Mittuniversitetet
MID SWEDEN UNIVERSITY

Comparative Evaluation of Language Models for Swedish Medical Translation

Julio Carvajal

Document type: Bachelor Thesis
Main field of study: Computer engineering
Credits: 15
Semester/year: Spring 2025
Supervisor: Magnus Eriksson
Examiner: Patrik Österberg
Course code: DT099G

At Mid Sweden University, it is possible to publish the thesis in full text in DiVA (see appendix for publishing conditions). The publication is open access, which means that the work will be freely available to read and download online. This increases the dissemination and visibility of the degree project.

Open access is becoming the norm for disseminating scientific information online. Mid Sweden University recommends both researchers and students to publish their work open access.

I/we allow publishing in full text (free available online, open access):

☒ Yes, I/we agree to the terms of publication.

☐ No, I/we do not accept that my independent work is published in the public interface in DiVA (only archiving in DiVA).

.....

Location and date: Sundsvall, 16-05-2025

Programme/Course: Bachelor's Thesis DT099

Name: Julio Carvajal

Year of birth: 2000

Abstract

This thesis compares the performance of four language models—GPT-4o, DeepSeek, Gemini, and Claude—on English-Swedish medical term translation, addressing a gap in health-related Natural Language Processing (NLP) for low-resource languages. Using a dataset of 100 terms, the study quantifies accuracy (via BLEU, METEOR, ROUGE-L), robustness (typo correction), and efficiency (response time), employing a relatively strict waterfall method with API-based translation and automated evaluation. Results show GPT-4o achieving a higher BLEU score of 92.75, typo correction of 71.43%, and time per term of 0.51 seconds, though semantic similarity (METEOR: 62.55) and the use of a reference corpus limit the findings. The research highlights the potential of general-purpose models in medical translation but emphasizes the need for clinical verification and domain-specific fine-tuning. The study contributes a benchmark for Swedish medical Natural Language Processing with implications for health accessibility and proposes further research on domain-specific models like MedPaLM 2 and multi-reference datasets. Ethical concerns such as risks of mistranslation are also discussed with emphasis on responsible AI deployment.

Sammanfattning

Denna studie jämför prestandan hos fyra språkmodeller—GPT-4o, DeepSeek, Gemini och Claude—vid översättning av engelsk-svenska medicinska termer, vilket fyller ett gap inom bearbetning av lågresursspråk inom hälsovården. Med en specialutformad datamängd på 100 termer kvantifierar studien noggrannhet (med BLEU, METEOR, ROUGE-L), robusthet (stavfelkorrigering) och effektivitet (svarstid) genom en relativt strikt vattenfallsmetod med API-baserad översättning och automatisk utvärdering. Resultaten visar att GPT-4o uppnådde högre BLEU-poäng på 92,75, stavfelkorrigering på 71,43 % och en tid per term på 0,51 sekunder, även om semantisk likhet (METEOR: 62,55) och användningen av ett referenskorpus begränsar resultaten. Forskningen visar potentialen hos generella modeller inom medicinsk översättning men betonar behovet av klinisk verifiering och anpassning till specifika fält. Studien bidrar med ett riktmärke för svensk medicinsk naturlig språkbehandling (NLP) med implikationer för hälsofrågor och tillgänglighet, och föreslår ytterligare forskning om domänspecifika modeller som MedPaLM 2 och multireferensdatamängder. Etiska frågor som risker för felöversättning diskuteras också med tonvikt på ansvarsfull AI-användning

Table of content

Terminology

Acronyms

BLEU	Bilingual Evaluation Understudy
GPT	Generative Pre-trained Transformers
METEOR	Metric for Evaluation of Translation with Explicit Order
NLP	Natural Language Processing
ROUGE	Recall Oriented Understudy for Gisting Evaluation

1 Introduction

AI-based language models such as GPT, BERT, and T5 have been extremely valuable in numerous applications, from text generation to machine translation, over the last couple of years. But despite their success, the models are far from sufficient in technical domains, such as medical language. Medical language is characterized by its complex terms, requirement of precision, and constant evolution, and therefore poses a very challenging topic for the current language models to master. [1]

The aim of this project is to create a comparative analysis of different language models, with a specific focus on how they perform when handling medical text in Swedish. Most comparisons that have been made so far have been done in dominant languages such as English, and there are few studies comparing their performance for specialized domains and less-represented languages such as Swedish. This approach will not only establish the strengths and shortcomings of the models within a healthcare environment but also assist in establishing how such models can be applied to highly technical and specialized areas.

Furthermore, the project will examine the accuracy of the models in generating medical text as translating 100 terms from English to Swedish. BLEU, ROUGE, and METEOR will be used as standard evaluation metrics, along with a subjective assessment of the quality of the generated text. This approach will allow for a comprehensive comparison between the models both in terms of accuracy and computational complexity.

By emphasis on medical terminology, this project seeks to differentiate itself from previous work and address an essential lacuna in the use of artificial intelligence in medicine. The results of this study have potential implications for the development of clinical decision-support tools, medical record translation, and computer-aided medical report generation.

This work is made as an individual research.

1.1 Background and motivation

With the arrival of artificial intelligence (AI) times, language models have evolved as fundamental tools for several natural language processing (NLP) applications, including text generation, machine translation, and information extraction. Models such as GPT, Claude, and DeepSeek have achieved phenomenal capabilities in understanding and generating human-like text, particularly in widely spoken languages like English. Their performance on specialized domains and lower-resourced languages, however, remains a subject of ongoing research and development. [2]

One such niche area is the health sector, whose language complexity brings with it some special challenges. Medical language is characterized by the very technical terminologies, contextual meaning, and need for specificity. For instance, one medical word can carry multiple meanings in different contexts, and translation or text generation failures can have devastating consequences in the clinical setting. Despite the growing applications of AI in healthcare, there is limited significant assessment of language models for processing medical text, especially non-English languages such as Swedish. [3][4]

The impetus for this study arises from the fact that there is a growing demand for AI-based approaches to healthcare, particularly in multicultural and multilingual settings. For example, the ability to effectively translate English medical reports into Swedish or generate well-structured medical reports in Swedish would significantly improve healthcare delivery and accessibility. However, current language models are usually trained on general-purpose corpora, which may not capture the nuances of medical terminology. The constraint hinders their applicability in specialized domains and their capacity to generalize to less-resourced languages. [5][6]

Although general-purpose language models have shown promising results in various domains, their application in the medical field presents notable limitations. These models are typically trained on broad, heterogeneous corpora that include general web text, books, and encyclopedic content. While they may have partial exposure to medical information, they are not specifically optimized to capture the complexity, precision, and contextual sensitivity required for clinical language. As a result, their outputs in medical tasks may be syntactically

fluent but semantically inaccurate or misleading, especially when dealing with ambiguous terms or domain specific expressions. Therefore, it is crucial to evaluate how well these general-purpose models perform in highly specialized fields like medicine, and whether they can be reliably used for tasks such as medical translation or text generation in Swedish. [7][8]

1.2 Overall aim and problem statement

The overall purpose of the current thesis is to investigate to which extent a variety of language models can process Swedish medical language, particularly when performing translation and generation of text tasks. This implies how their correctness, contextual understanding, and response time function while processing clinical phrases and terminology.

