Mohammed Nowshad Ruhani Chowdhury

Using Open Source LLM Model for Medical Transcription

Metropolia University of Applied Sciences

Master of Engineering

Information Technology

Master's Thesis

27 May 2025

PREFACE

This thesis is both the most difficult and satisfying part of my academic career. Working with large language models and clinical data in Finnish posed novel technical and linguistic challenges particularly considering the paucity of domain-specific datasets and native Finnish support for the LLaMA 3.1–8B model. In spite of the challenges, the process was a significant learning experience regarding deep learning, healthcare data, and the practical aspects of AI in low-resource language environments.

I am truly grateful to my thesis advisor for their guidance, to the Digital Medical Scribe project team for shared insights, and to CSC – IT Center for Science for providing access to computational resources needed to carry out this study. Special thanks are extended to my student colleagues who participated in the construction of the dataset, and to my family and friends for keeping me centered and motivated throughout despite my laptop seeming to have more visitors than they did.

Helsinki, 27 May, 2025
Mohammed Nowshad Ruhani Chowdhury

Abstract

Author:	Mohammed Nowshad Ruhani Chowdhury
Title:	Using Open Source LLM Model for Medical Transcription
Number of Pages:	38 pages + 2 appendices
Date:	27 May 2025
Degree:	Master of Engineering
Degree Programme:	Information Technology
Professional Major:	Medical Technology
Supervisors:	Sakari Lukkarinen, Senior Lecturer Mikael Soini, Principal Lecturer

---

In modern healthcare, clinical documentation is paramount for patient safety, accurate diagnoses, and continuity of care. However, physician burnout has been caused by the increasing overhead of electronic health record (EHR) systems, which take up less time for real human interaction. In less-resourced languages such as Finnish, in which natural language processing (NLP) tools are only beginning to emerge, this is an even bigger challenge. This thesis investigates the fine-tuning of the open-source LLaMA 3.1–8B language model on simulated Finnish clinical conversations that is, transcribed clinical dialogues created by Metropolia UAS students. The aim is to verify if a domain-aligned large language model (LLM) is able to reliably translate spoken Finnish medical discourse into formal clinical reports. With 7-fold cross-validation, the fine-tuned model achieved a BLEU score of 0.1242, ROUGE-L score of 0.4982, and BERTScore F1 score of 0.8373, showing satisfactory semantic performance using a small dataset and scalability of privacy-oriented NLP tools in Finnish medical environments.

Keywords: Medical scribes, Open source LLM, LLaMA 3

Contents

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
BART	Bidirectional & Autoregressive Transformer
BERTScore	Bidirectional Encoder Representations from Transformers Score
BLEU	Bilingual Evaluation Understudy
BART	Bidirectional & Autoregressive Transformer
CPU	Central Processing Unit
CSC	Center for Science Ltd
CT	Computed Tomography
CUDA	Compute Unified Device Architecture
EHR	Electronic Health Records
FHIR	Fast Healthcare Interoperability Resources
GB	Gigabyte
GDPR	General Data Protection Regulation
GEMINI	Google's Large Language Model
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
HBM	High Bandwidth Memory
I/O	Input/Output
ICD-10	International Classification of Diseases, Tenth Revision
JSON	JavaScript Object Notation
LCS	Longest Common Subsequence
LLM	Large Language Model
LLaMA	Large Language Model Meta AI
LoRA	Low-Rank Adaptation of Large Language Models
ML	Machine Learning
MP3	MPEG Audio Layer 3
NLP	Natural Language Processing
NVIDIA	Nvidia Corporation
OPUS	Publikationsverbund der Universität Stuttgart

Puhti	Puhti Supercomputer
RAM	Random Access Memory
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SLURM	Simple Linux Utility for Resource Management
SNOMED	Systematized Nomenclature of Medicine
TXT	Text File Document
UAS	University of Applied Sciences
VRAM	Video Random Access Memory

1 Introduction

In modern healthcare system, clinical documentation is absolutely essential for patient safety, correct diagnosis, support of medical billing, and preservation of legal documents. The increasing complexity of patient interactions and healthcare delivery systems has resulted in a greater burden of documentation for clinicians, which frequently causes physician burnout and a decrease patient interactions time and structural clinical documentation helps doctors and nurses to give appropriate treatment to patient and follow up. Traditionally, this responsibility has fallen heavily on physicians and nurses, contributing to widespread documentation burden and professional burnout. Physicians frequently have to split their time between talking to patients and entering information into electronic health records (EHRs), which can compromise both the depth of patient communication and the completeness of clinical documentation. This cognitive stress lowers the quality of care provided while also raising the risk of burnout and decreasing work satisfaction. On the other hand, patients may have recurrent questions, disjointed treatment pathways, and impeded continuity between experts as a result of inconsistent or poorly organised data, especially in complicated or chronic care cases [1].

Beside that, all of the stakeholders are becoming increasingly concerned about the burden of clinical paperwork since it has a direct effect on the calibre of interactions during consultations. These challenges underscore the critical need for structured clinical documentation, where patient information is organized into standardised, human-readable formats [2]. Structured documentation facilitates interoperability across health systems, supports evidence-based clinical decision-making, and enables secondary uses such as population health analytics and health services research. It makes ensuring that important clinical observations are regularly, accurately, and retrievably recorded, including diagnoses, prescriptions, allergies, and instructions for follow-up, thereby improving care outcomes and long-term patient safety.

The incorporation of medical scribes, whether human or digital, has surfaced as an effective strategy to alleviate administrative burdens and enhance workflow efficiency [3]. Human medical scribes are trained professionals who help doctors by recording patient interactions in real time. This frees up clinicians to spend more time interacting with patients rather than entering data. Typically these professionals work alongside the physician during consultations, scribes enter information directly into electronic health records (EHRs), including history, symptoms, physical exam findings, and care plans [4]. It has been demonstrated that their presence increases physician satisfaction, streamlines workflow, and cuts down on paperwork time. However, in addition to these advantages, using human scribes is problematic in terms of scalability, presents the threat of inconsistent quality of documentation, and exacerbates issues with cost, privacy of data, and consistency of training.

These limitations have given rise to automation in the form of artificial intelligence (AI)-based software, specifically in automatic speech recognition (ASR) and natural language processing (NLP) [5]. Such technologies work as a digital medical scribe by logically organizing clinical voice transcriptions into usable documentation in real-time. GPT-3 [6], GPT-NeoX [7] , LLaMA [8], and Falcon [9] are among the large language models that have recently achieved outstanding advancements and reported strong performance on a broad spectrum of NLP tasks. Besides that, healthcare systems looking to create unique, privacy-compliant clinical documentation solutions without depending on proprietary APIs have a strong chance with open-source LLMs in particular. These models may be refined using domain-specific data, such EHR corpora or medical notes, to accurately parse clinical language into structured forms like systematized nomenclature of medicine (SNOMED) CT representations or fast healthcare interoperability resources (FHIR).

However, in non-English languages, especially those with rich morphological traits like Finnish, the construction of such systems becomes much more challenging [10]. Finnish (Suomi) is a morphologically agglutinative language, indicating that a single word can represent several grammatical characteristics,

leading to a highly sparse vocabulary and complex parsing challenges. In the medical domain, this complexity is compounded by specialized terminology, code-switching (e.g., Latin, English loanwords), and the need for high accuracy due to patient safety concerns.

Open-source large language models (LLMs) are particularly appealing for healthcare systems subject to stringent privacy regulations like the GDPR since they provide transparency, customisability, and the possibility of local implementation, in contrast to proprietary models. Importantly, these models may be refined or fine-tuned on domain-specific corpora in any language, including Finnish, which enables them to capture subtle morphology, syntactic structure, and medical terminology strange to the language [11]. This adaptability serves as the basis for the current study, which attempts to show that an open-source, Finnish-language AI writing system can significantly benefit health professionals by reducing documentation burden and enabling more patient-centered consultations, improving care continuity and access to Finnish-language patients with language-specific clinical records [12]. Consequently, this thesis proposes an advance approach: fine-tune open-source LLMs to accurately the transformation of transcribed clinical speech into structured documentation, specifically for the Finnish language.

