable to Mixtral 8x7B and GPT-3.5. This demonstrates that modern SLMs, through optimized architectures and high-quality training data, can compete with models several times their size.

2.3.3 A Note on Terminology

The term “Small Language Model” warrants clarification. In current usage, “small” refers exclusively to parameter count — not to training data scope. A 3B parameter model trained on trillions of web-scale tokens is considered “small” only relative to 70B+ frontier models. This stands in contrast to *domain-specific* models such as ClinicalBERT or PubMedBERT, which are smaller in both parameters and training scope, having been trained on specialized medical corpora. Throughout this thesis, the term SLM refers to language models with fewer than 100 billion parameters, regardless of their training data origin. This broader definition encompasses both general-purpose compact models (Phi, Qwen, Llama) and domain-specialized models, allowing for comparison across deployment scenarios.

2.3.4 Capability Density and the Densing Law

Xiao et al. (2025) formalize this trend through the concept of *capability density* — defined as capability per parameter. Their empirical analysis reveals a “densing law”: capability density approximately doubles every 3.5 months (**xiao2025densing?**). This trajectory indicates that equivalent performance can be achieved with exponentially fewer parameters over time, making local deployment increasingly viable.

2.3.5 Edge Deployment Considerations

Recent work specifically addresses SLM deployment on resource-constrained devices. Hassanpour et al. (2025) systematically evaluate SLMs for edge scenarios, examining the trade-offs between model size, quantization levels, and task performance (**hassanpour2025edge?**). Their findings confirm that sub-4B parameter models can achieve practical utility for domain-specific tasks when properly configured — a key consideration for medical applications where data must remain on-device.

2.3.6 Implications for This Study

These developments frame the research question: given hardware constraints of on-device deployment for sensitive medical data, what is the smallest pre-trained model that can reliably perform clinical document classification? The answer depends not only on parameter count, but also on model generation and — as the following section explores — context engineering strategies that can augment smaller models at inference time.

2.4 Context Engineering Strategies

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

[] Alsentzer, Emily, John Murphy, Willie Boag, u. a. 2019. «Publicly Available Clinical BERT Embeddings». *arXiv preprint arXiv:1904.03323*.

[] Banerjee, Satanjeev, und Alon Lavie. 2005. «METEOR: An automatic metric for MT evaluation with improved correlation with human judgments». *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

[] Carlini, Nicholas, Florian Tramer, Eric Wallace, u. a. 2021. «Extracting Training Data from Large Language Models». *30th USENIX Security Symposium*, 2633–50.

[] Es, Shahul, Jithin James, Luis Espinosa Anke, und Steven Schockaert. 2024. «Ragas: Automated Evaluation of Retrieval Augmented Generation». *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.

[] Fu, Jinlan, See-Kiong Ng, Zhengbao Jiang, und Pengfei Liu. 2024. «GPTScore: Evaluate as You Desire». *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[] Guluzade, Aynur, Naguib Heiba, Zeyd Boukhers, u. a. 2024. «ELMTEX: Fine-Tuning LLMs for Structured Clinical Information Extraction. A Case Study on Clinical Reports». *arXiv preprint arXiv:2404.04475*.

[] Kim, Seungone, Jamin Shin, Yejin Cho, u. a. 2024. «Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models». *arXiv preprint arXiv:2405.01535*.

[] Kusner, Matt, Yu Sun, Nicholas Kolkin, und Kilian Weinberger. 2015. «From word embeddings to document distances». *International conference on machine learning*, 957–66.

[] Levenshtein, Vladimir I. 1966. «Binary codes capable of correcting deletions, insertions, and reversals». *Soviet physics doklady* 10 (8): 707–10.

[] Lin, Chin-Yew. 2004. «Rouge: A package for automatic evaluation of summaries». *Text summarization branches out*, 74–81.

[] Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, und Chenguang Zhu. 2023. «G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment». *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[] 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.

[] Manning, Christopher D, Prabhakar Raghavan, und Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[] 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.

[] Papineni, Kishore, Salim Roukos, Todd Ward, und Wei-Jing Zhu. 2002. «Bleu: a method for automatic evaluation of machine translation». *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–18.

[] Sokolova, Marina, und Guy Lapalme. 2009. «A systematic analysis of performance measures for classification tasks». *Information Processing & Management* 45 (4): 427–37.

[] Wu, Jiajun, Swaleh Zaidi, Braden Teitge, u. a. 2025. «Dual-stage and Lightweight Patient Chart Summarization for Emergency Physicians». *arXiv preprint arXiv:2510.02900*.

[] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, und Yoav Artzi. 2020. «BERTScore: Evaluating Text Generation with BERT». *International Conference on Learning Representations*.

[] Zhao, Wei, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, und Steffen Eger. 2019. «MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance». *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 563–78.

[] Zou, Andy, Zifan Wang, J Zico Kolter, und Matt Mattjung. 2023. «Universal and Transferable Adversarial Attacks on Aligned Language Models». *arXiv preprint arXiv:2307.15043*.

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:

______________________________ ______________________________