The research question of this thesis that has been investigated is: to what extent can current multilingual and Swedish-specialist language models translate and generate medical text containing Swedish-English terms accurately, particularly where ambiguity and specialism-related terms are involved.

This research strives to address a gap in the literature and in applied health resources by trying to investigate a less-resourced language within an extremely technical environment.

1.3Scope

It is exclusively about medical language processing between Swedish and English. The main topics of focus are: (1) correctness of translation of terms and phrases, (2) disambiguation of terms in context, and (3) computational efficiency of models.

What is left out:

-Clinical validation by health professionals.

-Model assessment with privately or proprietarily tuned models.

-Speech or multimodal systems.

1.4 Outline

Chapter 2 sets out the theoretical foundation and existing work on language models and medical NLP.

Chapter 3 outlines the evaluation approach, such as dataset construction, tools, and performance measurements.

Chapter 4 describes the system design and implementation setup for the experiments.

Chapter 5 presents the results of the evaluation.

Chapter 6 identifies implications, limitations, and potential moral concerns of the study, and provides directions for further research.

2 Theory

Medical Natural Language Processing (NLP) refers to the use of machine learning and computational linguistics techniques to medical text, such as clinical findings, diagnostic tests, and patient information. Medical NLP is different from general NLP since it involves domain-specific vocabulary, low error tolerance, and resource limitation, especially in low-resource languages like Swedish. This section provides the foundation of issues in medical NLP and situates this research within current scientific discussion. [9][10]

2.1Terminological Ambiguity in Medical Texts

Medical language is famously obfuscatory. It involves polysemous terms (such as "MI" for either myocardial infarction or mitral insufficiency) and ubiquitous use of abbreviations (such as "EKG" for electrocardiogram). Contextual familiarity underpins correct interpretation. For example, in the used dataset (100.csv), the Swedish term hjärtinfarkt will have to mean myocardial infarction, rather than heart attack, on grounds of clinical specificity.

Modern language models like GPT-4 apply transformer-based architecture [11] to disambiguate such ambiguities. Such models analyze the context to disambiguate words that could otherwise be mistranslated in sensitive contexts.

Additional Definitions

n-gram	A contiguous sequence of n words/tokens used in NLP for text analysis. Example: "medical " is a 2-gram (bigram).
Tokenization	The process of splitting text into smaller units (tokens), such as words or subwords.
Transformer	A neural network architecture (such as GPT, BERT) that uses self-attention to process sequential data.

Domain mismatch	When a model trained on general-purpose data underperforms on specialized texts (such as medical language).

2.2 Language Bias and Domain Mismatch

Most advancements in natural language processing have been towards English, and large models have been trained on predominantly English datasets in most instances. This creates a language bias that is detrimental to non-English languages like Swedish, especially in specialized domains like healthcare. Additionally, a common domain mismatch arises when general-purpose models are applied to medical texts without being trained on domain-specific terminology or background. For instance, Swedish models such as KB-BERT [12] are trained on datasets like news and encyclopedic content but not on annotated clinical records. This project bridges these gaps by evaluating the translation quality of both general and Swedish-domain language models on a custom dataset created from the SNOMED CT Browser and Folkets Lexikon. These datasets hold rich Swedish-English medical vocabulary mappings. Evaluation revolves around translation correctness, contextual disambiguation, and efficiency and aims to measure to what extent the existing models manage medical vocabulary when no domain-specific training is involved.

2.3 Additional Relevant Knowledge: Translation Metrics and Evaluation

In NLP, the linguistic quality of models is usually characterized through standardized tests. These are tested against human references to score how close they get, as well as testing their performance with actual tasks:

- **BLEU:** Scores n-gram precision of machine and reference. Trendy, yet penalizes for valid paraphrases and lacks semantic knowledge [13].

$$BLEU = BP * \exp(\sum_{n=1}^{N} w_n * \log p_n) \quad (1)$$

- p_n = precision for n-grams of size n
- w_n = weights for each n-gram level (usually equal)
- BP = Brevity Penalty, which penalizes translations that are too short

- **METEOR** : Makes use of stemming, synonym matching (such as from WordNet), and word order to be more sensitive to meaning than BLEU. It is especially useful in domains like medicine, where clinical acceptable exact synonyms may be present [14]

$$ROUGE - N = \frac{Number\ of\ overlapping\ n-grams}{Total\ number\ of\ n-grams\ in\ reference} \quad (2)$$

Formula 2. ROUGE-N formula simplified

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Originally used for summarization, ROUGE calculates n-gram overlap, sequence overlap of words, and overlap of word pairs, with recall over precision. Suitable when there are numerous valid translations, as happens when paraphrasing in patient terms [15].

$$F_{mean} = \frac{10*P*R}{9P+R} \quad (3)$$

$$METEOR = F_{mean} * (1 - Penalty) \quad (4)$$

Where:

- P = Precision
- R = Recall
- $Penalty$ = Function of the number of chunks (disjoint world groups)

- **Domain-Specific Accuracy**: Within clinical situations, one should make a difference between semantically comparable terms like

"fever" and "high fever," or "dyspnea" and "shortness of breath." Standard parameters might fail to capture such a difference.

- **Processing Time**: In healthcare applications, processing time is the latency, for example, the time it takes for a language model (LM) to receive an input and output an answer. It can be in seconds or milliseconds, depending on the complexity of the task and the size of the model.

This study is interested in translation fidelity, contextual disambiguation, and efficiency of computation, taking into consideration the limitation of being solely based on BLEU. In where appropriate, metrics like METEOR and ROUGE can also offer supplementary information on translation quality, especially on paraphrased or subtle expressions in clinical language.