2 Literature Review

2.1 Using Keenious for finding related literature

As I progressed through the different stages of writing my thesis, one of the biggest challenges I faced was locating academic sources that were relevant to this research. Instead of conducting keyword searches and manually reviewing irrelevant papers, I used Keenious to streamline the process [13]. By simply pasting sections of my writing such as my initial research objectives and technical descriptions of large language models into the Keenious interface, I received instant suggestions of peer-reviewed papers and scholarly sources that were contextually aligned with my topic. This saved me effort and significantly increased the relevance and quality of the sources to incorporate.

What made Keenious especially useful was its ability to understand the deeper meaning and context behind my writing, rather than just keyword matches. My standard procedure was to compose a paragraph or even a couple of sentences describing a narrative behind a technical aspect of my thesis, such as 'the use of large language models (LLMs) in medical transcription or the challenge of converting unstructured text into structured clinical information' and then pass it on to Keenious to process. From that result, Keenious provided me with a list of recommended academic papers which were contextually relevant. This process is illustrated in detail in Figure 1.

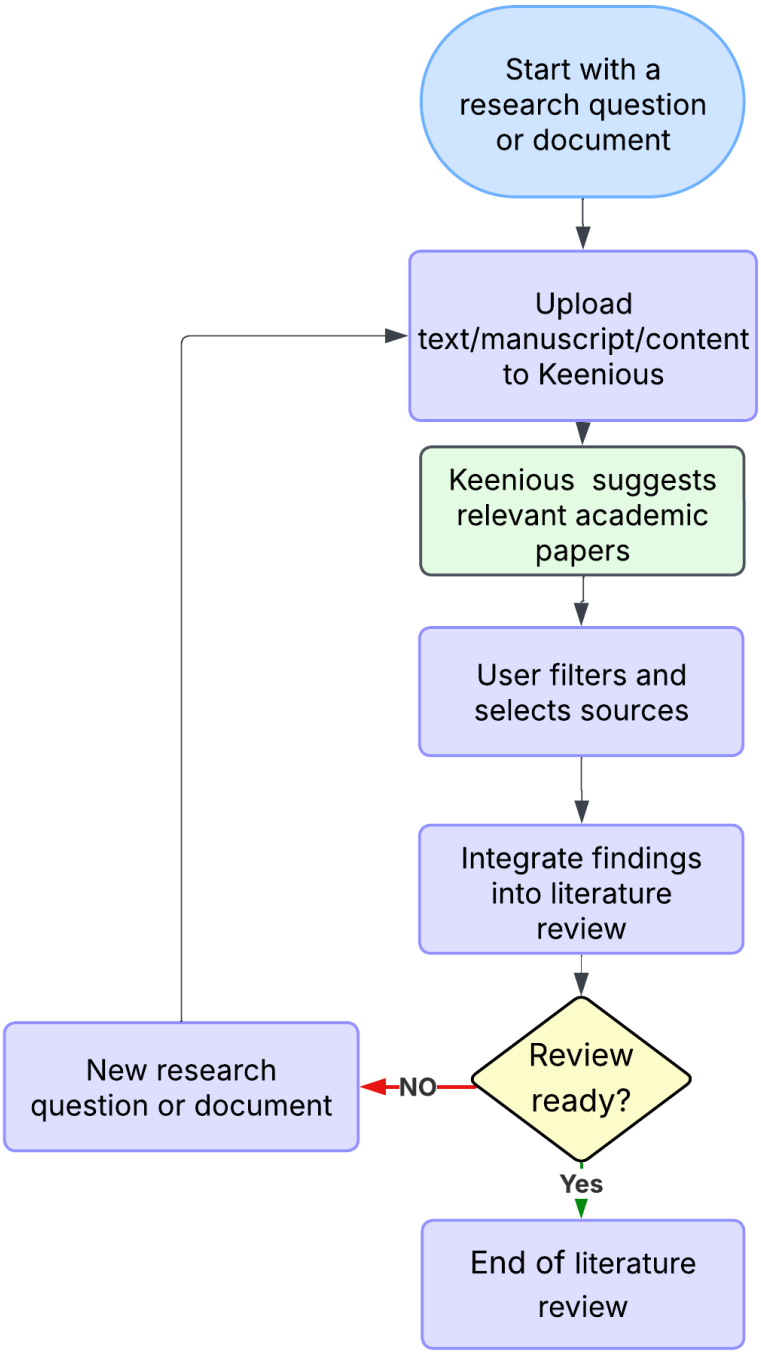

Figure 1: Workflow of Keenious for finding literature review.

For example, when I described how transformer models can be adapted to process spoken medical dictations, Keenious pointed me to “BioBERT: A pre-trained biomedical language representation model for biomedical text mining [14]”, which provided some insight into domain-aware pretraining techniques.

Another example was when I roughed out a part of the paper discussing the limitations of existing NLP tools in healthcare settings, Keenious pointed me to “Publicly Available Clinical BERT Embeddings [15]”, which educated me on the ways contextualized embeddings improve clinical text comprehension. In addition, while conducting research on how LLMs can be used to improve the accuracy of medical transcription, Keenious made me aware of “Recent Advances in End-to-End Automatic Speech Recognition [16]”, which directly supported my argument for the integration of ASR (automatic speech recognition) and structured output modelling. These targeted recommendations allowed me to discover relevant sources quickly, often showing me papers I may have difficulties discovering using traditional search engines alone. This approach greatly enriched my literature review with relevant, cutting-edge articles.

2.2 Exploring the Literature

Healthcare's digital transformation has placed increasing pressure on the efficiency, accuracy, and structure of clinical documentation. As physicians face mounting administrative demands, accurate record-keeping has become both a critical necessity and a major burden [17]. Clinical documentation is essential to quality assurance, billing, diagnosis, treatment continuity, patient safety, and compliance with rules. However, because of the enormous time commitment needed, many doctors are now more burned out and dissatisfied, spending more time documenting than connecting with patients [18].

In response, manual (human) medical scribes were created to help with real-time documentation during clinical visits, therefore reducing this stress [19]. These professionals are trained to enter information like patient history, physicals, clinical impression, and plans of care into the EHRs in conversations dialogue's document, thus allowing physicians to focus on delivery of care. Evidence has shown that having scribes present improves physician productivity, reduces time spent charting, and even patient satisfaction with more direct communication. Gidwanis’ [20] found that doctors working with scribes wrote more quickly and spent significantly less time after clinic hours in a randomized trial. But manual

scribes are problematic too such as training costs, inter-scribe variability, turnover rates, and data privacy and scalability issues in low-resource settings [21].

These limitations have spurred interest in automated or digital scribes powered by AI, Automatic Speech Recognition (ASR), and Natural Language Processing (NLP). ASR systems convert spoken language into text and have matured considerably with the advent of deep learning techniques, which is explained deeply on Li's end-to-end automatic speech recognition [16]. Open-source speech-to-text engines like Kaldi, Mozilla DeepSpeech, and OpenAI Whisper are a few of the intriguing options that Yifan's examination of publicly accessible datasets demonstrated [23]. Yet, achieving clinical-grade performance remains a challenge, particularly in environments with domain-specific vocabulary, background noise, and code-switching between languages [24].

After transcribing, these free-text conversational interactions must be structured in order to be useful for downstream applications such as clinical decision support, analytics, and hospital-hospital interoperability. Open-source LLMs such as GPT-NeoX [25], BLOOM [26], LLaMA [27], and Falcon [28] have been observed to demonstrate state-of-the-art performance in text generation, summarization, and classification tasks. Fine-tuned models on medical corpora can accurately detect important entities like diagnoses, medications, and symptoms even from noisy, unstructured conversational data, which done by Lee and his team. Perhaps more importantly, recent research has demonstrated the capability of LLMs to produce structured output from raw dialogue. Agrawal and team used GPT-3 to convert clinician-patient encounters into Subjective, Objective, Assessment, and Plan (SOAP) shaped documentation [22], while Raza and Nooralahzadeh proposed direct mapping of text into FHIR fields with the help of a Bidirectional & Autoregressive Transformer (BART) based mechanism enabling incorporation into EHR systems [29].

The conversational nature of clinical dialogue brings additional complexity over narrative formal notes. Questions, interjections, negations, and medical shorthand all introduce ambiguity. Yet, models fine-tuned on multi-turn dialogue

training datasets e.g., MedDialog which developed by Liu and team [30] and GEMINI have yielded promising results in structuring conversational inputs. Zhang and his team recently demonstrated that fine-tuned versions of LLaMA outperformed traditional rule-based NLP systems in dialogue-to-ICD-10 code mapping, leveraging the semantic and contextual capabilities of LLMs [31].

