Nazrul Kabir

Adapting Open-Source LLMs for Finnish Digital Scribe Systems: A Performance Evaluation

Metropolia University of Applied Sciences

Master of Engineering

Information Technology

Master's Thesis

9 June 2025

Abstract

Author:	Nazrul Kabir
Title:	Adapting Open-Source LLMs for Finnish Digital Scribe Systems: A Performance Evaluation
Number of Pages:	41 pages + 1 appendix
Date:	9 June 2025
Degree:	Master of Engineering
Degree Programme:	Information Technology
Professional Major:	Medical Technology
Supervisors:	Sakari Lukkarinen, Senior Lecturer Päivi Haho, Principal Lecturer

---

Digital scribe systems are emerging as promising tools to help doctors and nurses by documenting notes during patient visits so that they can spend more time on patient care but the challenge of adapting these systems for underrepresented languages remains largely unresolved. This thesis explores whether training open-source Large Language Models (LLMs) on specialty-specific data impacts the performance compared to general models. The study used simulated clinical conversations, which were turned into audio and text pairs to fine-tune and test how well the model can produce accurate and clear transcriptions.

Open-source LLaMA 3.1–8B model has been used with a 7-fold cross-validation setup. The model was evaluated using standard Natural Language Processing (NLP) metrics such as BLEU, ROUGE, and BERTScore to evaluate the similarity and meaning of the generated text.

The results show that with proper preprocessing and training open-source LLMs can be adapted for Finnish-language digital scribes. However, there are still challenges like dataset size, generalizability, and lack of clinical expert validation that highlight the need for future work. This research sets some early benchmarks and methodological approaches for creating AI-powered digital scribes in Finnish.

Keywords: Digital Scribe, AI, Large Language Models, healthcare AI

Contents

List of Abbreviations

List of Abbreviations

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
EHR	Electronic Health Record
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LoRA	Low-Rank Adaptation
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation

1 Introduction

The healthcare sector faces growing challenges with administrative workload, particularly with the clinical documentation, contributing significantly decreased time for direct patient interaction and clinician burnout. A study by AI in social and health care (SOTE) reveals that in one Finnish wellbeing services county, documentation takes between 5-30 minutes per visit. This is especially in psychiatry, where manual documentation demands are high [1].

To tackle this challenge, digital scribe systems have emerged as promising tools. These systems are designed to help doctors and nurses by documenting notes during patient visits so that the healthcare workers can spend more time on patient care. The industry reports show that more healthcare providers are starting to use digital scribes, particularly in the United States. Finland is not too far behind as there are several digital scribe pilots running as well [2]. These digital scribes utilize automatic speech recognition (ASR) and natural language processing (NLP) technologies to generate clinical documentation from patient-clinician conversations [3]. Digital scribes help by automatically transcribing and summarizing clinical notes. This saves clinicians time on documentation so they can focus more on patient care and potentially reducing burnout [4].

Despite their potential, the development and implementation of effective digital scribe systems comes with a number of challenges. A primary concern is the thoroughness and accuracy of the documentation generated by these systems [5]. Clinical conversations are pretty complicated, often non-linear, and rich in domain-specific terms that varies across medical specialties. The challenge of accurately capturing all relevant clinical information while filtering out irrelevant content remains a obstacle in digital scribe development [6].

This research aims to address a critical gap in the development of digital scribes by directly evaluating how different training methods affect the quality of digital scribe outputs on the fine-tuned open-source LLMs on multilingual European

datasets, particularly Finnish. The main objective is to determine whether training digital scribes with specialty-specific data leads to better performance compared to general models. The study focuses on evaluating key usability aspects such as how accurate, relevant, and complete these generated notes are within a specific medical domain.

Solving this problem is essential for the broader adoption of AI-driven documentation tools in European healthcare systems. Not only does it contribute in reducing the clinician workload, but it also ensures that patient data is recorded in a standardized, structured and legally compliant manner.

1.1 Problem Statement

Digital scribes are intended to support clinical workflows by enabling more efficient documentation processes, allowing healthcare professionals to dedicate more attention to patient care. However, the effectiveness of these systems depends heavily on their ability to generate accurate and comprehensive notes and due to the lack of proper usability healthcare workers are resistant to using it which mostly depends on high-quality transcribed data.

In multilingual settings and across various medical specialties, challenges arise due to differences in medical terminologies, documentation styles, and clinical focus across different medical specialties. A key question is whether training digital scribe systems on open-source LLMs could enhance their performance, particularly in low-resource languages—including most of the Nordic and Baltic languages—lack high-quality annotated datasets, benchmarks, and pretrained models for medical NLP. Furthermore, there is a lack of research on how optimized training strategies (e.g., custom loss functions) and structured output mechanisms affect the usability and compliance of automatically generated clinical documents.

1.2 Research Objectives and Questions

This study aims to enhance the performance of open-source LLMs for the generation of structured clinical notes from transcribed speech. The research addresses the following question: Does training open-source LLMs on specialty-specific data impact the performance compared to general models?
To support this research question, the study focuses on the following sub-objectives:

1. Fine-tune open-source LLM on multilingual datasets, with emphasis on Finnish-language medical text.
2. Evaluate the generated notes in terms of usability, structure, and compliance.
3. Assess the potential impact of the optimized models on clinical documentation workflows in European healthcare systems.

Such findings could directly inform the product development of digital scribes and training strategies, ultimately improving clinical usability and patient safety.

1.3 Scope and Limitations

The study focuses on evaluating the potential for performance improvement of digital scribes trained on specialty-specific data in the Finnish-language healthcare contexts. The primary objective is to evaluate whether specialized datasets to train LLMs leads to better documentation compared to general models. To test this, the research used simulated data that mimics real situations and an open-source LLaMA 3.1(8B) model, which has been fine-tuned on patient-doctor conversations. This helps to understand if a more focused approach can lead to better documentation compared to the generic models.

The objective is not to develop a fully functional, ready-for-the-market tool. Instead, it focuses on figuring out possibilities of training models to learn about the limits and looking at how good the documentation is. The focus is mainly on technical evaluation, such as accuracy, structure, and consistency. Speech-to-

text integration and structure note generation, although relevant to the future of digital scribes, they're not part of this phase. So, the results might contribute to early-phase development.

1.4 Importance of Study

Accurate and complete documentation ensures that all relevant clinical information is readily available to all members of the care team for safe and coordinated patient care. Insufficient or inaccurate data makes digital scribes less effective, impacting clinical workflows. However, clinicians often spend a considerable portion of their time on documentation, reducing time for patient interaction, which can lead to burnout [4]. As a potential solution, digital scribes use the Large Language Model (LLM) to generate documentations but they often struggle with specialization, especially in languages like Finnish.

This study addresses a timely and somewhat underexplored question: Can training digital scribes with specialty-specific data improve their performance in non-English, field-specific contexts? By concentrating on Finnish-language medical data, this research helps fill a gap in the world of NLP and LLM development. This is really important for smaller language communities where ready-made solutions usually fall short [7]. The results could help shape the future development of tailored documentation tools that make automated documentation easier for healthcare workers. Thus, speed up the adoption of digital scribes in clinical practice, reducing the administrative burden on clinicians and improving the overall efficiency of healthcare delivery.

2 Literature Background and Related Works

This section provides an overview of important topics relevant to the research, including challenges in clinical documentation, advancements in Large Language Models (LLMs), fine-tuning methods for domain-specific applications, and usability considerations. AI tools Keenious and NotebookLM were used to find relevant literature and selection was made by the author based on the title and abstract. The usage of these tool has been explained in details in section 2.5.

2.1 Documentation Challenges

Keeping good clinical documentation is important for delivering safe and high-quality patient care [8], but it is also time-consuming for healthcare professionals. A potential solution to this issue is automating clinical documentation which has thus led to the growing interest in AI-assisted documentation tools.

Speech recognition technologies have made it easier to transcribe but it can struggle with clinical terms, different voices, and switching languages. Even though traditional speech-based technology has improved, it still struggles with understanding context, accuracy and adapting to the changes still remains a challenge [9]. To address these problems, recent research has been looking at how to integrate advanced NLP models that can understand context, semantics and medical-specific language which has shifted efforts toward transformer-based architectures. A good example of this is ClinicalBERT, it uses the transformer model and is trained on extensive clinical notes [10].