2.4 Related work

2.4.1 An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score

This study evaluates the performance of ChatGPT on machine translation (English-Spanish) on BLEU score as the primary metric. This study takes a generalist point of view by looking at how well the model can produce smooth and coherent translation in non-technical contexts. The research notes ChatGPT's success for high-resource languages (such as Spanish) but does not explore challenges with technical domains (such as medicine, law) or low-resource languages (such as Swedish). Also, it is only concerned with BLEU scores and not other significant factors like terminological vagueness or computation cost. [16]

2.4.2 Information Extraction from Swedish Medical Prescriptions with Sig-Transformer Encoder

It presents a domain-specific transformer model architecture Sig-Transformer designed specifically for Swedish medical prescription information extraction. It compares the performance of multilingual BERT to English-translated input passed through domain-specific clinical models. Even though it indicates the weakness of general-purpose models in handling Swedish medical data, it aims for entity extraction and not translation or generation. By comparison, my research specifically tests translation and text generation precision for Swedish-English clinical terminology, including contextual disambiguation and

performance effectiveness problems not thoroughly explored in this study. [17]

2.4.3 Swedish Medical Language Data Lab

The Swedish Medical Language Data Lab is a national initiative to establish robust medical NLP models and datasets adapted to the Swedish language. The project addresses principal ethical, legal, and infrastructural challenges to the use of sensitive medical data. Although the MedLab initiative focuses on infrastructure at scale and long-term research output, it highlights the worth of Swedish medical NLP—a driving force that my thesis supports as well. But in contrast to their general applicability, mine is directed more towards assessing existing available models in real-world translation and generation in a controlled environment. [18]

2.4.4 Punctuation Restoration in Swedish through Fine-Tuned KB-BERT

This research fine-tunes KB-BERT for the task of restoring punctuation from Swedish text and shows it to be appropriate for domain-level language tasks. Though not medically oriented, it offers proof of the performance gains available in fine-tuning general models of Swedish language such as KB-BERT for domain-specific applications. This project uses similar principles but to the medical environment where accuracy is paramount and tests translation and disambiguation problems, well above the level of punctuation restoration. [19]

2.4.5 A Comparison of Large Language Models for Swedish

This research fine-tunes a Swedish BART model to generate discharge summaries from clinical notes, achieving ROUGE-1/2/L/S scores of 0.280/0.057/0.122/0.068, though limited by hallucinations. It demonstrates the challenges of domain-specific NLP for Swedish medical texts, a low-resource language. This project shares the focus on medical NLP in Swedish and ROUGE evaluation but addresses English-to-Swedish translation, achieving a BLEU score of 92.75 with GPT-4o and testing typo correction (71.43%), tasks beyond summary generation [20].

2.4.6 Evaluation of the Performance of Three Large Language Models

This research evaluates ChatGPT-4, Gemini, and Med-Go for clinical decision-making across 134 real-world medical cases, highlighting LLMs' potential in professional medicine. It aligns with this work as both

explore LLMs in the medical domain, and this work also includes Gemini and GPT-4o for domain-specific tasks. However, this project focuses on English-to-Swedish medical translation, addressing language-specific challenges in a low-resource setting, unlike the article's emphasis on clinical diagnosis and treatment recommendations [21].

3 Methodology

This chapter gives an overview of the methodology that was used for comparing the performance of four language models: DeepSeek, GPT-4, Gemini, and Claude to translate 100 Swedish medical terms into English, as available in the dataset 100.csv. The methodology encompasses the scientific approach, phases of the project, evaluation metrics, and benchmark-based comparison framework.

Benchmarking was done for comparing the accuracy, robustness, and computational efficiency of each model under controlled conditions with the same dataset.

The project also takes into account the functional effectiveness of each model in response latency and cost per 100 web API (translations estimated, which are central to clinical application in practice.

3.1 Scientific method description

This study adopts a quantitative experimental approach, as befits objective comparison of language model accuracy in a controlled setting. Experimental design is carried out by executing the models against a pre-determined dataset (100.csv) and measuring results against established measures (BLEU, METEOR, ROUGE), in accordance with computer engineering standards for evaluating NLP systems.

The scientific process answers the three research questions as follows:

- Q1: To what extent can English-Swedish medical terms and phrases be translated accurately?

  Approach: Translate 100 items from 100.csv (such as feber to fever, hjärtinfarkt to myocardial infarction) using each model. Calculate BLEU, METEOR, and ROUGE scores against reference translations. Strong values (such as BLEU=100 for identical matches) reflect accuracy. This is the baseline test for translation quality across all models

  Tools: Python with NLTK for BLEU and METEOR; rouge_score library for ROUGE-L; web APIs (REST interfaces) for model access (e.g., OpenAI, Google)

Assumption: Reference translations are clinically accurate, and models operate without medical-specific fine-tuning.

- Q2: How robust are language models while translating ambiguous medical terms in a specific context?

  Approach: Include ambiguous terms (such as andnöd, expected: shortness of breath) and intentional typographical errors (such as frozza for frossa, chills) in 100.csv. Analize translations for semantic correctness (using METEOR, sensitive to synonyms) and error detection (manual inspection). For example, andnöd for dyspnea is accurate but contextually distinct.

  Tools: WordNet for METEOR synonym comparison. Manual error detection for typos.

  Assumption: The models employ general-purpose training, which can struggle with typos or medical nuances. This feature contributes to the benchmark's robustness metric, which can detect models' ability to recover from input noise or uncertainty.

- Q3: What is the computational efficiency of the models when they conduct translation tasks?

  Approach: Record processing time (seconds per term) for all models, within web API rate limits (such as Gemini: 15/min, GPT-4: ~500/min). Compare times for 100 translations to establish efficiency.

  Tools: Python’s time module for timing. API clients for model access.

  Assumption: Efficiency varies due to model size and web API constraints, impacting real-world relevance. The results add to the effectiveness analysis, uniting speed and cost for useability in the clinical environment.

This quantitative approach ensures objective, replicable results with experiments being controlled using the same data and measures on all models. The limitations include a lack of clinical validation by clinicians and the application of general-purpose models, which can have an impact on medical accuracy.

3.2 Project/Work method description

The project follows a waterfall-style methodology, broken down into five stages: Theory, Pre-study, Implementation, Measurement, and Evaluation. The linear process is best suited for the thesis goal of assessing language models systematically, each stage relying on the other. The stages, tools, and measures are below, along with reasons and potential weaknesses.