While the developments are promising, most work to date remains English-centric, with comparatively little research on multilingual or low-resourced languages such as Finnish. Finnish is a morphologically rich, agglutinative language with free word order and composite words formed frequently. These traits pose severe parsing and tokenization challenges. Moreover, the absence of annotated Finnish clinical corpora complicates NLP development even further. Nevertheless, there are some ground resources to tap into: the Finnish version of BERT, FinBERT developed by Virtanen and his team [32], has worked exceptionally well on sentiment and classification tasks; OPUS-MT provides machine translation support; and the Language Bank of Finland (Kielipankki) has rich linguistic corpora. For ASR, AaltoASR software has also shown promise in Finnish speech-to-text applications, but domain adaptation to medical conversation remains to be carried out [33].

The theoretical basis of this thesis is the principle of transfer learning—the conjecture that large pretrained language models can be retrained or fine-tuned on language-specific and domain-specific data. Previous research has confirmed that LLMs trained on biomedical corpora, i.e., BioBERT and ClinicalBERT, perform significantly better than general models on clinical tasks [14, 15]. Translating this principle into Finnish, this thesis contends that open-source multilingual or Finnish-adapted LLMs are able to produce structured, accurate, and interoperable clinical notes from transcribed speech.

3 Methodological Approach

This section outlines the process for data creation for fine-tuning, selecting open-source LLMs, pre-train model finding from hugging face platform, and defining the evaluation methodology. Specifically, for Finnish datasets preparation, and finding model from hugging face based on selection criteria are most challenging parts of this thesis.

3.1 Data Creation

The data creation process began with the establishment of a foundational dataset comprising Finnish clinical case conversations. These baseline conversations were formatted simulated interactions between healthcare professionals such as doctors or nurses and patients. However, students from the Innovation Project [34], play those roles in creating this dataset. The aim was to capture authentic, context-rich dialogue reflective of real-world clinical scenarios. This baseline serves as the core reference for further development and training of conversational models or language-processing systems in the Finnish healthcare context.

To build the dataset, each clinical scenario was documented in two formats: an audio recording in MP3 format and a corresponding textual transcription. The MP3 files represent the actual spoken dialogues, while the text files offer human-readable versions of these recordings in Finnish. These transcriptions are aligned with the audio to ensure accuracy and preserve the nuances of clinical communication, including terminology, emotional tone, and conversational flow.

All records in the dataset follow the same naming pattern for traceability and for structural organisation. Files are encoded in the format I01-G01-C01.txt and I01-G01-C01.mp3, where each component possesses some metadata. The prefix I01 indicates iteration number of the batch of the dataset. The middle segment G01 delineates the specific group of students in the innovation projects who contributed to the creation of that data entry, facilitating the attribution of

performance and contribution between groups of students. The final segment C01 is the clinical case number, which links the file to an individual patient scenario. This naming pattern is applied uniformly to both audio (MP3) and text (TXT) files to facilitate simple management, retrieval, and referencing within research, analysis, or model training pipelines.

Students from different disciplines, like nursing, podiatry, paramedic, and gerontology, worked out to carry out the production process, and this process was monitored by their teachers. Under supervision, these students actively contributed to the creation and transcription of the clinical dialogues, guaranteeing clinical authenticity and realism. Their multidisciplinary contribution was critical in the creation of a dataset that achieves a balance between healthcare usefulness and pedagogical usefulness. The generated resource supports instructional use cases in nursing and healthcare simulation settings and offers a basis for training AI systems in clinical communication in the Finnish language.

3.2 Hugging Face Platform

The development of the Transformer architecture was a game-changer for the natural language processing (NLP) community. Transformers have significantly outperformed previous models, especially those based on recurrent neural networks (RNNs), because of their greater parallel processing, scalability, and performance. Although RNN-based models are still useful for some sequential tasks, Transformers have emerged as the de facto standard in cutting-edge NLP systems [35]. A growth in pre-training methods, which include first training models on extensive general-purpose corpora to enable them to acquire deep language representations before being optimised for particular tasks, also accompanied this architectural change [36].

As the capability and complexity of NLP models grew, there was a certain needs for open platforms that could democratize access to these advanced tools. This requirement was met by the creation of Hugging Face, a company and open-source community that aims to make machine learning models reusable, adaptable, and accessible to developers and researchers. Hugging Face provides a unified ecosystem for model sharing, experimentation, and collaboration across the community, eventually channelling robust NLP models into everyday applications [37].

The Hugging Face platform is now the central location of the global machine learning (ML) community, popularly renowned for its heavily utilised Transformers library. It is an open-source library with a single application programming interface (API) and an increasing range of pre-trained transformer-based models that have been created by the Hugging Face team and by an enormous user community. With contributions from hundreds of open-source developers across its repositories, Hugging Face fosters collaborative development by offering an open ecosystem for sharing, fine-tuning, deploying, and evaluating machine learning models, particularly in the domain of natural language processing [38]. Modularity of the library makes integration of tokenizers, Transformer models, and task-

specific heads simple and is extremely well-suited to a wide range of applications such as in educational environments, health systems, and chatbots.

The Hugging Face offers an end-to-end environment to build and deploy domain-specific language models with its main building blocks: datasets, models, spaces, and language support. The datasets hub allows researchers to share and access curated corpora, including domain-specific resources like Finnish clinical dialogue. The model's hub contains thousands of pre-trained and fine-tuned models, which can be further fine-tuned for various natural language processing tasks [39]. Spaces provide a collaborative environment to implement and showcase machine learning models with ease of configuration, where there is more convenience in experimenting and assessing model performance. Hugging Face models also supports large numbers of languages like Finnish, which are required for developing NLP solutions that comply with national healthcare programs. Hugging Face, equipped with these packaged tools and open community assets, is particularly beneficial for developing and optimizing smart agents in streamlined clinical documentation workflows. Figure. 2 depicts the most common elements from the Hugging Face universe.

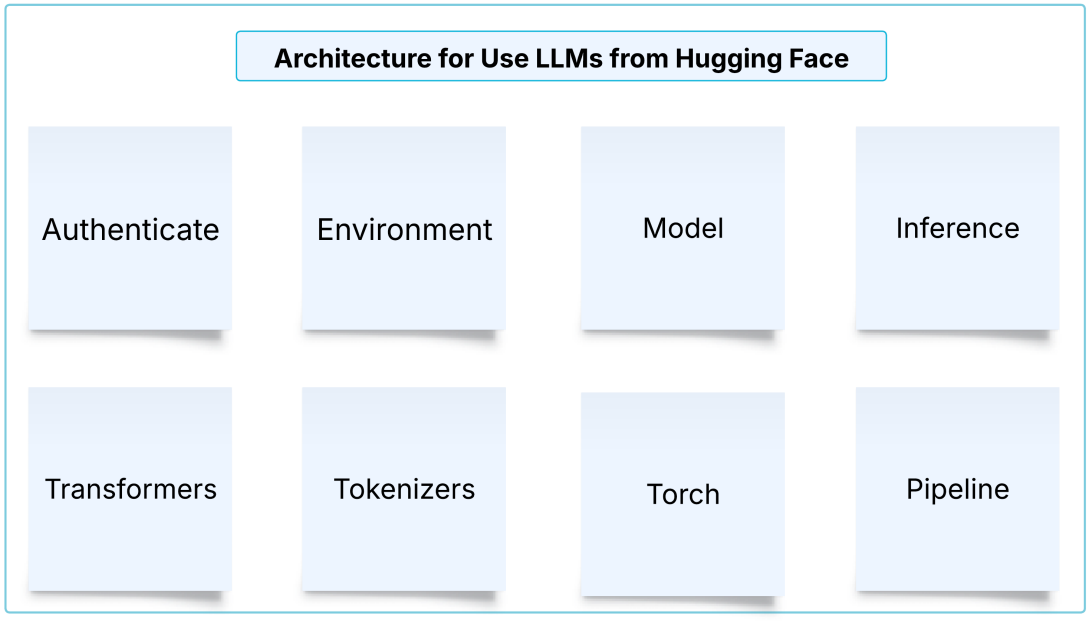

Figure 2: Architecture Components for Use LLMs from Hugging Face.