2.2 Evolution of Large Language Models

The field of Natural Language Processing (NLP) has made great progress by using neural network-based representations, also known as embeddings. Older methods for word embedding like GloVe, word2vec and fastText, learned a single vector representation for each unique word [11]. However, these methods have limitations because they don't consider the context of the words. The need to

model long-range dependencies and how distant words interact, especially relevant for complex texts like clinical notes [12], led to the evolution of contextual representations.

The development of contextual embeddings, such as BERT (Bidirectional Encoder Representations from Transformers) and LLaMA have revolutionized Natural Language Processing (NLP) by offering word representations that change meaning depending on the word around them. After the introduction of attention-based Transformer architecture, BERT was introduced and showed better results in many NLP tasks [12]. And since then BERT and related models have become a de facto standard for embedding text segments and single words in context. Thus the rise of contextual embeddings has really boosted clinical NLP tasks like concept extraction [11].

Pretraining Large Language Models (LLMs), like BERT, with data specific to certain areas has really helped improve results on many NLP tasks. However, initial efforts were often aimed at general domain corpora [13]. There are two main branches of pre-trained LLMs in the general domain: BERT variants and Generative Pre-trained Transformer (GPT) variants. BERT variants have been heavily studied in the biomedical domain for tasks that involve distinguishing between options, GPT variants have been pre-trained on large-scale biomedical literature to address generation capabilities. A key milestone in this evolution was the launch of ClinicalBERT, which built on the BERT architecture by pre-training on millions of clinical notes [10]. This adaptation really improved the model to understand better medical terms, abbreviations, and specific details in patient records [10]. Following this progress, BioGPT used generative transformer architecture and was trained using biomedical literature, like PubMed abstracts which enabled BioGPT to excel in generating biomedical text, mining information, extracting relations supporting applications like summarizing research articles and answering biomedical questions [14]. Recently, Me-LLaMA has made some exciting progress by using the LlaMA architecture and pulling in a variety of medical data sources. It has brought in special adapters that help it adapt to different clinical tasks, like diagnosis prediction and note generation, showing

great results in multiple medical Natural Language Processing (NLP) benchmarks [15]. Transformers are at the cutting edge of AI, are trained on large datasets across various domains like text and speech. They're especially popular in Advanced Speech Recognition (ASR) frameworks for their ability to capture complex relationships in the data [16].

Despite these significant advancements, there's still one issue: most of these models including ClinicalBERT, BioGPT and Me-LLaMA mainly use English-language data for training. Which restricts their usability in places like Finland, where healthcare documents need to be in Finnish or Swedish and often mix languages. While addressing language issues specific to Finnish might not solve universal issue, but it shows problem that smaller or less common languages face in clinical NLP. Addressing this gap is a key for the adoption of Large Language Model (LLM)-driven solutions in healthcare around the world, especially in multilingual and limited resources.

2.3 Fine-tuning LLMs for Clinical Contexts

To make general Language Models (LMs) better for the medical field, researchers have focused on creating specialized medical language models [18]. This development process usually involves tailoring existing general Large Language Models (LLMs) using techniques like pre-training, fine-tuning, and prompting.

By pre-training, general LLMs can gain valuable medical knowledge and become aligned with what the medical domain needs. Fine-tuning LLMs with clinical and multilingual data is proving to be a promising way to improve their performance in specific NLP tasks. In healthcare, regular LLMs usually struggle with the complex terminology, formats, and terms that are common in medical settings [11]. This means existing LLMs underperform with very specific types of text, such as medical text, that contain many uncommon words compared to the general vocabulary.

Fine-tuning involves adapting a pre-trained model to a specific task or dataset, allowing it to get a better grip on the medical terms, abbreviations and documentation styles that matter in different specialties. It produces much better results when models are trained on well-annotated medical data in tasks like Named Entity Recognition (NER), document classification and summarization [17]. Also, in multilingual settings, fine-tuning with data for specific languages helps the model to handle clinical texts that aren't in English, which is really important for healthcare in countries like Finland. Recent studies shows that fine-tuned models do a better job than general-purpose LLMs for clinical tasks, especially when the training data matches real-life documentation formats [13]. This approach helps the model recognize words that sound similar and deal unclear phrases better. For digital scribes, fine-tuning is important to reduce hallucinations and improving the accuracy of the generated summaries. This study focuses on fine-tuning LLMs with simulated Finnish clinical conversations to see how effective customized digital scribes can work in real-world settings.

Besides fine-tuning the model weights, methods like Retrieval-Augmented Generation (RAG) are used to enhance LLMs in healthcare. RAG helps LLMs to pull information from outside sources, like medical guidelines, for better and evidence-based reliable answers. This retrieval relies on embedding models that measure the similarity between document chunks and queries [18].

2.4 Summary of Gaps in Current Research

Although medical Natural Language Processing (NLP) has made notable progress, several gaps remain: limited support for minority and European languages in LLMs, underexplored impact of LLM-assisted documentation on real-world clinical workflows and lack of standard evaluation benchmarks for medical LLMs in clinical settings [15] and large, high-quality clinical datasets for multilingual clinical note generation [19].

Large Language Models (LLMs) have achieved remarkable success in NLP tasks but they mostly focus on a few major languages like English and Chinese. When

it comes to smaller European languages like Finnish, Swedish or Estonian, their performance is still not up to the mark [20]. In healthcare, it's essential to understand and generate language accurately. Take Finnish, for instance; it has a complex syntax that many models still struggle with [21]. Without a proper understanding of the language, there is a risk that these LLMs might create unsafe documentation.

Since most existing medical LLMs are still in the research phase, with limited application and validation in real-world clinical settings, we are not really clear on how LLMs affect everyday clinical work. Most of the researches focus on technical performance like how well a model generates a summary—not on how it helps clinicians or changes workflows [22] nor reporting on clinical validity and usability. There is a gap between research and real-world usage because companies offering digital scribes often do not share detailed validation information [4].

Another challenge is the lack of open, multilingual datasets for training and testing clinical text generation. Most of the publicly available datasets, like MIMIC-III or i2b2 are in English and focus on U.S. healthcare contexts [23]. Finding clinical datasets in languages like Finnish is rare due to strict data governance rules [24]. Without the right resources, it is quite difficult to make meaningful comparisons between models.

2.5 Literature Discovery using AI

At the beginning of this research, finding relevant and quality literature was critical for building the theoretical framework, choosing methods and fine-tuning the research questions. Because this research covers areas like digital scribes, LLM, language model development, model evaluation, usability evaluation, and digital health documentation, a standard keyword-based literature search was often insufficient. To improve the literature search, I used an AI-assisted tool called Keenious [25] alongside traditional sources like PubMed, Scopus, and Google Scholar. It helps to find contextually relevant papers based on the user's written

content like drafts or paragraphs. Instead of just using keywords like regular search engines, Keenious evaluates the context of the user's writing, enabling it to find articles that might not be found through manual query formulation. It can be integrated with Google Docs and Microsoft Word, to get real-time suggestions while writing.

Since the topic mixed both technical (ASR, NLP) and human-centered (usability, healthcare) domains, there was a need to explore literature beyond domain-specific databases. Keenious helped me find articles from related fields that enriched the research foundation.

While working on the background section about how the digital scribes aim to reduce the time clinicians spend on documentation I submitted the paragraph to Keenious by highlighting the texts and found several articles that I didn't come across in my initial PubMed keyword searches, for example: 'MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records' [26] because this paper's title doesn't have the keywords I used to search which was really helpful for the study. Overall Keenious helped find related literature, improved the theoretical background, and helped avoid potential gaps in coverage.

3 Current State Analysis

The process of clinical documentation is changing rapidly, driven by a growing need for solutions that can simplify the documentation process while enhancing the accuracy of medical records. Traditional documentation practices which often relies on manual note-taking or after-visit dictation, requires a lot of work and sometimes result in missing information. To tackle this, major technological trends have emerged in this space: digital scribe systems, Automatic Speech Recognition (ASR), and transformer-based Natural Language Processing (NLP) models to automate the documentation workflow.

3.1 Digital Scribe Systems

Digital scribe systems are designed to make note-taking in clinics healthcare by quietly listening conversations between patients and clinicians and creating organized notes from those conversations. Examples of commercially available systems like Dragon Copilot previously Dragon Ambient eXperience [27], Amazon HealthScribe [28], Notable [29], DeepScribe [30] and Augmedix [31] are created to automate clinical documentation. These systems use a combination of Automatic Speech Recognition (ASR) and natural language processing (NLP) techniques to convert clinical conversations into informative documentation suitable for Electronic Health Records (EHRs). Some platforms such as Dragon Copilot and Amazon HealthScribe offer real-time documentation while others like DeepScribe and Notable generate summaries asynchronously after visits.