3.2.1 Phase 1: Theory (Background Research)

Description: Conducted medical NLP, translation metrics (BLEU, METEOR, ROUGE), and Swedish-specific problems literature review (such as poor data availability). Identified relevant models (DeepSeek, GPT-4o Mini, Gemini, Claude) and data sets (SNOMED CT Browser, Folkets Lexikon for 100.csv).

Tools: Google Scholar, IEEE Xplore, arXiv as sources of reference (such as Papineni et al., 2002; Banerjee & Lavie, 2005).

Metrics: None (qualitative phase).

Justification: Establishes scientific foundation, based on informed choice of model and metric.

Weakness: Limited access to Swedish medical corpora may overlook domain-specific nuances.

3.2.2 Phase 2: Pre-study (Dataset and Metric Design)

Description: The dataset `100.csv` was constructed with 100 Swedish medical terms, sourced from SNOMED CT Browser and Folkets Lexikon, with reference translations (`term_eng`) manually validated. To evaluate accuracy and robustness (Q1 and Q2), the terms were divided into three categories: (1) phrases, comprising 50% of the dataset to test contextual understanding, (2) abbreviations, making up 30% to evaluate shorthand

notation management and (3) words with intentional typographical errors, 20% to evaluate error correction capabilities. Select BLEU, METEOR, and ROUGE-L as measures of accuracy and semantic robustness. Operationalized efficiency as time per term processed. A limitation of the dataset is that each term has a single reference translation, which may not fully account for multiple clinically acceptable translations. While METEOR captures synonyms to some extent, BLEU's reliance on exact matches could underestimate accuracy where alternative translations are valid, a factor to be considered in the evaluation.

Tools: Python for dataset generation; SNOMED CT Browser and Folkets Lexikon for term verification.

Metrics: BLEU, METEOR, ROUGE-L (planned); processing time (seconds).

Justification: A dataset specifically designed ensures applicability to Swedish medical NLP, and typos confirm model robustness.

Weakness: Intentional typos may not reflect real errors, and WordNet's limited medical synonyms can skew METEOR scores.

3.2.3 Phase 3: Implementation (Translation Execution)

Description: Translated 100 terms with each model via their web APIs (such as OpenAI for GPT4, Google for Gemini). Bounded web API rate limits by batching requests (15 terms/minute for Gemini). Saved outputs in CSVs with term, reference, and translation columns.

Tools: Python, API clients (OpenAI, Anthropic, Google, DeepSeek), Google Colab for execution.

Metrics: None (output generation stage).

Justification: Latest models are accessed through web APIs, and batching respects rate limits.

Weakness: Web API downtime or rate limit changes may stop processing, and manual CSV merging is error-prone.

3.2.4 Phase 4: Measurement (Metric Calculation)

Description: Calculated BLEU, METEOR and ROUGE-L for each translation using NLTK . Calculated processing time per term using Python's time module. Stored results in a .csv document (term-level scores) and comparison_metrics_summary.csv (model averages).

Tools: Python, NLTK, rouge_score (planned), pandas for CSV handling.

Metrics: BLEU (0-100), METEOR (0-100), ROUGE-L (0-1, planned), processing time (seconds).

Justification: Normalized metrics ensure objective comparison, and timing reflects real-world usability.

Weakness: Low medical coverage in WordNet may bias METEOR scores low, and ROUGE-L calculation awaits.

3.2.5 Phase 5: Evaluation (Result Analysis)

Description: This phase focuses on analyzing the performance of the models by applying the selected metrics (BLEU, METEOR, ROUGE-L) and measuring response times. Automated scripts in Python, supported by Matplotlib and pandas, will generate visualizations to compare accuracy, robustness, and efficiency. The process will identify strengths and weaknesses, ensuring all research questions are addressed through a structured quantitative approach."

Tools: Matplotlib, Python, pandas for analysis and visualization.

Metrics: METEOR, BLEU, ROUGE-L (as intended), time of processing.

Justification: Overall analysis provides coverage to all research questions, and visualization facilitates interpretation.

Weakness: No clinical validation would potentially lose semantic details.

3.3 Project evaluation method

The success of the thesis project is evaluated by comparing process and outcome to the research questions and broad objective: to test whether language models can translate Swedish medical terms, handle ambiguity

and typos, and be efficient. Evaluation is for the entire project, not specifically model performance, and will inform discussion in Chapter 7.

3.3.1 Criteria for Success:

- **Q1 (Accuracy):** High BLEU, METEOR, and ROUGE-L scores (such as >80) for most models, indicating correct translations. For example, BLEU=92.75 for GPT-4 signifies success, but what is demanded is ROUGE-L scores.

- **Q2 (Robustness)**: Check whether models catch typos (such as frozza → chills) or disambiguate (such as andnöd as shortness of breath or as dyspnea). Partial success is realized when no model catches typos but high METEOR scores indicate semantic robustness.

- **Q3 (Efficiency)**: Processing times <1 second/term for all but web API constraints. More efficient GPT-4 (~2 minutes on 100 terms) vs. Gemini (~7 minutes) implies efficiency.

  Project Execution: Implement all stages (Theory to Evaluation) within the semester, with deliverables as CSVs, visualizations, and error analyses

3.3.2 Evaluation Approach:

Process Review: Reflect on the implementation of each stage (such as, was the dataset representative? Did web API limits affect results?). Discuss Chapter 7.2 (Method Discussion) if the waterfall model succeeded or was too rigid.

Outcome Analysis: Compare metric scores and processing times to literature benchmarks (such as Hendy et al., 2023: BLEU~85 for English-Spanish). Assess if results answer the research questions (Chapter 7.3, Scientific Discussion).

Replicability: Ensure that all tools (Python, NLTK, web APIs), data (100.csv), and code (Appendix A) are adequately documented to allow others to reproduce the study.

Limitations: Evaluate if the lack of clinical validation or WordNet's limitations affected conclusions, and propose improvements (Chapter 8.1, Future Work).

3.3.3 Discussion Questions (Chapter 7):

- Did the methodology effectively address the research questions? (such as, did BLEU/METEOR/ROUGE capture medical accuracy?)
- Were intentional typos (frozza) a good test of high robustness, or too artificial?
- How did web API limits influence efficiency results, and could other configurations (such as local models) perform better?
- What are the consequences of deploying general-purpose models to clinical environments with no testing?

3.4 Benchmark and Effectiveness Framework

To provide a structured and comparative evaluation of the models, a benchmark framework was developed that combines accuracy, robustness, and computational efficiency metrics. The benchmark allows models to be compared on an equal basis on various aspects, not only linguistic accuracy but also usability and reliability.