Moreover, Hugging Face brings significant value to academic and applied research with its community-maintained model hub, reproducibility, and open-source culture. Researchers and developers working on specialised domains such as education, healthcare, or low-resource languages benefit from Hugging Face's infrastructure that simplifies general-purpose LLMs into domain-specific tasks [40]. With such capabilities as the Trainer API, Low-Rank Adaptation of Large Language Models (LoRA) fine-tuning, and model spaces in hosting, even non-commercial clients can now readily customise such gigantic models like LLaMA 3 to certain ends, narrowing the gap between high-end NLP performance and practical application in real-world systems under high stakes.

3.3 Selection of LLM Model

For this research, selecting the appropriate Large Language Models (LLMs) was guided by a set of criteria that align with the specific requirements of structured clinical documentation in Finnish. The goal was to identify models that not only demonstrate a high level of accuracy in medical language processing but also provide flexibility for customisation and support for the Finnish language. The core selection criteria included:

1. Ability to handle medical text optimisation,

2. Support for structured clinical documentation,

3. Customisation and fine-tuning capabilities, and

4. Finnish language support, pre-training or adaptability for continued learning.

Ability to handle medical text optimisation

Large language model parameters are model internal variables weights and biases obtained during training. The parameters determine the efficacy of the model in identifying linguistic patterns, sensing context, and generating meaningful language. The size and configuration of the parameters dictate the complexity and efficacy of the model [41].

Support for structured clinical documentation

Structured clinical documentation is the process of taking unstructured medical information such as free-text conversation or handwritten orders and placing it into a standardized, structured format. It enables quicker data retrieval, clinical decision-making, health system interoperability, and secondary use such as analytics and research [42].

Customisation and fine-tuning capabilities

A pre-trained and fine-tuneable Finnish model has initially gone through an initial training process with a wide corpus of Finnish-language text, instructing it in the grammar, syntax, vocabulary, and language idiosyncrasies of the language [43]. On top of its base competence, the model can also be further trained or fine-

tuned on specialized Finnish datasets in order to enhance its performance on expert tasks such as clinical reporting or medical information extraction. This twofold attribute allows for high accuracy and context sensitivity in Finnish-language.

Finnish language support, pre-training or adaptability for continued learning

An adaptable model can be modified or extended to accommodate specific research objectives or domain requirements [44]. It includes concepts like modification of the model structure, addition of new data sets, or tuning of the training objectives for optimal task-specific performance.

Models such as BioGPT [45] and ClinicalBERT [46] offer strong performance in medical domain understanding due to pre-training on clinical texts; however, they lack native Finnish language capabilities. However, models like SiloGen Finnish GPT [47] and TurkuNLP FinBERT [48] are particularly made with support for the Finnish language and, with further refinement, have great promise in structured data extraction. When tailored for domain-specific activities like transcription structure and electronic health record summarisation, open-access architectures like Teuken-7B [49] and DeepSeek [50] provide a solid balance between scalability and flexibility. Based on all criteria and models' availability on Hugging face we developed a models comparison table to solve this selection stage, table 1 explain it in detail.

Table 1. Models' comparison based on selection criteria.

Model Name	Parameters B(Billion), M(Million)	Structured Clinical Documentation	Pre-trained or Support Finnish for Fine-tune	Customisable
Llama 3 (Meta) [51]	8B, 70B, 400B	Yes	Yes	Yes
SiloGen Finnish GPT [47]	7B, 13B, and 33B	Yes	Yes	Yes
Teuken-7B (OpenGPT-X) [57]	7B	Yes (with fine-tuning)	Yes	Yes
Hippocrates LLM [49]	7B, 13B	Yes	No	Yes
DeepSeek [50]	7B, 67B	Yes (with fine-tuning)	Yes	Yes
TurkuNLP FinBERT [48]	110M	Yes (with fine-tuning)	Yes	Yes
BioGPT [45] (Microsoft)	1.2B	Yes	Yes	Yes
ClinicalBERT [46]	110M	Yes	No	Yes

Based on this comparative analysis, LLaMA 3 (Meta) [51] is selected model for this research, we used LLaMa 3.1-8B(Billion) model for this study. It is a versatile and powerful open-source LLM available in 8B, 70B, and 400B parameter sizes, offering high accuracy, coherence, and scalability across a wide range of tasks. It supports fine-tuning and LoRA-based customisation, making it adaptable for medical scribing through domain-specific Finnish clinical conversations. Additionally, the model is readily available through platforms like Hugging Face, enabling easy access, integration, and experimentation within standard machine learning workflows. Its strong general-language capabilities, combined with flexibility in training and deployment, make it an ideal choice for generating clinical conversation documentation in Finnish healthcare settings.

3.4 Evaluation Methodology

This section of the thesis gives a clear explanation of the method that was used to assess the performance of the fine-tuned model. Specifically, it outlines how model outputs were contrasted with reference clinical notes made by humans for the purposes of determining the quality of structured documentation that is generated from clinical conversations in the Finnish language.

Three commonly used automatic assessment metrics Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and Bidirectional Encoder Representations from Transformers Score (BERTScore) are used to objectively evaluate the quality of the clinical documentation produced by the fine-tuned large language model (LLM). By comparing the produced output with actual clinical notes, these metrics allow for the examination of many morphological and semantic aspects.

The BLEU score measures the n-gram precision between the generated and reference texts, indicating how many segments of the generated content are lexically identical to the reference [52]. The percentage of appropriately produced words (or segments) that also occur in the reference text is referred to as precision. Despite being created initially for machine translation, BLEU is frequently used in text creation tasks to assess the consistency and syntactic accuracy of model output. It is used to evaluate the model's adherence to standard clinical vocabulary and wording in this study.

The ROUGE metric, in particular ROUGE-N and ROUGE-L [53], is widely used to evaluate the quality of generated text by measuring the number of overlapping sequences between system and reference texts. ROUGE-N counts the number of perfect matches of n-grams (i.e., unigrams for ROUGE-1, bigrams for ROUGE-2), while ROUGE-L computes the longest common subsequence (LCS), with non-contiguous matches with word order allowed. This sensitivity to overlapping sequences is particularly valuable for clinical NLP tasks, e.g., annotating unstructured medical dialogue with structured clinical notes, where accurate

representation of medical terms and preservation of semantic relations are critical. In these domains, even incomplete phrase overlaps e.g., "shortness of breath" and "difficulty breathing" can convey clinically equivalent information, and ROUGE's granularity detects whether key content is preserved. By rewarding matching medical phrases and punishing omissions and distortions, ROUGE provides an estimate of content fidelity that is good enough to serve as a suitable metric to evaluate summarization-like systems in clinical reports.

To assess the semantic similarity between the produced and reference material, BERTScore is included to go beyond surface-level lexical matching [54]. Instead of utilising precise word or sequence overlap like BLEU and ROUGE do, BERTScore uses contextualised word embeddings from transformer models that have already been trained, like BERT or ClinicalBERT, to calculate similarity. These embeddings allow BERTScore to compare the meaning of words in sentences based on their context, rather than relying solely on exact word matches. Due to the prevalence of paraphrased terms with the same meaning (e.g., "elevated blood pressure" vs. "hypertension"), it is perfect for clinical natural language processing tasks.

These three indicators work together to offer a comprehensive framework for evaluation. ROUGE gauges content recall, BERTScore records semantic integrity, and BLEU quantitatively assesses syntactic accuracy. This complete evaluation technique facilitates an impartial assessment of the model's capacity to produce accurate, therapeutically pertinent, and cohesive documentation from Finnish clinical dialogues.

3.5 K-Fold Cross-Validation

Cross-validation is a standard method of machine learning model evaluation that is used to assess the generalizability and robustness of models on small data sets. Among its forms, K-fold cross-validation becomes extremely useful when the data set is too small to accommodate a clear training, validation, and test split. In this case, the data set is divided into K equal-sized parts, or K folds. The model is then K times trained, once using K−1 folds to train and the other to utilize as validation [55]. The performance score is determined by averaging the results across all the folds. The process makes both training and testing stages take advantage of every sample so it's a useful solution for small-data cases.