Initial research suggests that there are benefits of time saved and higher user satisfaction [22], however, most healthcare systems designed are commercial, with limited transparency and little proof on real-world deployment or adaptation, especially for different languages or public healthcare settings.

3.2 Advancements in ASR and NLP

The last few years have seen significant progress in ASR and NLP, which has notably enhanced the capabilities of digital transcription systems, particularly with open-source models such as Whisper (by OpenAI), resulting in significant improvement in transcription accuracy [32]. Whisper is built with encoder-decoder architecture that can handle multiple languages, which makes it a better fit for real-world clinical settings rather than the traditional rule-based ASR systems.

Research has shown that context-sensitive word embeddings and attention-based neural networks enhance the performance ASR systems [16]. Despite these improvements, ASR systems still struggle in medical domains due to specialized vocabulary, speaker inconsistency, interruptions and background noise [33]. ASR by itself usually isn't sufficient enough for complete documentation. It must be paired with NLP tools which is capable of identifying key entities (e.g., medications, symptoms) and generating summaries. This has led to the need to combine ASR with domain-specific NLP models such as ClinicalBERT [14].

3.3 Transformer-based NLPs

Recent progress in clinical Natural Language Processing (NLP), especially with transformer-based models, including BioGPT [10], ClinicalBERT and Me-LLaMA have made a big difference by advancing various NLP tasks. ClinicalBERT [14] is trained on medical texts and clinical notes, and it does a better job than general NLP models when it comes to recognizing and classifying entities. BioGPT, made by Microsoft, extends transformer capabilities show strong results in generating biomedical text using PubMed abstracts. Me-LLaMA [15] is an adaptation of Meta's LLaMA model designed for clinical contexts.

These models use prior training on biomedical texts to improve contextual understanding, which has enhanced functionalities in summarization, named

entity recognition and clinical question answering. However, most of these are limited to English-language datasets, thereby limiting their practicality in smaller language regions, like Finnish. Additionally, most clinical NLP studies focus on performance metrics such as ROUGE, BERT or BLEU scores which are valuable for benchmarking but do not fully capture real-world usability or workflow integration in clinical environments [18]. Bridging this gap between finding a way to improve performance and keeping it relevant is a challenge for future development. To turn good prototypes into helpful healthcare tools, future research needs to focus on real-life evaluations, doing long-term studies in actual clinical settings and assessing how these systems affect clinician workload, the quality of paperwork, and patient results.

3.4 Challenges in Clinical Documentation

Clinicians often spend a considerable portion of their time on documentation, contributing to burnout and reducing time for patient interaction. Research shows that in a Finnish wellbeing service region, it takes 5-30 minutes to document each visit [1]. This burden gets worse because of the complexity and manual nature of clinical documentation processes [17].

For effective patient care, clinical documentation must be comprehensive and precise. Incomplete or inaccurate documentation can lead to gaps in care, miscommunication within care teams, and compromised patient safety. According to the review of clinical information extraction applications researchers looked into how clinical information extraction works and pointed out that the quality of the input data, especially clinical narratives, really matters [19]. If the documentation is incomplete it may result in critical data being missed during automated extraction, while inaccurate data can then spread through clinical processes, causing problems with analytics or leading to wrong conclusions. Therefore, keeping accurate and complete records is really important for taking care of patients and for other uses like research and measuring quality. Digital documentation tools need to be carefully designed and validated to make sure they keep the clinical information reliable throughout the whole process.

Many clinical NLP models rely on training data in English, thereby restricting their applicability in environments where multiple languages are spoken or English is not the primary language. This language barrier makes it tough for countries that use local languages for clinical notes to adopt NLP-based documentation tools. According to a study, general-purpose models like multilingual BERT may struggle with domain-specific tasks unless they are fine-tuned on localized data. Their development of Finnish BERT-based models shows why it's crucial to have language-specific tools that consider both the language and its context [34]. Without these changes, even the best models might not work well in clinical settings outside English-speaking areas. This points to the need for creating and sharing high-quality training datasets in underrepresented languages, so everyone can benefit from AI-driven innovations in healthcare.

3.5 Existing Solutions

Currently, many digital scribe systems are trained on general medical texts [29]. This provides them a decent grasp of medical language, but it might not cover all the specific terms and details found in different specialties. As a result, the documentation can be less accurate and comprehensive.

Training digital scribes on specialty-specific clinical notes can really boost their performance. By modifying the training to suit the particular needs and language of each specialty, these systems can produce clearer and more accurate documentation. This approach can overcome the limitations of general medical text training and make digital scribes more effective in healthcare settings [27].

Several healthcare technology companies have piloted digital scribe solutions and shown promising results. For example, Autoscriber and PremierScribe have reported significant time savings and improved documentation quality with the help of AI-powered digital scribes [35]. These case studies show how training focused on specialty-specific can help improve clinical workflows and overall patient care.

4 Methodological Approach

This chapter outlines the methodological approach of how the study was done, followed by a detailed description of the materials and datasets used to fine-tune an open-source large language model (LLM) on Finnish clinical data influences the accuracy, compliance, and usability of automatically generated clinical notes. The section begins with research design, then gives a clear description of the data sources and how the data was prepared. After that, it covers how the model was fine-tuned and evaluated. Finally, the chapter ends with a discussion on usability and compliance in real-world healthcare settings.

4.1 Research Design

The research follows a comparative experimental design to check how well a fine-tuned open-source large language model (LLM) works in a digital scribe system. The main goal is to assess how accurately the model transcribes and organizes clinical conversations in Finnish, and whether the generated output meet standards for clarity, thoroughness and regulatory compliance.

To achieve this, the research process is organized into three main phases:

1. Data Collection and Preprocessing: Sourcing a collection of simulated clinical conversations in Finnish that captures real interactions between patients and healthcare professionals. Then preprocessing them to ensure consistency, serving as the foundation for model training and evaluation.

2. Model Fine-Tuning and Optimization: Selecting suitable open-source large language model (LLM) model based on its performance potential and fine-tuning it using domain-specific Finnish dataset, while experimenting with customized loss functions. Designing a training pipeline that keeps clinical goals in mind while enhancing the model's ability to generate clear and useful results.

3. Evaluation and Benchmarking: Measuring the model's performance using quantitative metrics such as BLEU, ROUGE, BERTScore to assess alignment with reference data and language quality.

4.2 Integration of NotebookLM in Research Workflow

During this project, NotebookLM [36] quickly became much more than just another tool by acting as a genuine research assistant, making the literature review and research tasks easier as I went through a complex landscape of literature and data. NotebookLM allows to upload and organize different types of materials like PDFs, Google Docs, and web articles thus helping to create a dedicated, project-specific notebook. This centralization helped to easily keep track of more than twenty important research articles about Finnish-language LLMs and digital scribe systems which I found through literature review. Instead of manually searching through papers and citations, I was able to ask NotebookLM specific questions such as, “Which metrics can be used to evaluate text generated by a model trained on Finnish doctor-patient conversation?”. The system provided a clear, citation-backed answers supported by the sources uploaded.

NotebookLM's summarization tool was also helpful. The platform creates summaries, study guides, and timelines from the uploaded materials, making it easy to quickly identify research trends, best practices and gaps. For example, while comparing different evaluation metrics, I used NotebookLM to clarify the technical details of BERTScore. The tool highlighted that, when evaluating non-English text, leveraging contextual embeddings from multilingual BERT (BERTmulti) was essential. This was really relevant for our Finnish datasets, as BERTmulti’s cross-lingual capabilities enabled more accurate semantic similarity scoring than monolingual models—an insight that directly helped my metric selection and result interpretation.

Another standout feature was the Audio Overview tool which allows to convert lengthy, complex documents into structured podcast-style audio summaries.

Listening to these summaries helped me quickly get the main topics and different opinions without having to take notes even while I was commuting or taking a break.

Overall, NotebookLM made it easier to come up with hypotheses, understand different studies, and keep track of all research results. Its ability to centralize, analyze, and repurpose knowledge from a wide array of content formats made it a must-have tool for both thorough literature reviews and dynamic, ongoing research analysis.

4.3 Data Sources and Preprocessing

The performance of Large Language Model (LLM) is highly dependent on the quality, diversity and representativeness of training data, particularly in domain-specific applications like clinical documentation. Due to privacy regulations that limits access to real clinical conversations, a set of simulated Finnish clinical conversations was generated in collaboration with the Department of Healthcare and Nursing. These simulations were designed to mimic real patient-provider interactions in various situations like regular checkups, managing chronic conditions etc.