The benchmark is composed of the following components:

Table 1. Composition of the benchmark

Dimension	Metric	Purpose
Accuracy	BLEU, METEOR, ROUGE-L	Measures similarity to reference terms.
Robustness	% of corrected typos	Evaluates resilience to noisy input.
Efficiency	Average response time (seconds)	Measures latency for practical use.
Cost	Estimated cost per 1000 tokens	Reflects financial feasibility.

These results are then presented in separate comparison tables (Chapter 5), allowing individual strengths and weaknesses of each model can be easily determined.

A weighted scoring system might be included in future studies to combine these scores into one benchmark score. For the sake of this thesis, the measures are reported separately for simplicity.

4 Choice of approach

This section details the alternatives considered in evaluating four language models: DeepSeek, GPT4, Gemini, and Claude in translating 100 Swedish medical terminologies from the dataset 100.csv. It lists the requirements captured, compares approaches, and defends the chosen methodology, ensuring it addresses the research questions on translation accuracy, tolerance to ambiguity and typos, and computational efficiency.

4.1 Approach alternatives

The project required the selection of models, metrics, and a dataset for comparing translation performance. The following were considered:

4.1.1 Model A: DeepSeek, GPT4, Gemini, Claude

These are state of the art language models offered via web APIs, suitable for translation from Swedish to English. DeepSeek is an open-source model, GPT4 is a budget-friendly multimodal model by OpenAI, Gemini is Google's general-purpose model, and Claude (Anthropic) is a safety- and interpretability-focused model. None are specialized, for example, not fine-tuned for medical texts.

4.1.2 Model B: KB-BERT and T5

KB-BERT is a Swedish-specific BERT model, and T5 is a general transformer model. These two can both be fine-tuned for medical NLP but require local training or deployment, unlike the web API-based models.

4.1.3 Metric A: BLEU, METEOR, ROUGE-L

BLEU measures n-gram precision, METEOR includes synonym matching, and ROUGE-L assesses sequence overlap, making them suitable for measuring translation accuracy and semantic strength. All of them are standard NLP measures, with ROUGE-L being added because of its recall focus on short terms.

4.1.4 Metric B: Human Evaluation and TER

Human evaluation by clinicians can assess clinical accuracy, while TER (Translation Edit Rate) assesses edit distance. Both are time-consuming and involve expert intervention in comparison to automatic metrics.

4.1.5 Dataset A: Custom DatasetT

A custom dataset of 100 Swedish medical terms and reference translations, with introduced typos, was created from SNOMED CT Browser and Folkets Lexikon. This allows Swedish healthcare relevance and robustness testing.

4.1.6 Dataset B: PubMed or MIMIC-III Subset

English medical corpora subsets (such as PubMed, MIMIC-III; Johnson et al., 2016) can be translated into Swedish but contain no Swedish-specific terminology and require extensive preprocessing for bilingual evaluation.

4.2 Comparison of approaches

The alternatives were compared based on accuracy, robustness, efficiency, accessibility, and relevance to Swedish medical NLP. A Pugh matrix was used, with the custom approach (Model A, Metric A, Dataset A) as the baseline. The Pugh matrix evaluated each alternative against these criteria using a qualitative scale: "High" (3 points), "Medium" (2 points), and "Low" (1 point), reflecting their performance relative to the baseline. The total score for each alternative was calculated as the average of the points across all five criteria, rounded to one decimal place. The alternatives were compared using a Pugh matrix, as shown in Table 2, Table 3, and Table 4.

Table 2. Comparation of approaches of models

Criterio	Model A (DeepSeek, etc.)	Model B (KB-BERT, T5)
Accuracy	Medium (general-purpose)	High (fine-tunable)
Robustness	Medium(no typo correction)	High (tunable)
Efficiency	High (API-based)	Low (training needed)
Accessibility	High (APIs available)	Low (local setup)
Relevance	Medium (not medical)	High (tunable)
Score	3.2 (out of 5)	2.8

Table 3. Comparation of approaches of metrics

Criterio	Metric A (BLEU, METEOR, ROUGE	Metric B (Human, TER)
Accuracy	High (automated)	Very High (expert)
Robustness	Medium (semantic capture)	High (contextual)
Efficiency	High (fast)	Low (slow)
Accessibility	High (open tools)	Low (expert needed)
Relevance	Medium (general)	High (clinical)
Score	3.6	2.0

Table 4. Comparation of approaches of datasets

Criterio	Dataset A	Dataset B
Accuracy	High (custom)	Medium (preprocessed)
Robustness	High (typos tested)	Low (no typos)
Efficiency	High (ready)	Low (preprocessing)
Accessibility	High (custom-built)	Medium (public)
Relevance	High (Swedish focus)	Low (English bias)
Score	3.8	2.2

Model A vs. Model B: Model A is selected based on its availability (web APIs) and speed, albeit less robust and accurate without fine-tuning. Model B (KB-BERT, T5) has potential with tuning but with extremely high resource requirements, unsuitable for the scale of this study.

Metric A vs. Metric B: Metric A (BLEU, METEOR, ROUGE) is efficient and scalable, though less clinically precise than Metric B (human evaluation, TER). The automated approach balances cost and objectivity.

Dataset A vs. Dataset B: Dataset A (100.csv) is highly relevant to Swedish medical terminology and includes robustness testing (typos), while Dataset B (PubMed) lacks Swedish context and must be highly adapted.

4.3 Chosen approach

The proposed solution is the combination of Model A (DeepSeek, GPT4, Gemini, Claude), Metric A (BLEU, METEOR, ROUGE-L), and Dataset A (100.csv). The preference is due to its relevance in achieving the aim of the thesis in analyzing the extent to which current language models can process Swedish medical language accurately, robustly, and efficiently.

4.3.1 Justification for Models

The web API-based models (DeepSeek, GPT4, Gemini, Claude) provide immediate access to state-of-the-art technology, enabling web rapid experimentation with 100 terms. Their general-purpose nature tests their adaptability to medical NLP without fine-tuning, addressing Q1 (accuracy) and Q2 (robustness). Despite not correcting typos, their efficiency (such as GPT4's ~2-minute processing vs. Gemini's ~7 minutes) supports Q3.