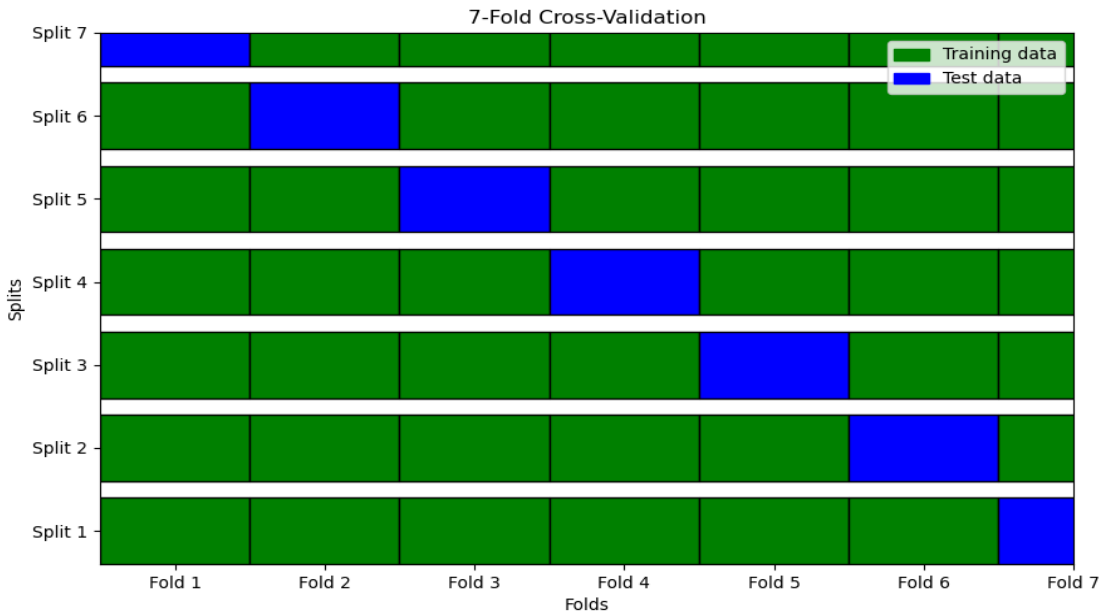

Figure 3: K-Fold Cross-Validation (k=7) Structure.

In this study, the data set is limited to only seven clinical dialogues in Finnish with their corresponding MP3 files and manually annotated reference documents. Due to the limited number of samples, standard single-split validation would lead to unstable and biased performance estimates. To combat this limitation, a 7-fold cross-validation (leave-one-out) method is adopted where a single conversation is employed as test case while learning is conducted on the remaining six, which

is dispatch in the figure 3. This allows for a greater overall understanding of the model's behaviour under various patient interactions and clinical subjects, which is essential in the healthcare sector where data diversity is as essential as data volume.

K-fold cross-validation offers several advantages that complement the objectives of fine-tuning large language models (LLMs) for domain-specific applications. It reduces variance caused by random sampling, provides stability in performance across heterogeneous samples, and is used to identify outliers or inconsistencies in the data [56]. For LLMs, which are prone to overfitting in low-resource settings, cross-validation is a safeguard mechanism by exposing model generalization to multiple tests. In addition, accompanied by evaluation metrics such as BLEU, ROUGE, and BERTScore, it supports both syntactic and semantic testing of output creation. K-fold validation in this case not only provides reliability to testing but also the reproducibility and scientific value of analysis on the model's performance.

4 Experimental Study and Analysis

This section presents the experimental setup, dataset configuration, fine-tuning procedure, and evaluation method used in the study. It discusses how the model shown was tested based on Finnish clinical conversations and provides quantitative results along with analytical findings obtained from automatic evaluation metrics.

4.1 Dataset Summary

As explained in Section 3.1, the study utilized a custom dataset drawn from Finnish-language clinical MP3 records and their manually written textual equivalents. The dataset was specifically curated to reflect clinical conversation in Finnish, capturing a variety of medical scenarios and communication styles. Audio recordings were generated from existing text data to ensure that the resulting transcriptions maintained both clinical appropriateness and linguistic precision, supporting effective natural language processing.

The dataset is comprised of seven complete clinical dialogues. Pre-processing was applied to each sample and transferred into structured JSON format specially to cater to the training requirements of large language models [57]. The JSON format has a clear segmentation of dialogue turns, speaker designation (e.g., doctor or patient), and annotated fields that indicate the target output for clinical documentation. This format allowed for supervised fine-tuning, which allowed the model to learn how to convert free-form clinical discussion into well-formed, documentation that met the standards of healthcare.

4.2 Experimental Setup

This section outlines the technical configuration used for model training and evaluation. Experiments were conducted on the CSC Puhti supercomputer [58], which provides a reliable computing environment for tackling intricate machine learning operations. The infrastructure was selected in order to support the fine-tuning of large language models in a fast and secure manner, particularly for domain-specific application scenarios such as Finnish clinical documentation.

The setup involved a proper balance of GPU-optimized hardware and machine learning libraries required for model training, testing, and deployment. Different tools and frameworks were utilized to attain transformer-based model compatibility and scalable fine-tuning pipelines. This installation allowed for reproducibility, efficient experimenting, and proper adaptation of models to the Finnish healthcare environment.

4.2.1 Software

This work uses Python Ver. 3.11.5, PyTorch version ≥ 2.0 with Compute Unified Device Architecture (CUDA) 11.7, and open-source ML libraries on a GPU based Puhti supercomputer for fine-tuning, validation, and quantitative evaluation.

This makes use of core machine learning libraries such as PyTorch for model training and GPU consumption and CUDA for parallel computations on the Puhti supercomputer. The Hugging Face Transformers library is used to fetch pre-trained LLaMA models and tokenizers, and the datasets library to enable easy data handling, thus making the fine-tuning procedure efficient and scalable. The voice recordings used in this study were first transcribed using the Whisper large-v3 model and later pre-processed into a uniform format for training large language models.

4.2.2 Hardware

The fine-tuning experiments were conducted on the CSC Puhti supercomputer, a high-performance computing facility located in Finland and operated by CSC – IT Center for Science. Hardware configuration that was used in this study consisted of one NVIDIA A100 GPU with 40 GB VRAM, which had sufficient memory capacity to execute LLaMA 3.1–8B in 4-bit quantized mode with the help of LoRA adapters [59]. The node consisted of 16 CPU cores and 128 GB of system memory to support preprocessing, loading of data, and executing tasks in parallel. Approximately 200 GB of disk space was occupied by maintaining model weights, training outputs, logs, and dataset files [60].

The setup was tuned for a compromise between compute efficiency and the model's requirements to enable fine-tuning with good performance without inefficient use of resources. The CSC Puhti supercomputer has ensured stable operation, high-speed I/O, and reproducibility from the whole training and testing.

4.3 Experimental Setup on CSC Puhti Supercomputer

In order to begin the final experimental setup, user must first create an account on the CSC cloud service [61]. After making an account, user must build a project there. The setup procedure starts with a project creation based on the thesis title. In order to approve project, user must answer a few questions and receive a project ID. As a student of Metropolia University of Applied Science, we received 100,000 billing units by CSC service to begin our project. The billing units gave access to the Puhti supercomputer and related CSC cloud services that enabled us to carry out the whole model training, fine-tuning, and evaluation phases without incurring any computational charges.

The project must be executed on a computing environment that fully satisfy the hardware and software requirements of the GPU node, which include one NVIDIA A100 Tensor Core GPU (40 GB (HBM), High Bandwidth Memory), 16 Xeon CPU cores, 128 GB of RAM, and around 200 GB of local disc space [62, 63]. For the

purpose of training a large language model without excessive overhead, this arrangement balanced memory and computation. Specifically, an 8-billion-parameter LLaMA 3.1 model might be loaded in 4-bit quantised format using low-rank (LoRA) adapters and the A100's 40 GB VRAM, which has been found to significantly reduce memory needs. The 16 CPU cores of the node allowed for parallel I/O and data pre-processing tasks. Specifically, Puhti also provides high-speed interconnect and storage: its Lustre file system provides the order of 50 GB/s aggregate I/O bandwidth, and the center's controlled software modules guarantee reliable, reproducible performance. These hardware capabilities allowed successful fine-tuning runs.

Before beginning the coding and tuning phases of the project, obtaining access to the LLaMA 3.1–8B model on Hugging Face through Meta's official repository was required [64]. Since the LLaMA models are released under a bespoke license with some restrictions regarding how they may be utilized, access is not granted by default [65]. To proceed, the research team were required to request access to the models by submitting an application through the Hugging Face platform. This involved consenting to Meta's terms of license, which outline permissible use cases, limitations, and citation policies. The applicants were also required to sign off on the responsible use form, confirming compliance with legal and ethical standards, including non-commercial use and safe deployment. Once approved, model weights were downloadable and could be included in the project environment. This ensured all model usage was done under terms described by Meta to enable proper and secure deployment of LLaMA 3.1–8B in the fine-tuning pipeline.