The dataset used in this research was developed in collaboration with a parallel thesis project [37]. Each clinical scenario was documented in both audio recording in MP3 format and a corresponding human-generated transcription in Finnish. The transcripts matched up with the audio to keep important details like medical terms, situation, feelings and the flow of conversation. All data files followed a consistent naming structure into clearly defined folders for the ease of organization and tracking.

Special care was taken to ensure the conversations sound realistic, simulating typical conversations between clinicians and patients. These conversations formed the basis for model training. The entire dataset preparation and alignment

were carried out by a team member of Digital Medical Scribe project whose thesis provides a detailed breakdown of the methods and data processing pipeline [37].

4.4 Model Fine-Tuning and Optimization

After getting the dataset ready, the next step involved fine-tuning an open-source Large Language Model (LLM) to understand and structure Finnish clinical conversations. For this purpose, Meta's LLaMA 3.1–8B model was selected and set up in a 4-bit quantized format with Low-Rank Adaptation (LoRA) adapters to enable efficient training on limited hardware resources. This configuration provided enough memory bandwidth and processing power to support model training and data loading in parallel.

For the fine-tuning, CSC's Puhti supercomputer, which has an NVIDIA A100 GPU (40 GB VRAM) with 16 CPU cores and 128 GB of RAM was used. This setup gave enough memory and processing power to run model training and load data at the same time. About 200 GB of disk space was used to keep the model's weights, training logs, results, and dataset. The model was trained for several rounds using the clinical conversations with hyperparameters adjusted to maintain generalization and making sure to keep the important clinical details [37].

4.5 Evaluation Framework and Criteria

This research focuses on assessing how well fine-tuned LLM work in clinical documentation. The primary objective is to determine whether fine-tuning an open-source LLM on Finnish-language clinical data improves the accuracy, compliance, and usability of automatically generated clinical notes compared to general-purpose models.

4.5.1 Quantitative Evaluation Metrics

To measure the textual similarity and accuracy of the notes generated by the model compared to reference notes, three main NLP metrics were used in this evaluation: BLEU, ROUGE, and BERTScore. Each metric is described in more details below.

BLEU (Bilingual Evaluation Understudy) is a quick, inexpensive and language independent method that evaluates how many n-grams (n consecutive words) in the generated text match the reference text. It is widely used for translation and summarization tasks and provides a score for each translated segment, usually sentences —by comparing them against a set of good-quality reference. The output is always a number between 0 and 1 where 1 is highest and represents more similar texts [38].
The formula for BLEU is the following:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} P_n\right)$$

Where:

- BP (Brevity Penalty) is a penalty term that adjusts the score for translations. It is calculated as minimum of 1, (reference_length/ translated_length), where reference_length is the total number of words in the reference and translated_length is the total number of words in the generated text.
- p_n is the precision of n-grams, which is calculated as the number of n-grams matches the reference text divided by the total number of n-grams in generated text.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a family of metrics (e.g., ROUGE-1, ROUGE-L) used to evaluate the quality of generated text by comparing the similarity between the generated text and reference (ground truth) text, making it ideal for assessing how well the model captures information from conversations. It's often used in summarization and text generation tasks [39].

BERTScore checks how similar generated text is compared to reference text by using BERT (Bidirectional Encoder Representations from Transformers) model embeddings by leveraging the semantic understanding. It evaluates how well the generated text captures the meaning of the reference text, not just the surface-level token overlap [40].

It also provides an option to do contextual analysis which involves a deeper examination of the semantic and contextual differences between the generated and ground truth text. It depends on contextual embeddings that understand words based on their surrounding text which helps identify why certain parts of the text are not aligning well.

Table 1. Summary of the main properties of the selected metrics.

Metric	**Focus**	**Advantage**	**Limitation**
BLEU	Precision	Word level, easy to calculate	It heavily relies on n-grams and may not capture the fluency or overall meaning
ROUGE	Recall	It considers overlap of n-grams, which helps in capturing the important content	It relies solely on the n-gram overlap, which may miss context.
BERT-Score	Semantics similarity	Handles synonyms and paraphrasing	Less interpretable, computationally heavier than ROUGE

4.5.2 Thresholds for Acceptable Performance

This study used benchmarks from previous work in natural language generation and clinical NLP to interpret evaluation results. These benchmarks provide practical targets instead of strict limits and contextual understanding.

- BLEU: A score between 0.3 and 0.5 are generally acceptable in summarization tasks [38].
- ROUGE Score: ROUGE-1 ≥ 0.5, ROUGE-2 ≥ 0.3, and ROUGE-L ≥ 0.4 are often considered baseline in clinical context [39].

- BERTScore: For paraphrased clinical notes, a BERTScore of ≥ 0.85 is recommended to ensure the meaning is accurate [40].

By using these metrics in combination, the study aims to provide a robust quantitative foundation for judging the overall model performance, balancing both syntactic correctness and semantic accuracy.

4.5.3 Excluded Metrics and Justification

Other metrics like F1, METEOR were considered but decided not to include them in the main evaluation for the following reasons:

F1 score is a way to measure how well a system does at identifying specific pieces of information and requires labeled classification data including true positives (TP), false positives (FP) and false negatives (FN) to compute precision and recall. This research focuses on generated clinical conversation instead of labeled data, so applying F1 would be inappropriate and misleading for this type of generative task [33].

METEOR uses synonym matching and stemming, but it’s performance in multilingual and specific areas, like Finnish clinical data is not consistent [12]. Also, it doesn't have strong support for Finnish.

5 Results and Analysis

This chapter presents and compares the results of the experimental evaluation conducted in this study using three widely adopted metrics: BLEU, ROUGE, and BERTScore, which assessed the performance of the fine-tuned open-source LLaMA 3.1–8B model on Finnish domain-specific datasets. The main focus is on determining whether fine-tuning with Finnish clinical dialogue improves the quality, accuracy, and usability of automatically generated clinical documentation. Two sets of evaluations were conducted independently by two researchers—referred to as Evaluation A (Author's evaluation) and Evaluation B (Team member's evaluation) [37]. Even though both evaluations used the same model and dataset, differences in evaluation metrics and implementation details led to some differences in the scores reported which is discussed and presented at the later section of this chapter. The results are organized into quantitative evaluations using standard NLP metrics and a discussion that connects findings to the original research objectives.

5.1 Author's Evaluation (Evaluation A)

The quantitative analysis in this section is based on the evaluation framework introduced in Section 4.3. The selected metrics—BLEU, ROUGE, and BERTScore were chosen to reflect different aspects of output quality. BLEU and ROUGE scores assess lexical overlap and similarity of the structure, while BERTScore provides a semantic comparison between generated text compared to the reference text. These metrics were selected to provide a comprehensive view of model performance, considering both surface-level and contextual accuracy. Both sets of evaluations results presented below offer insight into how fine-tuning and data preprocessing impact the accuracy, completeness, and clinical relevance of the generated documentation. This directly addresses main goal of the research of evaluating whether fine-tuning open-source LLMs on specialty-specific, multilingual datasets improves the performance of digital scribes compared to general-purpose models.

The evaluation process was conducted using Python, using the following libraries:

- BLEU is calculated using the nltk.translate.bleu_score module, which assesses how many n-grams matches [38].
- ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) is calculated with the rouge-score package, which measures lexical and structural overlap [39].
- BERTScore is calculated via the bert-score Python package, which evaluates semantic similarity using contextual embeddings [40].