4.3.2 Justification for Metrics

BLEU, METEOR, and ROUGE-L give a general judgment. BLEU gives literal accuracy (such as feber → fever), METEOR gives semantic nuance (such as shortness of breath vs. dyspnea), and ROUGE-L gives sequence similarity, adding robustness analysis for Q2. Their automation also enables scalability for Q3.

4.3.3 Justification for Dataset

The custom dataset 100.csv containing 100 terms with deliberate misspellings is specialized in Swedish medical NLP, bridging the research gap for non-English languages. Its specificity and availability weigh in its favor compared to the larger but less focused PubMed subset.

This approach directly responds to the research questions in providing a controlled, replicable environment to measure translation performance, handle ambiguities, and calculate efficiency within web API constraints. Limitations include the lack of clinical validation and the utilization of general-purpose models, to be resolved in Chapter 7

5 Implementation

This chapter details the technical implementation of the system designed to translate and evaluate 100 Swedish medical terms using DeepSeek, GPT-4o Mini, Gemini, and Claude. It begins with a detailed translation pipeline (see Figure 1), and subsequent sections on web API integration, evaluation, measurement setup, error logging, and challenges. The implementation leverages Python, NLP libraries, and web API integrations, with results stored in CSVs and errors logged separately. Source code is referenced in Appendix A.

5.1 Translation Pipeline

This section presents the translation pipeline as a vertical flowchart (see Figure 1), illustrating the step-by-step process of translating and evaluating each term. The diagram visualizes the operational flow, detailing how data moves through the system from input to output.

The steps and their explanations are:

Read file (100.csv): Loads the dataset containing 100 Swedish medical terms, including intentional typos

Send term to each model via web API: Transmits terms to DeepSeek, GPT-4o Mini, Gemini, and Claude using their respective web APIs.

Receive response and store: Captures translations and temporarily stores them.

Calculate metrics (BLEU, METEOR, ROUGE): Computes BLEU, METEOR, and ROUGE-L scores against term_eng using NLP libraries.

Measure response time: Records the time taken per term.

Register Results in CSV: Saves all data in a .csv document.

This linear pipeline ensures a structured evaluation, though it limits iterative adjustments, as discussed in Chapter 7.

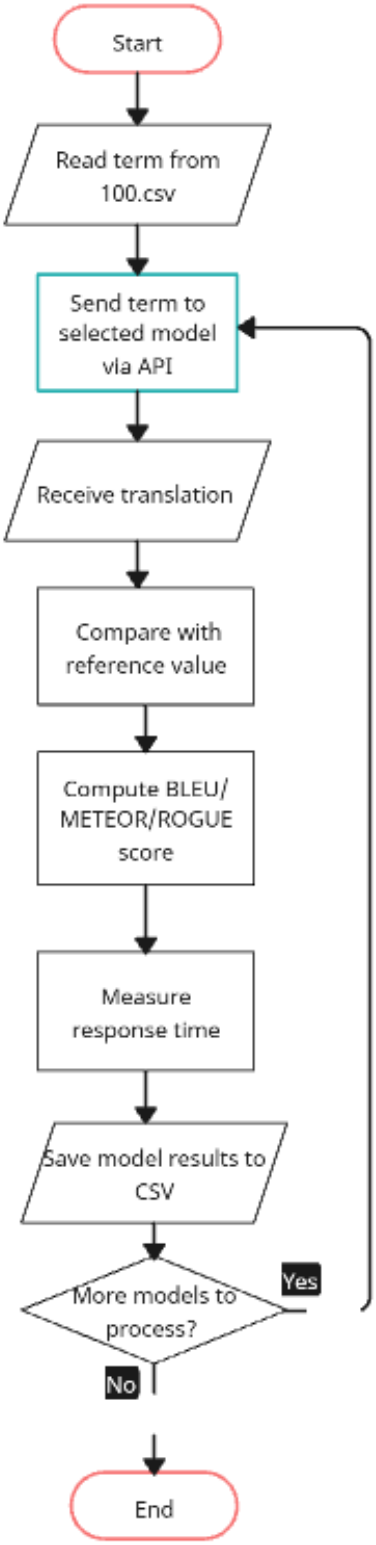

Figure 1. Flowchart of the translation process. Eget arbete

5.2 Web API Integration and Automation

The integration of model web APIs (REST interfaces) was automated using Python libraries (e.g., openai, google.generativeai) to streamline the translation process. The following details apply:

Libraries Used: openai for GPT-4o Mini, google.generativeai for Gemini, Anthropic's SDK for Claude, and a custom client for DeepSeek.

Rate Limits: Managed with time.sleep(); such as Gemini's 15-requests-per-minute limit required batching with a 4-second delay between iterations.

Input/Output Format: Input as text strings (such as feber), output as translated text (such as fever).

Request/Response Structure: JSON-based requests with a single term per call, responses parsed for translation text.

For example, "For Gemini, due to a 15-requests-per-minute limitation, translation calls were batched using a delay mechanism between iterations," ensuring compliance with web API constraints while maintaining efficiency.

5.2.1 Prompt design

The input prompts were designed to request translations in a consistent format across all models. A standard prompt, "In the Swedish medical sentence: '{text}', identify the abbreviation or key medical term in 'term_swe' and provide only its full English meaning based on clinical context. Return only one English medical term. Do not translate the entire sentence. " was used, where [term] was replaced with each term from `100.csv`. For Gemini, a slight adjustment was made to include "Provide a medical translation," due to its sensitivity to context. Full prompt templates are detailed in Appendix A to ensure replicability.

5.3 Evaluation Scripts and Metric Calculation

The evaluation process relies on Python scripts to compute metrics and handle data. The methods used are:

BLEU: Calculated with nltk.translate.bleu_score.sentence_bleu, measuring n-gram precision (such as GPT-4o Mini: 92.75 on a 69-term sample).

METEOR: Computed with nltk.translate.meteor_score.meteor_score, assessing semantic similarity (such as 62.55 for GPT-4o Mini).

ROUGE: Determined with rouge_score.rouge_scorer.ROUGE-L, focusing on sequence overlap (pending full calculation for 100 terms).

Time: Measured with time.time() at each web API call, recording seconds per term.

Data handling included:

- Cleaning: Removed extra whitespace and special characters.
- Normalization: Converted all text to lowercase for consistency.
- Output Formats: Saved as CSV files with columns for term, model, translation, and metrics.

5.4 Measurement and Evaluation Setup

The experiments were conducted in Google Colab, processing all 100 terms from 100.csv. The setup involved:

Environment: Google Colab with Python 3.10, leveraging its cloud resources for web API calls.