4.4 Finetuning

The fine-tuning of the LLaMA 3.1–8B model represented an important step in adapting a general-purpose LLM into a specialized system for processing spoken clinical conversations [66], with the broader aim of supporting healthcare-related educational goals. The training dataset consisted of MP3-formatted audio recordings of clinical conversation and their textual transcripts, where, students of the Innovation Project, act healthcare professional (e.g. doctor, nurse, therapist) and patient both roles in creating this dataset. The audio recordings were transcribed using the Whisper large-v3 model and later pre-processed into a standardized format suitable for training LLM. Every pair of transcription and ground-truth reference was merged into one text field as model input. Five of the audio-transcript files were selected as initial training and test files for two files each. Subsequently, a 7-fold cross-validation scheme was adopted to facilitate robust evaluation and improve generalizability of the model across the relatively small dataset. This validation procedure enabled the model to be repeatedly trained and tested against various data partitions, reducing the likelihood of overfitting and enhancing its dependability for use downstream in clinical documentation and conversational AI.

The fine-tuning was conducted using the meta-llama/Llama-3.1–8B model, accessed from a local directory after obtaining the necessary license from Meta. The model was added to memory with *float16* precision and *device_map="auto"* to utilize available GPU memory capacity on the CSC Puhti supercomputer. The LLaMA architecture-compatible tokenizer was also initialized with the original tokenizer model and utilized to convert textual training data into sequences of tokens. Records were padded or truncated up to a maximum of *512 tokens* to maintain input lengths consistent during training. This 512 tokens was chosen as the input length limit to maintain transformer structure compatibility, with tractable memory usage in training while still gathering adequate clinical context.

The training involved using Hugging Face's *Trainer API* in half-precision supervised fine-tuning mode. Significant hyperparameters included a batch size of one (1), three (3) epochs of training, logging at every 10 steps (e g. 5 training samples × 3 epochs = 15 steps per fold), and saving the checkpoint at every 100 steps with a limit of 2 saved checkpoints. Gradient checkpointing was not used under this setup because the training setup was made minimal for the sake of clarity and stability. The models were trained in half-precision floating point *(fp16=True)* to maintain the calculation below the memory constraint of the available *NVIDIA V100 GPUs* on Puhti.

The training script included a custom tokenizer and data collator for causal language modelling. The *DataCollatorForLanguageModeling* was used with masked language modelling disabled (*mlm=False*), aligning with the LLaMA model's causal architecture. No evaluation was conducted during training (*evaluation_strategy="no"*), as the focus was on completing a robust initial fine-tuning pass on the full training set.

Following the process of training, the fine-tuned model was saved in different forms. First, the model and tokenizer were saved in Hugging Face transformers format through *save_pretrained()* for convenient reloading as well as continued development. Additionally, the *PyTorch* state dictionary for the model was saved in *.pth* format for non-Hugging Face out-of-the-box applications or non-Hugging Face frameworks compatibility. While LLaMA models themselves aren't natively *TensorFlow/Keras* compatible, there was a placeholder *.h5* file included to show potential future need for a conversion, though full *.h5* export is not currently supported.

All computations were done with SLURM job scripts on the Puhti GPU cluster, with GPU reservations requested by *--gres=gpu:v100:1*, *16 CPU* cores, and *128 GB RAM*. The virtual Python environment was also activated within the job script to assure stable dependency management and reproducibility.

This fine-tuning process produced a domain-tuned LLaMA 3.1–8B model with the ability to generate Finnish clinical conversation texts in the given contexts. The model can now translate spoken Finnish medical discourse between patients and healthcare professionals into neatly organized clinical conversation documents. Trained model and its training and preprocessing scripts are made available publicly at GitHub facilitating future research and deployment in clinical AI tools.

5 Results and Evaluation

To evaluate the model ability after fine-tuned, research conducted automatic testing using BLEU, ROUGE-L, and BERTScore measures [40, 41]. These analyses were applied between generated documents and ground-true documents of Finnish clinical conversations. The evaluation process took a 7-fold cross-validation method to ensure proper performance estimation on different subsets of the data and prevent overfitting with respect to the finite amount of data available, table 2 illustrated the results of k-fold validation. This helped in ensuring stringent evaluation of the model's capability to generalize to clinical conversations generation and documentation tasks.

Table 2: 7-Fold Cross-Validation Results.

File Names	**BLEU**	**ROUGE-L**	**BERTScore**
I01-G01-C01.txt	0.0739	0.3156	0.7642
I01-G01-C02.txt	0.1579	0.5151	0.8384
I01-G01-C03.txt	0.1154	0.5759	0.8621
I01-G02-C01.txt	0.1330	0.5829	0.8472
I01-G02-C02.txt	0.1247	0.4558	0.7905
I01-G03-C01.txt	0.1151	0.5447	0.8390
I01-G03-C02.txt	0.1295	0.4974	0.8197
Average (k-fold result)	0.1242	0.4982	0.8373

BLEU was employed to evaluate the precision of n-gram overlaps between the model's outputs and reference transcripts. It served as a key indicator of how well the model replicated human-authored clinical conversation. ROUGE-L emphasizing recall, was used to assess how comprehensively the generated text matched the reference content. Additionally, we included BERTScore, a semantic similarity metric based on contextual embeddings, to better capture the meaning-

level fidelity of the outputs, especially important in the healthcare context where paraphrasing and variation in wording are common.

Across 7-fold cross-validation, the model achieved a mean BLEU score of 0.1242, ROUGE-L score of 0.4982, and BERTScore of 0.8373. These all reflect moderate syntactic similarity and high semantic matching with human-transcribed references due to the complexity and domain-specific nature of the input data. In general, 10–20% BLEU scores are generally moderate, particularly in low-resource and high-variety settings [52]. ROUGE-L scores over 0.5 indicate adequate overlap between generated and reference text [53]. By contrast, BERTScore scores of above 0.8 represent high semantic similarity, with scores of 0.85 or higher being generally regarded as representative of high-quality semantic-level correspondence [54]. These test results confirm that fine-tune LLMs model, LLaMA 3.1–8B's capability to generate contextually appropriate and linguistically coherent outputs from Finnish clinical audio inputs. Besides that, a thoroughly analysis is conducting by team member of Digital Medical Scribe project [67]. The 7-fold cross-validation complemented the credibility of the performance metrics by demonstrating stable outcomes across diverse data partitions. In conclusion, the evaluation supports the model’s readiness for downstream applications in Finnish-language clinical NLP systems.

6 Discussion and Conclusions

The fine-tuning of the LLaMA 3.1–8B model for Finnish clinical conversation documents generation highlights the effectiveness of domain-specific adaptation of large language models (LLMs). Through the integration of hand-curated clinical conversation transcripts and Whisper-generated transcriptions, the model demonstrated a measurable capacity for understanding and generating clinical content in Finnish. The performance scores reflect that the fine-tuned model was able to capture both syntactic structure and semantic coherence, even with a relatively small dataset. These results support the feasibility of using specialized LLMs for tasks such as clinical documentation or medical education tools in low-resource languages.

The fine-tuned LLaMA 3.1–8B model was more in line with the linguistic nuances of Finnish healthcare communication. While larger models tend to outperform smaller models on general benchmarks due to large-scale pretraining, this study demonstrates that task-specific fine-tuning even on a limited dataset.

The k-fold cross-validation improved the model's strength and reduced overfitting issues, especially with domain-specific data like clinical texts. It improved the generalizability of the model towards diverse input patterns. While BERTScore (0.8373) reported high semantic retention, BLEU (0.1242) and ROUGE (0.4982) suggested there were areas to be worked on in precision and recall. Using training on the CSC Puhti supercomputer with fp16 optimization and LoRA validated the process to be sustainable using limited resources, albeit parameter fine-tuning is required for replicating the process in less-capable machines.