5.1.1 Results Summary

BLEU Score:

Table 2. BLEU score evaluation result

File	**Score**
I01-G01-C01.txt	0.1421
I01-G01-C02.txt	0.1558
I01-G01-C03.txt	0.1147
I01-G02-C01.txt	0.1284
I01-G02-C02.txt	0.1544
I01-G03-C01.txt	0.1150
I01-G03-C02.txt	0.1289
Average (Mean ± SD)	0.1342 ± 0.0176

ROUGE Scores:

Table 3. ROUGE score evaluation result

File	ROUGE-1	ROUGE-2	ROUGE-L
I01-G01-C01.txt	0.5654	0.3108	0.5605
I01-G01-C02.txt	0.5815	0.3649	0.5801
I01-G01-C03.txt	0.5623	0.2927	0.5604
I01-G02-C01.txt	0.5814	0.3287	0.5814
I01-G02-C02.txt	0.5869	0.3521	0.5835
I01-G03-C01.txt	0.5685	0.3096	0.5656
I01-G03-C02.txt	0.5932	0.3518	0.5911
Average (Mean ± SD)	0.5770 ± 0.0112	0.3301 ± 0.0250	0.5747 ± 0.0112

BERTScore:

Table 4. BERTScore evaluation result

File	Precision	Recall	F1
I01-G01-C01.txt	0.6783	0.6804	0.6797
I01-G01-C02.txt	0.5701	0.5729	0.5721
I01-G01-C03.txt	0.6670	0.6724	0.6701
I01-G02-C01.txt	0.5934	0.6022	0.5983
I01-G02-C02.txt	0.5845	0.5917	0.5886
I01-G03-C01.txt	0.5739	0.5807	0.5779
I01-G03-C02.txt	0.5438	0.5453	0.5451
Average (Mean ± SD)	0.6016 ± 0.0486	0.6065 ± 0.0457	0.6045 ± 0.0453

5.1.2 Interpretation

A detailed analysis of the quantitative results shows that the average performance of the fine-tuned open-source LLaMA 3.1–8B model on seven Finnish conversational datasets is as follows:

Table 5. Average result from Evaluation A

Metric		**Result (Mean ± SD)**
BLEU		0.1342 ± 0.0176
ROUGE	ROUGE-1	0.5770 ± 0.0112
	ROUGE-2	0.3301 ± 0.0250
	ROUGE-L	0.5747 ± 0.0112
BERTScore	Precision	0.6016 ± 0.0486
	Recall	0.6065 ± 0.0457
	F1	0.6045 ± 0.0453

The BLEU score of 0.134 is low, and that's normal for open-domain conversations since the scope of valid responses is broad, and token overlap between generated and reference texts is often limited [41]. It's worth noting that BLEU gives lower scores if words are changed too much, so texts with a lot of paraphrases (which might happen with smaller, varied training sets) might score lower even if it's right [38]. With an observed ROUGE-L score of 0.574, evaluation indicates moderate success in retaining sentence-level context and capture longer common subsequences between the generated text and reference texts of natural Finnish conversation. The BERTScore F1 of 0.604 further suggests moderate semantic similarity, reflecting an ability to capture good understanding of the conversation despite the words used are different [41].

5.1.3 Comparative Analysis

When comparing these metrics to benchmarks presented in section 4.5.2 and domain-specific Language Models for Finnish, it is clear that the open-source

LLaMA 3.1–8B model exhibits both strengths and limitations. We have used K-fold cross-validation, which is usually helpful for generalization, but it might have led to inconsistent training due to limited data in each fold. When there's not enough data, language models often struggle to learn meaningful patterns. Also, the differences in the conversation content and length in each fold could have led to unstable results. Prior research shows that bigger, clearer datasets tend to produce more reliable summarization models [40]. Comparisons with studies on multilingual Large Language Models fine-tuned on low-resource languages show that a BLEU score below 0.15 may be expected due to diverse conversational variations [42]. In the broader context of Icelandic, Basque, or other low-resource language models [43], our outcome indicates that there are similar challenges in Finnish such as having limited lexical overlap and high linguistic variability.

Therefore, the moderate ROUGE-L and BERTScore values observed in our evaluation show that fine-tuning open-source LLMs on Finnish conversational data can give competitive semantic and structural outcomes. However, the limitation of lexical overlap remains.

5.1.4 Result Analysis

BLEU Score:

Table 6. BLEU Score distribution by N-Gram

File	**1-grams**	**2-grams**	**3-grams**	**4-grams**
I01-G01-C01.txt	0.4345	0.2811	0.1723	0.0997
I01-G01-C02.txt	0.3890	0.2533	0.1629	0.1090
I01-G01-C03.txt	0.3663	0.2054	0.1217	0.0720
I01-G02-C01.txt	0.3950	0.2467	0.1481	0.0890
I01-G02-C02.txt	0.4146	0.2691	0.1742	0.1167
I01-G03-C01.txt	0.3940	0.2263	0.1299	0.0646
I01-G03-C02.txt	0.3897	0.2392	0.1482	0.0878
Average (Mean ± SD)	0.3976 ± 0.0200	0.2459 ± 0.0236	0.1510 ± 0.0187	0.0913 ± 0.0175

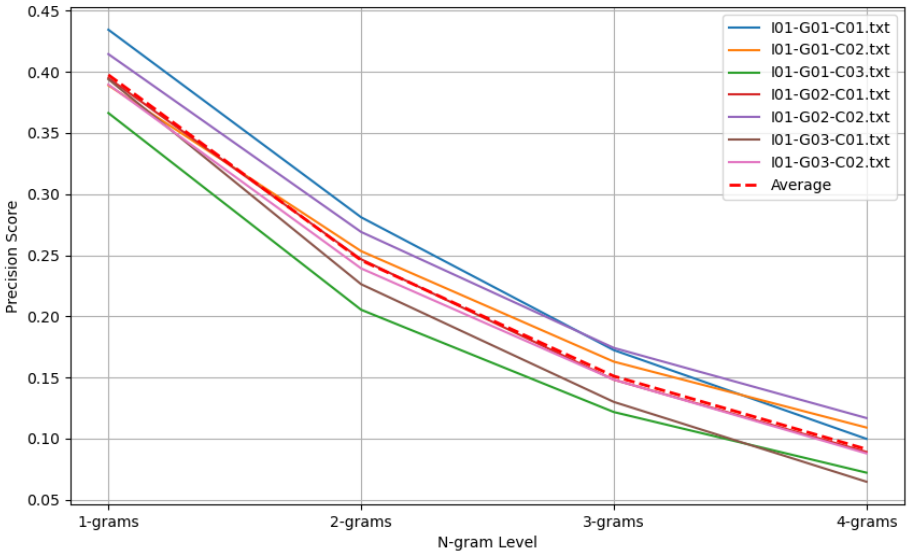

Figure 1. BLEU Score N-gram Precision Trends Across Model Outputs

Detailed Analysis:

1-grams (Unigrams): These scores show a decent level of accuracy which indicates how well individual words in the generated conversation match the ground truth. Scores around 0.4 mean that roughly 40% of the words in the generated text are the same as those in the actual text.

2-grams (Bigrams): The bigram scores show how well pairs of words in the generated notes match the ground truth. Scores shows that about 25% of the word pairs in the generated notes match the word pairs in the ground truth indicating that the generated text have some issues with the correct ordering and combination of words.

3-grams (Trigrams): The trigram scores show closely groups of three words in the generated text match the ground truth text. About 15% of the word sequences match which shows that the generated notes have more significant issues with longer sequences of words, indicating problems with fluency and coherence.

4-grams: The 4-gram scores show how well sequences of four words match. About 10% of the word sequences match. This indicates that the generated notes struggle with longer sequences, which might affect the overall readability.

BERT Score:

Table 7. Contextual similarity of BERTScore from Evaluation A

File	**Contextual similarity**
I01-G01-C02.txt	0.9590
I01-G02-C01.txt	0.9506
I01-G02-C02.txt	0.9451
I01-G03-C01.txt	0.9477
I01-G03-C02.txt	0.9459
Average (Mean ± SD)	0.9497 ± 0.0050

Contextual Similarity provides a high-level measure of semantic alignment without considering the exact token matching. The contextual similarity values ranging from 0.9451 to 0.9590, indicate a strong semantic alignment between the generated and ground truth clinical conversation at the embedding level. This means that even though the wording or structure might change, the overall meaning and context of the texts are well-preserved.

5.2 Teammate's Evaluation (Evaluation B)

5.2.1 Evaluation Tools & Metrics

The evaluation process was conducted using Python, using the following libraries:

- BLEU is calculated using the sacrebleu library [44].
- ROUGE-L is calculated with the rouge-score package [45].
- BERTScore is calculated via the bert-score package [46].