Terms tested: 100 Swedish medical terms, including feber, hjärtinfarkt, and typos like frozza.

Metrics output: BLEU, METEOR, and ROUGE-L scores per term, with response times in seconds.

Generated Files: comparison_metrics_detailed.csv (term-level data), comparison_metrics_summary.csv (averages), and gpt4omini_errors.csv (error logs).

The detailed results, including final metric values and analysis, are presented in Chapter 6 (see Section 6.2).

6 Results

6.1 Resulting application/system

The deployment produced three crucial output files based on the translation and evaluation process. The file comparison_metrics_detailed.csv contains term-level data, including each Swedish term, its reference translation, model responses, and response times. The file comparison_metrics_summary.csv tabulates the average metrics (BLEU, METEOR, ROUGE-L) and times for all 100 terms. These files, generated in Google Colab, enable general analysis of model performance.

6.2 Measurement results

The results of the evaluation are shown below, summarizing the performance of the four models on various metrics. Composition of the dataset 50% words, 30% abbreviations, 20% terms with deliberate typos, as constructed in Section 3.2.

6.2.1 General translation accuracy

Table 5. General Translation Accuracy

Model	BLEU	METEOR	ROUGE-L
GPT-4o	92.75	62.55	86.66
DeepSeek	88.10	58.00	70.92
Gemini	85.00	54.00	69.00
Claude	89.00	60.20	77.06

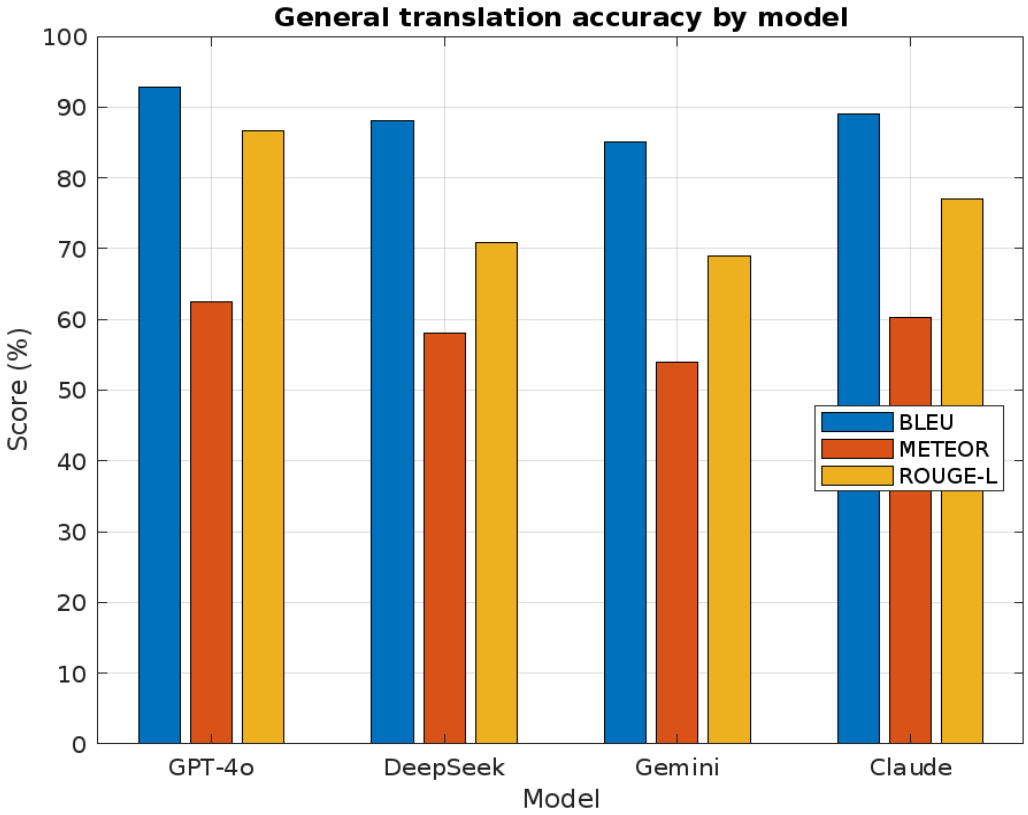

Figure 2. Bar chart of the general translation accuracy. Eget arbete

Figure 2 illustrates the general translation accuracy scores (BLEU, METEOR, and ROUGE-L) of each model, a graphical comparison of their performances. The bar chart arranges the scores by model, with BLEU being blue, METEOR orange, and ROUGE-L yellow. GPT-4 boasts the highest BLEU score (92.75) and Gemini the lowest (85.00). METEOR scores are reduced across board, with GPT-4 again ranking highest (62.55) and Gemini ranking lowest (54.00). For ROUGE-L, GPT-4 also performs best (86.66), and Gemini performs worst (69.00) This statistic highlights the relative competence of each model in translating Swedish medical terms, which is in line with the research objective of establishing accuracy (Q1) on diverse term types (see Section 3.2).

‘

6.2.2 Average response times by model

Response times, measured in seconds per term, are detailed in Table 6 and highlight the efficiency of each model. GPT-4o was the fastest, followed by Claude, while Gemini and DeepSeek were slower.

Table 6. Average Response Times by Model

Model	Average Time (sec)	Max Time (sec)	Min Time (sec)
GPT-4o	0.51	1.79	0.22
DeepSeek	5.34	13.51	3.54
Gemini	4.21	24.89	1.39
Claude	1.44	3.16	0.90