A key limitation of this study is the small size of the dataset, which consisted of only seven Finnish clinical dialogues, and the fact that the LLaMA 3.1–8B model which is the only model used, is not natively pre-trained in Finnish. These constraints may affect the model's ability to learn linguistic nuances as well as clinical domain-specific patterns. Future work could involve expanding the training dataset with more authentic or synthetic Finnish clinical conversation to

further enhance model generalization, and there should be train others models and comparison between the models, the baseline results (without pre-training) will also need to evaluate. Moreover, additional architectural refinements, such as incorporating structured end-of-sequence tokens or leveraging retrieval-augmented generation, may increase output accuracy and reduce hallucination. As shown in related works using models like TurkuNLP FinBERT [48], or ClinicalBERT [46], smaller LLMs can be made more effective through such augmentation, suggesting potential for lighter deployment scenarios in healthcare systems with resource limitations.

In the end, while this study focused on automatic evaluation, it also opens the door for integrating human-in-the-loop feedback in future iterations. Manual validation by healthcare professionals and Finnish language experts could provide more nuanced insight into the model's output quality. Coupling linguistic metrics with clinical validity assessments would ensure the practical applicability of such models in real-world environments. The success of this project suggests that fine-tuned LLMs, like LLaMA 3.1–8B can serve as foundational tools in the development of intelligent medical scribes and conversational AI for Finnish clinical contexts.

7 Acknowledgement

The authors would like to thank Digital Medical Scribe project team members, Principal Lecturer Mikael Soini, Principal Lecturer Päivi Haho, Senior Lecturer Sakari Lukkarinen, Student from Masters' of IT Nazrul Kabir, Student from Masters' of IT Huy Wõ students and supervisors of Nursing, Bachelor's Degree, Nursing Bilingual Top-Up Degree, Biomedical Laboratory Science, Bachelor's Degree, Laboratory Science, Bachelor's Degree, Bachelor of Health Care, Physiotherapist (top-up), and Bachelor of Health Care, Physiotherapist for their support in collecting the dataset, and direct and indirect contribution to the project.

References

1. Smith, Roger O., et al. “Assistive Technology Products: A Position Paper from the First Global Research, Innovation, and Education on Assistive Technology (GREAT) Summit.” Disability and Rehabilitation Assistive Technology, vol. 13, no. 5, June 2018, pp. 473–85, https://doi.org/10.1080/17483107.2018.1473895.

2. Meyers, Steven, et al. “Structured Clinical Documentation to Improve Quality and Support Practice-Based Research in Headache.” Headache The Journal of Head and Face Pain, vol. 58, no. 8, Aug. 2018, pp. 1211–18, https://doi.org/10.1111/head.13348.

3. Mishra, Pranita, et al. “Association of Medical Scribes in Primary Care With Physician Workflow and Patient Experience.” JAMA Internal Medicine, vol. 178, no. 11, Sept. 2018, pp. 1467–1467, https://doi.org/10.1001/jamainternmed.2018.3956.

4. Taylor, Kimberly A., et al. “Medical Scribe Impact on Patient and Provider Experience.” Military Medicine, vol. 184, no. 9–10, Feb. 2019, pp. 388–93, https://doi.org/10.1093/milmed/usz030.

5. Nawab, Khalid. “Artificial Intelligence Scribe: A New Era in Medical Documentation.” Unknown Journal, vol. 0, no. 0, Sept. 2024, pp. 3103–3103, https://doi.org/10.36922/aih.3103.

6. Brown TB, et al. Language Models are Few-Shot Learners. NeurIPS 2020. Available from: https://arxiv.org/abs/2005.14165

7. Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, et al. GPT-NeoX-20B: An open-source autoregressive language model. EleutherAI; 2022 Apr [cited 2025 May 18]. Available from: https://arxiv.org/abs/2204.06745

8. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv [Preprint]. 2023 Feb [cited 2025 May 18]. Available from: https://arxiv.org/abs/2302.13971

9. Penedo G, Almazrouei E, Liu Y, Poulain R, Scao TD, Akiki C, et al. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and the Falcon 40B Model. arXiv [Preprint]. 2023 Jun [cited 2025 May 18]. Available from: https://arxiv.org/abs/2306.01116

10. Liu, Rui, et al. “Text-to-Speech for Low-Resource Agglutinative Language With Morphology-Aware Language Model Pre-Training.” IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 32, Jan. 2024, pp. 1075–87, https://doi.org/10.1109/taslp.2023.3348762.

11. Wu, Chaoyi, et al. “PMC-LLaMA: Toward Building Open-Source Language Models for Medicine.” Journal of the American Medical Informatics Association, vol. 31, no. 9, Apr. 2024, pp. 1833–43, https://doi.org/10.1093/jamia/ocae045.

12. Heilmeyer, Felix, et al. “Viability of Open Large Language Models for Clinical Documentation in German Health Care: Real-World Model Evaluation Study.” JMIR Medical Informatics, vol. 12, June 2024, pp. e59617–e59617, https://doi.org/10.2196/59617.

13. Keenious [Internet]. Oslo, Norway: Keenious AS; c2020 [cited 2025 May 10]. Available from: https://www.keenious.com/

14. Lee, Jinhyuk, et al. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” Bioinformatics, vol. 36, no. 4, Sept. 2019, pp. 1234–40, https://doi.org/10.1093/bioinformatics/btz682.

15. Alsentzer, E., et al. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv: https://arxiv.org/abs/1904.03323.

16. Li, Jinyu. “Recent Advances in End-to-End Automatic Speech Recognition.” APSIPA Transactions on Signal and Information Processing, vol. 11, no. 1, Jan. 2022, https://doi.org/10.1561/116.00000050.

17. van Buchem, Marieke, et al. “Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study.” JMIR AI, vol. 3, Sept. 2024, pp. e60020–e60020, https://doi.org/10.2196/60020.

18. Quiroz, Juan C., et al. “Challenges of Developing a Digital Scribe to Reduce Clinical Documentation Burden.” Npj Digital Medicine, vol. 2, no. 1, Nov. 2019, https://doi.org/10.1038/s41746-019-0190-1.

19. Shah, Lisa M., et al. “Effects of Medical Scribes on Patients, Physicians, and Safety: A Scoping Review.” Knowledge Management & E-Learning An International Journal, Dec. 2021, pp. 559–629, https://doi.org/10.34105/j.kmel.2021.13.030.

20. Gidwani, Risha, et al. “Impact of Scribes on Physician Satisfaction, Patient Satisfaction, and Charting Efficiency: A Randomized Controlled Trial.” The Annals of Family Medicine, vol. 15, no. 5, Sept. 2017, pp. 427–33, https://doi.org/10.1370/afm.2122.

21. Pearson, Elsa, and Austin B. Frakt. “Medical Scribes, Productivity, and Satisfaction.” JAMA, vol. 321, no. 7, Feb. 2019, pp. 635–635, https://doi.org/10.1001/jama.2019.0268.

22. Krishna, Kundan, et al. “Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques.” arXiv (Cornell University), Jan. 2020, https://doi.org/10.48550/arxiv.2005.01795.

23. Peng, Yifan, et al. “Reproducing Whisper-Style Training Using An Open-Source Toolkit And Publicly Available Data.” 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2023, pp. 1–8, https://doi.org/10.1109/asru57964.2023.10389676.

24. Lee, Wookey, et al. “Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review.” Sensors, vol. 21, no. 4, Feb. 2021, pp. 1399–1399, https://doi.org/10.3390/s21041399.

25. Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, et al. GPT-NeoX-20B: An open-source autoregressive language model. EleutherAI; 2022 Apr [cited 2025 May 18]. Available from: https://arxiv.org/abs/2204.06745

26. Scao TD, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, et al. BLOOM: A 176B-parameter open-access multilingual language model. arXiv [Preprint]. 2022 Oct [cited 2025 May 18]. Available from: https://arxiv.org/abs/2211.05100

27. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv [Preprint]. 2023 Feb [cited 2025 May 18]. Available from: https://arxiv.org/abs/2302.13971

28. Penedo G, Almazrouei E, Liu Y, Poulain R, Scao TD, Akiki C, et al. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and the Falcon 40B Model. arXiv [Preprint]. 2023 Jun [cited 2025 May 18]. Available from: https://arxiv.org/abs/2306.01116

29. Raza S, Nooralahzadeh F. Towards automatic FHIR extraction from clinical notes using sequence-to-sequence learning. Appl Clin Inform. 2022;13(1):72–83. Available from: https://doi.org/10.1055/s-0041-1741456

30. Liu T, Du Z, Ding Y, Wang Z, Tang B, Fu Y, et al. MedDialog: A large-scale medical dialogue dataset [Preprint]. arXiv:2010.07497 [Internet]. 2021 [cited 2025 May 12]. Available from: https://arxiv.org/abs/2010.07497

31. Bhutto, Sajida Raz, et al. “Automatic ICD-10-CM Coding via Lambda-Scaled Attention Based Deep Learning Model.” Methods, vol. 222, Dec. 2023, pp. 19–27, https://doi.org/10.1016/j.ymeth.2023.11.017.