5.2.2 Results

BLEU Score:

Table 8. BLEU score evaluation result

File	**Result**
I01-G01-C01.txt	0.0739
I01-G01-C02.txt	0.1580
I01-G01-C03.txt	0.1154
I01-G02-C01.txt	0.1330
I01-G02-C02.txt	0.1247
I01-G03-C01.txt	0.1151
I01-G03-C02.txt	0.1295
Average (Mean ± SD)	0.1214 ± 0.0235

ROUGE-L Scores:

Table 9. ROUGE-L score evaluation result

File	**Result**
I01-G01-C01.txt	0.3156
I01-G01-C02.txt	0.5151
I01-G01-C03.txt	0.5759
I01-G02-C01.txt	0.5829
I01-G02-C02.txt	0.4558
I01-G03-C01.txt	0.5447
I01-G03-C02.txt	0.4974
Average (Mean ± SD)	0.4982 ± 0.0853

BERTScore (F1):

Table 10. BERTScore evaluation result

File	**Result**
I01-G01-C01.txt	0.7642
I01-G01-C02.txt	0.8384
I01-G01-C03.txt	0.8621
I01-G02-C01.txt	0.8472
I01-G02-C02.txt	0.7905
I01-G03-C01.txt	0.8390
I01-G03-C02.txt	0.8197
Average (Mean ± SD)	0.8230 ± 0.0319

5.2.3 Interpretation

Table 11. Average result from Evaluation B

Metric	**Average Result (Mean ± SD)**
BLEU	0.1214 ± 0.0235
ROUGE-L	0.4982 ± 0.0853
BERTScore F1	0.8230 ± 0.0319

The results presented above show a lower BLEU score, slightly higher ROUGE-L and significantly higher BERTScore than the author's evaluation. A detailed comparison is presented in the next section.

5.3 Comparative Analysis of Evaluation A and B

5.3.1 Metric Comparison

Below table provides a side-by-side comparison of Evaluation A and Evaluation B across metrics, highlighting key differences in BLEU, ROUGE, and BERTScore performance.

Table 12. Metric Comparison Evaluation A v/s B

Metric		Evaluation A (Average)	Evaluation B (Average)	Difference
BLEU		0.1342	0.1214	0.0128
ROUGE	ROUGE-1	0.5770	-	-
	ROUGE-2	0.3301	-	-
	ROUGE-L	0.5747	0.4982	0.0765
BERTScore	Precision	0.6016	-	-
	Recall	0.6065	-	-
	F1	0.6045	0.8230	-0.2185

The above comparison highlights key differences in how Evaluation A and Evaluation B assess model performance across various metrics. Evaluation A shows higher scores for BLEU and ROUGE-L. On the other hand, BERTScore shows a bit different scenario: while Evaluation A shows decent precision and recall, Evaluation B has a much higher F1 score, indicating a difference in semantic similarity evaluation.

Below bar chart provide a quick visual reference for understanding the performance comparison between Evaluation A and Evaluation B across three key metrics: BLEU, ROUGE-L, and BERTScore F1. Each pair of bars represents the respective scores from the two evaluations, while the percentage annotations above indicate the relative change from Evaluation A to Evaluation B.

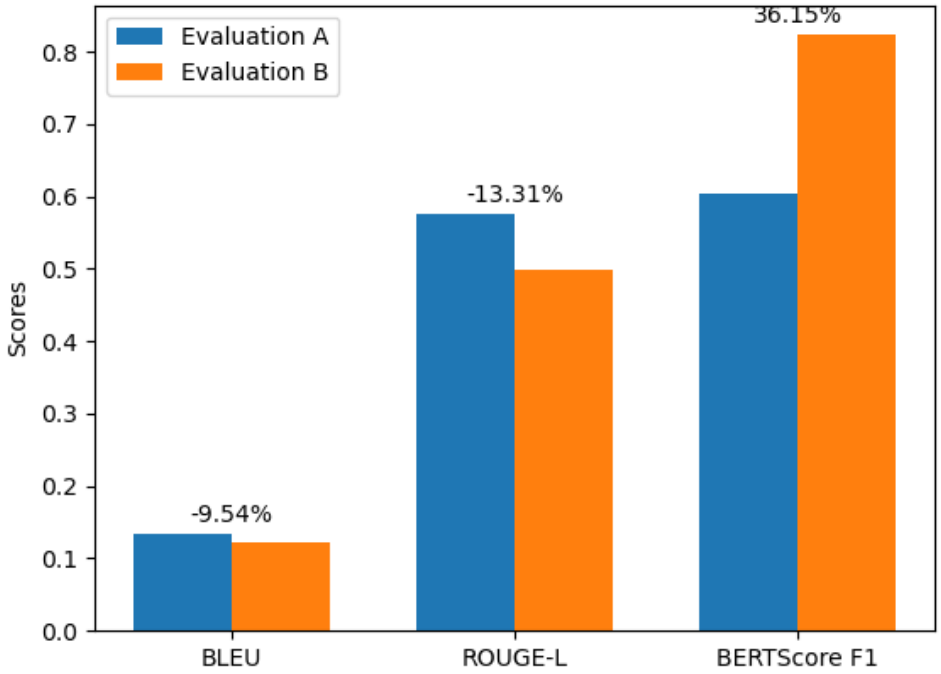

Figure 2. Metric Score Comparison: Evaluation A vs Evaluation B

5.3.2 Interpretation of Differences

BLEU:

Differences in implementation details, particularly the calculation level, method of text tokenization and smoothing are the primary reasons why Evaluation A and Evaluation B produced different BLEU scores when evaluating the same set of generated and reference ground truth texts.

Table 13. Implementation difference of BLEU evaluation

Feature/ Criterion	Evaluation A	Evaluation B
Library	nltk.translate.bleu_score	sacrebleu
Tokenization	Explicit word_tokenize from NLTK	Assumes pre-tokenized input or uses internal tokenizer
Calculation Level	Per-segment level	Corpus-level

Text Preprocessing (Tokenization)	nltk.tokenize.word_tokenize	SacreBLEU's standardized tokenization/normalization
Smoothing Method	NLTK's SmoothingFunction().method4	SacreBLEU's default smoothing

A key difference is at the level at which the BLEU score was calculated. The original BLEU metric was designed for aggregate counts over an entire test corpus (system-level evaluation) level rather than individual sentences (sentence-level evaluation). Evaluation B followed this standard corpus-level calculation whether Evaluation A calculated a separate BLEU score for each pair of hypothesis and reference files within its loop. So, calculating BLEU scores independently for smaller units (like individual files or sentences) and then averaging them individually is not mathematically equivalent to computing a single BLEU score over the entire corpus. Thus, Evaluation A scores can be less reliable [10] but later study shows that sentence-level scores can still be useful indicators [34].

Another key factor for the difference is how the tokenization is done. The BLEU score is heavily dependent on counting matching sequences of tokens (n-grams) between the generated and reference texts. Evaluation B (using sacrebleu) used standardized tokenization and normalization process and Evaluation A used splited text into tokens (using nltk.tokenize.word_tokenize) which directly affected the n-grams that are counted and matched, leading to different precision scores.

Furthermore, Standard BLEU can result in a zero score if any n-gram (normally up to 4-grams) precision is zero which is rare for corpus-level evaluation on reasonable data sizes but it's common for sentence-level evaluation or small datasets. If a system produces a text that doesn’t match any n-grams of a certain length, the precision score for that n-gram length will be zero. Taking the logarithm of zero precision makes the term negative infinity, which means the final geometric mean (and thus, BLEU score) of zero [38]. To get around this problem, smoothing methods are used to change those zero counts to tiny non-

zero numbers or apply other adjustments. This helps prevent a zero score, especially in shorter texts or higher n-grams where zero counts are common. Evaluation A (using nltk.translate.bleu_score) used SmoothingFunction().method4 to address this whether Evaluation B (using sacrebleu) used its own default smoothing technique which is designed to align with common practices for corpus-level evaluation.

Table 14. Implementation difference of ROUGE evaluation

Feature/ Criterion	Evaluation A	Evaluation B
Library Used	rouge_score	rouge_score (same)
ROUGE Variants	Computes ROUGE-1, ROUGE-2, and ROUGE-L	Computes only ROUGE-L
Text Normalization	Extensive: lowercasing, drug/abbreviation/temporal normalization, stopword removal	None shown; uses raw references and predictions
Stopword Handling	Removes Finnish stopwords	Not applied

There are a few important differences in how ROUGE scores are calculated, and these can affect evaluation result. ROUGE comes in several variants: ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap) and ROUGE-L (longest common subsequence) which measure overlap at different granularity levels. Evaluation A computed multiple variants provide a more comprehensive view of the similarity between the generated text and the reference than relying solely on ROUGE-L (Evaluation B).

ROUGE relies on exact matches of n-grams (for ROUGE-1, ROUGE-2) or character sequences (for ROUGE-L's longest common subsequence). Different ways of capitalizing, using punctuation, using abbreviations, or number/date formats that have the same meaning can prevent these exact matches and artificially lower the score. While studies discussing ROUGE directly didn’t mention specific normalization but the principle of text standardization is

discussed about other metrics affected by surface form (like WER) [32]. Normalization steps (drug/abbreviation/temporal) applied in Evaluation A aimed to make ROUGE scores less sensitive to these surface variations and more reflective of semantic overlap which is relevant for domain-specific text like clinical conversations where abbreviations and specific formats are common. On the other hand, Evaluation B's usage of raw text can sometimes overlook valid content because of small formatting differences.

Another factor is the treatment of stopwords which are common words (like "minä", "mutta", "kyllä") that often carry less semantic weight, removing them is a common practice in NLP tasks to focus more on important words [8]. While studies don't explicitly state whether ROUGE should remove stopwords, doing so would mean the ROUGE score primarily measures the overlap of the more meaningful words in the text. This approach, used in Evaluation A, which potentially led to an evaluation metric that aligns more closely with human judgments about the similarity of core content. Evaluation B includes stopwords, which means a lower score compared to Evaluation A influenced by the presence or absence of these words.

Table 15. Implementation difference of BERTScore evaluation

Feature/ Criterion	**Evaluation A**	**Evaluation B**
Library	BERTScorer class	bert_score.score
Model	Explicitly bert-base-multilingual-cased	Default (xlm-roberta-base)
Preprocessing	Full-document, newline/space normalization	Line-by-line, no cleaning
Layers Used	Custom: num_layers=8	Default (last layer)
Baseline Rescaling	rescale_with_baseline=True	File-level (single pair)

Differences in implementation details like the embeddings, layers at which it has been evaluated, tokenizers and rescale are the primary reasons why Evaluation A and Evaluation B produced different BERTScore.

BERTScore relies on extracting contextual embeddings from pre-trained Transformer models like BERT or RoBERTa which are trained on different data and learn distinct vector representations for words based on their context. BERTScore mainly relies on calculating cosine similarity between embeddings. So, if the embeddings change, the similarity scores and BERTScore values will change too. The BERTScore paper itself evaluates using various models, highlighting that performance can differ. Also, suggests using the multilingual BERT model (bert-base-multilingual-cased) as a suitable choice [40] which has been used in Evaluation A.

Another factor is layer, transformer models create embeddings in every layer. By default, the BERTScore library uses the last layer, but it allows specifying alternative layers (like layer 8 in Evaluation A). Research shows that the layers in the middle tend to capture linguistic details better for meaning-related tasks compared to the last layer and also provide recommended layers for various models, including layer 9 for bert-base-multilingual-cased. Evaluation A used layer 8, which is very close to this supported recommendation while Evaluation B used the default last layer, which the paper suggests is typically less informative for semantic tasks compared to intermediate layers [40].

Furthermore, BERT model has their own internal subword tokenizers example: WordPiece that are optimized for the model's pre-training. Using a tokenizer before feeding text to BERT can potentially interfere with the model's intended tokenization process, which might negatively impact embedding quality.

The Baseline Rescaling (rescale_with_baseline=True) is an optional step applied in Evaluation A to the final BERTScore to map it to a more intuitive range (normally 0-1) for improving readability but it does not affect the scoring.

6 Discussions and Conclusions

This chapter interprets and explains the results obtained from evaluating the fine-tuned LLaMA 3.1(8B) model across two separate evaluations, using common Natural Language Generation (NLG) metrics BLEU, ROUGE and BERTScore and compares them with a peer tested on the same dataset. It also discusses their significance, and relates them to the broader implications for future development of Finnish-language LLMs for clinical documentation. It further look into the quality of the training data, points out a issue with the instructional sentences in the ground truth transcript and considers how this affects model performance and evaluation results.

6.1 Summary of Key Findings

The current evaluation and analysis gave several important insights for the further development of digital scribe systems in low-resource languages like Finnish. The evaluation used the 7-fold cross-validation approach, using a comprehensive set of metrics: BLEU, ROUGE (1, 2, and L) and BERTScore (precision, recall, and F1). This detailed analysis pointed out some clinically important issues, including lexical mismatch, which was not fully captured by aggregate metrics alone. My teammate’s evaluation focused on BERTScore F1, BLEU, and ROUGE-L showed higher scores, but it didn’t reveal the weaknesses found in this analysis.

A review of the training data found that there were descriptions or instructions like “Potilas kävelee magneettitilan pukuhuonetta kohden. (In English: The patient walks towards the dressing room of the magnet room.)” in the ground truth documents. These types of phrases aren't clinical and might have introduced noise into the training process, possibly impacting both model learning and evaluation.

6.2 Dataset Content Observation and Its Potential Impact

Looking more closely at the data, it has been found that some ground truth transcripts included instructional or descriptive sentences that were not part of actual spoken language but appeared to guide the transcription process. Such as ”Potilas kävelee magneettitilan pukuhuonetta kohden (In English: The patient walks towards the dressing room of the magnet room).” or ”Potilas vaihtaa vaatteet ja koputtaa oven (In English: The patient changes clothes and knocks on the door)”.

During training, the model learns from the patterns present in the ground truth so it is likely to include similar details in its generated output. This is consistent with the principle that machine learning models learn to copy the style and content from the datasets they were trained on. These instructions are not generated by the model because they are not spoken words, but they appear in the reference text, which can affect precision-focused metrics such as ROUGE by lowering the score. Studies suggest that data quality issues may have contributed to the lower-than-expected performance observed in this research [4].

6.3 Limitations

This research is subject to several limitations, primarily related to the nature of clinical data and evaluation methodologies. As discussed, the presence of instructional texts in the ground truth datasets might have impacted model learning and evaluation, although it was not measured how much it affected. ROUGE metrics have been potentially biased though BERTScore attempts to mitigate this by focusing on semantic similarity. There’s no automated metric that perfectly captures human judgment of clinical text quality, which includes factors like clinical accuracy, relevance, and conciseness.

Getting access to data in healthcare is often restricted due to privacy concerns, limiting the diversity and scale of training and evaluation datasets, especially for specific languages and tasks. The process of creating high-quality, annotated

gold standards for clinical text is also time-consuming, costly and can also vary from one person to another, especially when the language is complex or descriptive. This evaluation in a pilot setting, using simulated conversations on a limited dataset, may not fully capture different challenges and variations encountered in real-world clinical practice.

6.4 Implications and Future Directions

Findings from this research have some key takeaways for both clinical practice and research. In clinical settings, using detailed evaluation metrics can help ensure that NLP models generate clinically relevant and accurate results, which can reduce the risk of errors that could harm patients. Cleaning the dataset carefully is key to reducing errors and improving model robustness.

Based on the findings and limitations of this study, a few future research areas are suggested. First, refining dataset guidelines to standardize the representation of common actions or notes in the clinical text could help make the ground truth dataset more consistent and lower any differences in the results. Exploring alternative evaluation metrics tailored for clinical texts that evaluate more than lexical or semantic matches to assess clinical accuracy, completeness, and relevance is also crucial.

Domain-specific fine-tuning on larger, more diverse Finnish clinical datasets would likely enhance model performance [19]. Since clinical documentation can vary, investigating methods robust to paraphrasing and writing style differences is important.

Ultimately, it's important to evaluate the clinical utility and usability of NLP systems in real-world Finnish clinical settings to determine their actual impact on documentation speed, clinician workflow, and patient care. This would involve future studies that assess not just the technical accuracy of the output generated by the model but also how doctors use the system and how it impacts the overall clinical documentation process and patient outcomes. Bringing in "doctor-in-the-

loop" approaches, where the AI gives a draft or extracts important information for human review and editing, could be a practical good approach before going all the way to full automation.

7 Summary

This study explored the feasibility of integrating a fine-tuned open-source Large Language Model (LLM) into a Finnish-language digital scribe system for clinical use. By focusing on evaluation metrics aligned with clinical standards, the research aimed to determine whether training digital scribes with specialty-specific data improves performance compared to general models to explore how well such models could transcribe Finnish clinical dialogues.

This research lays the groundwork for the future development of language-adapted digital scribes in healthcare in Finnish. This study shows how Natural Language Processing (NLP) can help in processing Finnish clinical text, but the domain-specific language used in the field and the difficulty of getting consistent ground truth data highlight the complexity of the task. Continued research focusing on robust methodologies, better evaluation techniques and real-world clinical validation is necessary to fully realize the benefits of NLP for improving clinical documentation and access to information.

References

1 DigiFinland. Artificial intelligence pilot projects in the health and social services sector enhance the work of professionals and improve customer service and quality of care. <https://digifinland.fi/sote-sektorin-tekoalyn-kokeiluprojekteilla-tehostetaan-ammattilaisten-tyota-seka-parannetaan-asiakaspalvelua-ja-hoidon-laatua/>. Accessed 21 April 2025.