32. Virtanen, A., et al. (2020). Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv: https://arxiv.org/abs/1912.07076.

33. Aalto University. Aalto-ASR: Aalto University Automatic Speech Recognition for Finnish [Internet]. Espoo: Aalto University; [cited 2025 May 18]. Available from: https://github.com/aalto-speech

34. Metropolia University of Applied Sciences. Innovation projects [Internet]. Helsinki, Finland: Metropolia UAS; [cited 2025 May 11]. Available from: https://www.metropolia.fi/en/rdi/innovation-projects

35. Hassan, Abid, et al. “Development of NLP-Integrated Intelligent Web System for E-Mental Health.” Computational and Mathematical Methods in Medicine, vol. 2021, Dec. 2021, pp. 1–20, https://doi.org/10.1155/2021/1546343.

36. Mikolov, Tomáš, et al. “Advances in Pre-Training Distributed Word Representations.” arXiv (Cornell University), Jan. 2017, https://doi.org/10.48550/arxiv.1712.09405.

37. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv [Preprint]. 2020 Jul 14 [cited 2025 May 11]; arXiv:1910.03771. Available from: https://arxiv.org/abs/1910.03771

38. Choquette, Jack, et al. “NVIDIA A100 Tensor Core GPU: Performance and Innovation.” IEEE Micro, vol. 41, no. 2, Feb. 2021, pp. 29–35, https://doi.org/10.1109/mm.2021.3061394.

39. Tunstall L, von Werra L, Wolf T. Natural language processing with transformers. Revised ed. Sebastopol (CA): O'Reilly Media, Inc.; 2022.

40. Shen, Yongliang, et al. “HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face.” arXiv (Cornell University), Jan. 2023, https://doi.org/10.48550/arxiv.2303.17580.

41. Hu, Zhiqiang, et al. “LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models.” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2023, https://doi.org/10.18653/v1/2023.emnlp-main.319.

42. Ebbers, Tom, et al. “The Impact of Structured and Standardized Documentation on Documentation Quality; a Multicenter, Retrospective Study.” Journal of Medical Systems, vol. 46, no. 7, May 2022, https://doi.org/10.1007/s10916-022-01837-9.

43. Jain, Abhilash, et al. “Finnish Language Modeling with Deep Transformer Models.” arXiv (Cornell University), Jan. 2020, https://doi.org/10.48550/arxiv.2003.11562.

44. Zeng, Guo-Qing, et al. “Adaptable and Precise: Enterprise-Scenario LLM Function-Calling Capability Training Pipeline.” arXiv (Cornell University), Dec. 2024, https://doi.org/10.48550/arxiv.2412.15660.

45. Luo R, Sun Y, Xiang Y, Qin B, Liu T, Li Z. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. arXiv [Preprint]. 2022. arXiv:2210.10341. Available from: https://arxiv.org/abs/2210.10341

46. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019 Jun;

Minneapolis, MN. Association for Computational Linguistics; 2019. p. 72–8. Available from: https://aclanthology.org/W19-1909

47. Silo AI. SiloGen and Tietoevry Care are developing a Finnish-speaking AI assistant for healthcare professionals [Internet]. AMD Blog; 2023 June 29 [cited 2025 May 18]. Available from: https://www.amd.com/en/blogs/2023/silogen-and-tietoevry-care-are-developing-a-finnish-speaking-ai-assistant-for-healthcare-professionals.html

48. Virtanen A, Kanerva J, Elo M, Luoma J, Luotolahti J, Miekka N, et al. FinBERT: A Finnish pretrained BERT model for financial and biomedical text. arXiv [Preprint]. 2019. arXiv:1912.07076. Available from: https://arxiv.org/abs/1912.07076

49. Acikgoz, An Open-Source Framework for Advancing Large Language Models in Healthcare [Internet]. [cited 2025 May 18]. Available from: https://cyberiada.github.io/Hippocrates/

50. DeepSeek. DeepSeek LLM Models [Internet]. [cited 2025 May 18]. Available from: https://huggingface.co/deepseek-ai

51. Meta AI. LLaMA Model Card [Internet]. Hugging Face; [cited 2025 May 18]. Available from: https://huggingface.co/meta-llama

52. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*; 2002 Jul; Philadelphia, PA. Association for Computational Linguistics; 2002. p. 311–8. https://doi.org/10.3115/1073083.1073135

53. Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop; 2004 Jul; Barcelona, Spain. Association for Computational Linguistics; 2004. p. 74–81. https://aclanthology.org/W04-1013

54. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR); 2020. arXiv preprint arXiv:1904.09675. https://arxiv.org/abs/1904.09675

55. Scikit-learn Developers. Cross-validation: evaluating estimator performance — scikit-learn 1.3.2 documentation [Internet]. Available from: https://scikit-learn.org/stable/modules/cross_validation.html

56. Kislay, Kaustubh, et al. “Evaluating K-Fold Cross Validation for Transformer Based Symbolic Regression Models.” arXiv (Cornell University), Oct. 2024, https://doi.org/10.48550/arxiv.2410.21896.

57. Menon, Niharika G., et al. “Deep Learning Based Transcribing and Summarizing Clinical Conversations.” 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, https://doi.org/10.1109/i-smac52330.2021.9640683.

58. CSC – IT Center for Science. Puhti Supercomputer [Internet]. Espoo, Finland: CSC; [cited 2025 May 18]. Available from: https://docs.csc.fi/computing/systems-puhti/

59. Ye, Zhengmao, et al. “ASPEN: High-Throughput LoRA Fine-Tuning of Large Language Models with a Single GPU.” arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2312.02515.

60. Zhang, Longteng, et al. “LoRA-FA: Memory-Efficient Low-Rank Adaptation for Large Language Models Fine-Tuning.” arXiv (Cornell University), Jan. 2023, https://doi.org/10.48550/arxiv.2308.03303.

61. Vakaloudis, Alex, et al. “Preparation and Execution of Final Year Student Projects on the Cloud.” 2021 IEEE Frontiers in Education Conference (FIE), 2020, pp. 1–7, https://doi.org/10.1109/fie44824.2020.9273971.

62. Narayanan, Deepak, et al. “Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM.” arXiv (Cornell University), Jan. 2021, https://doi.org/10.48550/arxiv.2104.04473.

63. Choquette, Jack, et al. “NVIDIA A100 Tensor Core GPU: Performance and Innovation.” IEEE Micro, vol. 41, no. 2, Feb. 2021, pp. 29–35, https://doi.org/10.1109/mm.2021.3061394.

64. Pol, Urmila R. “Hugging Face: Revolutionizing AI and NLP.” International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 8, Aug. 2024, pp. 1121–24, https://doi.org/10.22214/ijraset.2024.64023.

65. Castaño, Joel, et al. “Lessons Learned from Mining the Hugging Face Repository.” arXiv (Cornell University), Feb. 2024, https://doi.org/10.48550/arxiv.2402.07323.

66. Gema, Aryo Pradipta, et al. “Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain.” arXiv (Cornell University), Jan. 2023, https://doi.org/10.48550/arxiv.2307.03042.

67. Kabir, Nazrul, “Measuring the Effectiveness of Domain-Specific Language Models for Enhancing Digital Scribe Thoroughness”, Master of Engineering thesis, Metropolia University of Applied Sciences, on processing (2025).

Application of Scholar-GPT for Academic Language Enhancement in Thesis Writing

I used OpenAI's Scholar GPT (GPT version 4) to come up with ideas for research design and thesis organization, and enhancing writing and citation style. It was only used to boost clarity, coherence, and academic style. All analysis, research, and content generated were original and created by me, and I take full responsibility for the final product in this thesis.

Code and Resources Repository

All code, scripts, and related resources used within this project are stored on the following GitHub repository: https://github.com/sakluk/digital-scribes. The repository contains data pre-processing scripts used to clean and organize the input data, the main codebase handling training and evaluation tasks, and SLURM job scripts used to run experiments on the high-performance computing (HPC) environment. Setup, dependency, and usage guidance are described in the README file on the repository.