2 Apotti. The development of AI tools in information systems in the health and social services sector is accelerating. <https://www.apotti.fi/ai-tyokalujen-kehitys-sote-alan-tietojarjestelmissa-kiihtyy/>. Accessed 21 April 2025.

3 Coiera, E., Kocaballi, B., Halamka, J. et al. The digital scribe. npj Digital Med 1, 58 (2018). https://doi.org/10.1038/s41746-018-0066-9.

4 van Buchem, Marieke & Boosman, Hileen & Bauer, Martijn & Kant, Ilse & Cammel, Simone & Steyerberg, Ewout. (2021). The digital scribe in clinical practice: a scoping review and research agenda. npj Digital Medicine. 4. 10.1038/s41746-021-00432-5. https://www.nature.com/articles/s41746-021-00432-5.

5 Edwards, S. T., Neri, P. M., Volk, L. A., Schiff, G. D., & Bates, D. W. (2014). Association of note quality and quality of care: a cross-sectional study. BMJ quality & safety, 23(5), 406–413. https://doi.org/10.1136/bmjqs-2013-002194.

6 Quiroz, J.C., Laranjo, L., Kocaballi, A.B. et al. Challenges of developing a digital scribe to reduce clinical documentation burden. npj Digit. Med. 2, 114 (2019). https://doi.org/10.1038/s41746-019-0190-1

7 Névéol, A., Dalianis, H., Velupillai, S. et al. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semant 9, 12 (2018). https://doi.org/10.1186/s13326-018-0179-8.

8 Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics, 128–144. https://doi.org/10.1055/s-0038-1638592.

9 Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., El-Sappagh, S., Islam, M. S., & Islam, S. M. R. (2020). Medical Diagnostic Systems Using Artificial intelligence (AI) Algorithms: Principles and Perspectives. IEEE Access, 8, 228049–228069. https://doi.org/10.1109/access.2020.3042273.

10 Huang, Kexin & Li, Jaan & Ranganath, Rajesh. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. https://doi.org/10.48550/arXiv.1904.05342.

11 Yuqi Si, Jingqi Wang, Hua Xu, Kirk Roberts, Enhancing clinical concept extraction with contextual embeddings, Journal of the American Medical Informatics Association, Volume 26, Issue 11, November 2019, Pages 1297–1304, https://doi.org/10.1093/jamia/ocz096.

12 Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (n.d.). Multilingual is not enough: BERT for Finnish. https://doi.org/10.48550/arXiv.1912.07076.

13 Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans. Comput. Healthcare 3, 1, Article 2 (January 2022), 23 pages. https://doi.org/10.1145/3458754.

14 Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6). https://doi.org/10.1093/bib/bbac409.

15 Xie, Q., Chen, Q., Chen, A., Peng, C., Hu, Y., Lin, F., Peng, X., Huang, J., Zhang, J., Keloth, V., Zhou, X., He, H., Ohno-Machado, L., Wu, Y., Xu, H., & Bian, J. (2024). Me-LLaMA: Foundation Large Language Models for Medical Applications. Research square, rs.3.rs-4240043. https://doi.org/10.21203/rs.3.rs-4240043/v1.

16 Kheddar, H., Hemis, M., & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. Information Fusion, Elsevier. https://doi.org/10.1016/j.inffus.2024.102422.

17 HFMA. Strategies for success: Tackling common clinical documentation integrity challenges head-on. <https://www.hfma.org/revenue-cycle/strategies-for-success-tackling-common-clinical-documentation-integrity-challenges-head-on>.

18 Zhou, H., Liu, F., Gu, B., et al. (2023). A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. arXiv preprint arXiv:2311.05112. https://doi.org/10.48550/arXiv.2311.05112.

19 Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. Journal of biomedical informatics, 77, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011.

20 Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for Natural Language Processing in Low-Resource Scenarios. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. https://doi.org/10.18653/v1/2021.naacl-main.201.

21 Luoma, J., Chang, L.-H., Ginter, F., & Pyysalo, S. (2021). Fine-grained named entity annotation for Finnish. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021) (pp. 135–144). Linköping University Electronic Press. https://aclanthology.org/2021.nodalida-main.14.

22 Sasseville, Maxime & Yousefi, Farzaneh & Ouellet, Steven & Bergeron, Frédéric & LeBlanc, Annie. (2025). Impacts of AI Scribes on Clinical Outcomes, Efficiency, and Documentation A rapid review Prepared for Canada Health Infoway. http://dx.doi.org/10.13140/RG.2.2.20047.19364.

23 Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35.

24 Névéol, A., Dalianis, H., Velupillai, S. et al. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semant 9, 12 (2018). https://doi.org/10.1186/s13326-018-0179-8.

25 Keenious. <https://keenious.com>. Accessed 04th June 2025.

26 Fleming, S. L., Lozano, A., Haberkorn, W. J., Jindal, J. A., Reis, E., Thapa, R., Blankemeier, L., Genkins, J. Z., Steinberg, E., Nayak, A., Patel, B., Chiang, C.-C., Callahan, A., Huo, Z., Gatidis, S., Adams, S., Fayanju, O., Shah, S. J., Savage, T., Goh, E., Chaudhari, A. S., Aghaeepour, N., Sharp, C., Pfeffer, M. A., Liang, P., Chen, J. H., Morse, K. E., Brunskill, E. P., Fries, J. A., & Shah, N. H. (2024). MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22021-22030. https://doi.org/10.1609/aaai.v38i20.30205.

27 Microsoft Dragon Copilot. <https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot>. Accessed 04th June 2025.

28 Amazon HealthScribe. <https://aws.amazon.com/healthscribe>. Accessed 04th June 2025.

29 Notable. <https://www.notablehealth.com>. Accessed 04th June 2025.

30 DeepScribe. <https://www.deepscribe.ai>. Accessed 04th June 2025.

31 Augmedix. <https://www.augmedix.com>. Accessed 04th June 2025.

32 Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv. https://arxiv.org/abs/2212.04356.

33 Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of biomedical informatics, 73, 14–29. https://doi.org/10.1016/j.jbi.2017.07.012.

34 Banerjee, S., & Lavie, A. (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 MT and/or Summarization, 65–72. https://aclanthology.org/W05-0909.

35 Autoscriber. Autoscriber for Healthcare Professionals. <https://autoscriber.com/healthcare-professionals>. Accessed 04th June 2025.

36 Google NotebookLM. <https://notebooklm.google>. Accessed 04th June 2025.

37 Chowdhury, Mohammed Nowshad Ruhani, “Using Open Source LLM Model for Medical Transcription”, Master of Engineering thesis, Metropolia University of Applied Sciences, (2025). https://urn.fi/URN:NBN:fi:amk-2025052716517.

38 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135.

39 Moradi, M., Dashti, M., & Samwald, M. (2020). Summarization of biomedical articles using domain-specific word embeddings and graph ranking. Journal of Biomedical Informatics, 107(103452), 103452. https://doi.org/10.1016/j.jbi.2020.1034522.

40 Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675. https://arxiv.org/abs/1904.09675.

41 Tuuli Laitinen, “Generating Healthcare Reports Using Natural Language Processing – Fine tuning Finnish Language Models”, Master’s Degree Programme in Computer Sciences Thesis, Tampere University (2025). https://urn.fi/URN:NBN:fi:tuni-202503112703.

42 Purason, T., Kuulmets, H., & Fishel, M. (2024). LLMs for extremely Low-Resource Finno-Ugric languages. https://arxiv.org/abs/2410.18902.

43 Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., & Soroa, A. (2024). Latxa: An Open Language Model and Evaluation Suite for Basque. https://arxiv.org/abs/2403.20266.

44 Post, M. (2018). SacreBLEU: A standardised BLEU implementation. [Software]. <https://github.com/mjpost/sacrebleu>. Accessed 04th June 2025.

45 Python Package Index. Python ROUGE: A native python implementation of ROUGE. [Software]. <https://pypi.org/project/rouge-score>. Accessed 04th June 2025.

46 Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT [Software]. <https://github.com/Tiiiger/bert_score>. Accessed 04th June 2025.

Code and Repository

All code, scripts and resources used in this research are stored on the following GitHub repository: https://github.com/sakluk/digital-scribes. The repository contains ground truth and model generated datasets, pre-processing scripts and evaluation scripts. Setup, dependency, and usage are described in the README file on the repository.