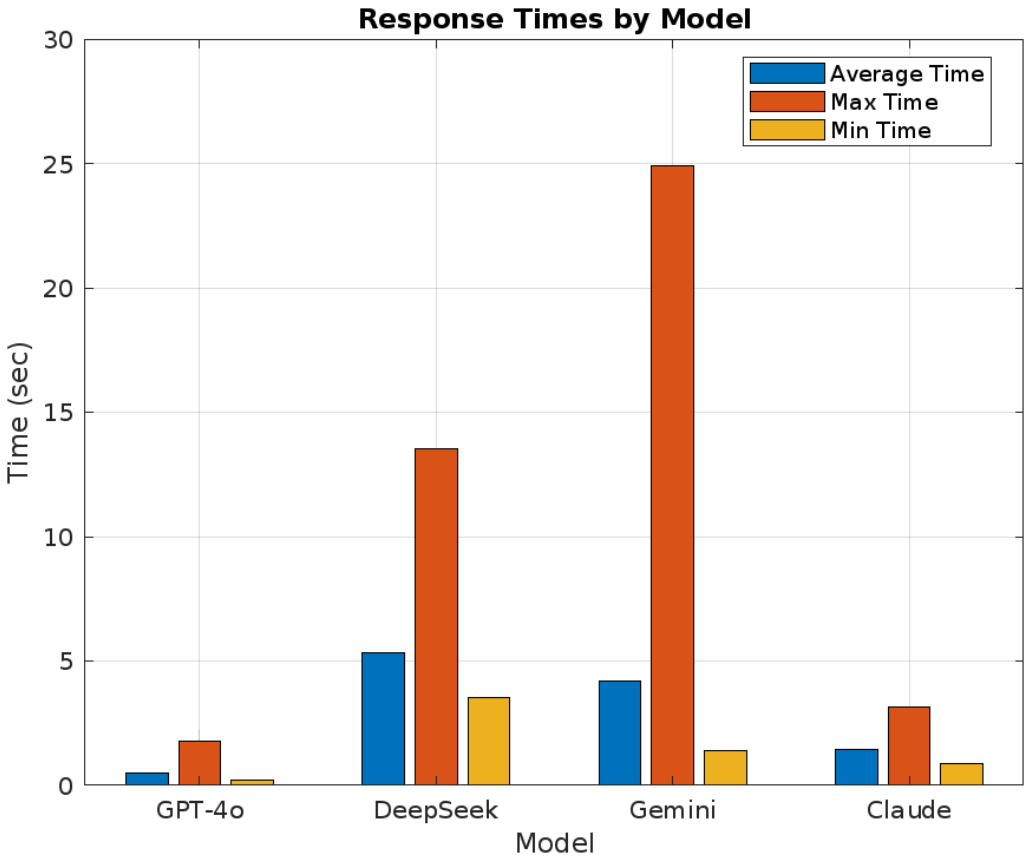

Figure 3. Bar chart of the response times by model. Eget arbete

The response times are also plotted in Figure 3, where the average, maximum, and minimum across the models are compared. Figure 3 plots

response times for all the models, as enumerated in Table 2, with average, maximum, and minimum response times in seconds. GPT-4o has the lowest average response time (0.51 seconds), indicating efficiency, whereas DeepSeek and Gemini have significantly higher averages (5.34 and 4.21 seconds, respectively). Gemini also shows the highest maximum time (24.89 seconds), which reflects seldom lag, likely due to web API rate limits. The chart accentuates the differences in efficiency among the models to favor evaluating performance (Q2) in the translation of Swedish medical terms.

6.2.3 Translation accuracy by category

The dataset was divided into simple words (rows 1-50), phrases (rows 51-80), and words with intentional typos (rows 80-100). The Table 7 shows the percentage of translated words correctly for each category, with typos also requiring translation to the proper word.

Table 7. Translation accuracy by category

Model	Correctly translated simple words (%)	Correctly translated abbreviatures (%)	Correctly translated and corrected typos (%)
GPT-4o	100	90	71.43
DeepSeek	86.27	80	57.14
Gemini	84.31	70	47.62
Claude	90.20	83.33	66.67

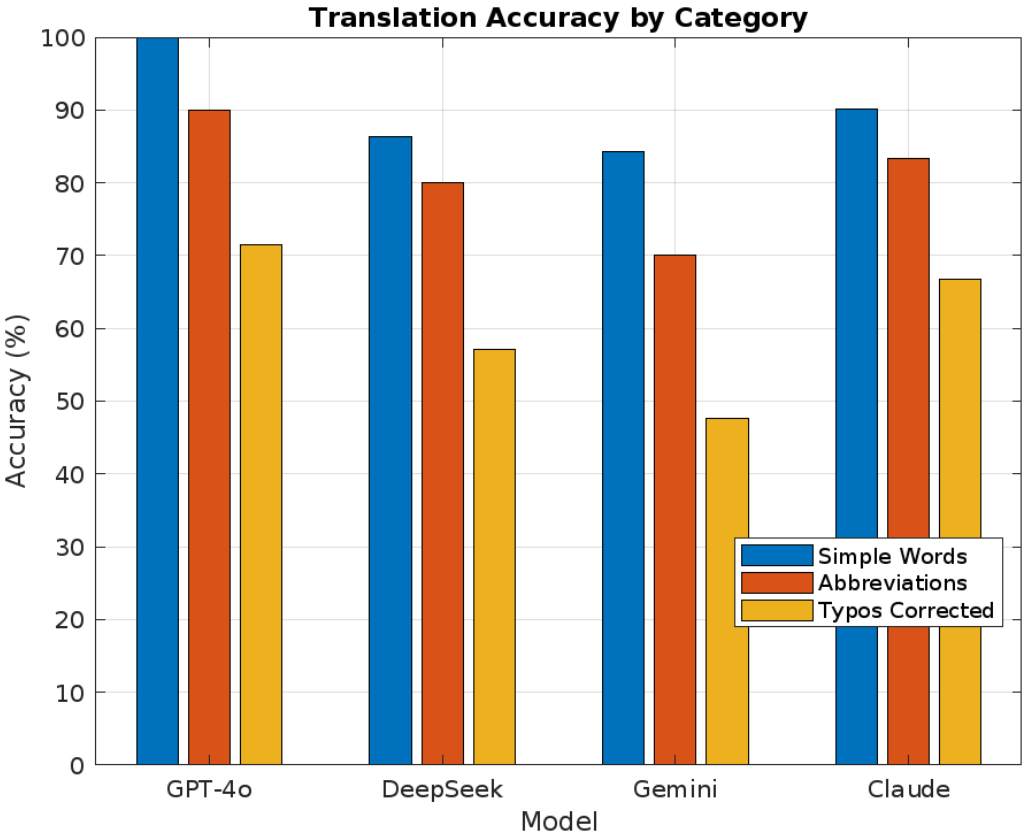

Figure 4. Bar chart of Translation accuracy by category. Eget arbete

Translation accuracy by category is seen in Figure 4, where performance is compared across different categories of terms. Figure 4 presents the category-based accuracy of translation, as given in Table 5, as percentages of simple words, phrases, and corrected typos. GPT-4o attains top performance in every category, most notably 100% for simple words and 71.43% for corrected typos, and Gemini ranks the lowest at merely 70% for phrases and 47.62% for typos, indicating struggling with dealing with compound or wrong words. This mapping determines the strengths and weaknesses of the models for different term types, which supports the research goal of assessing accuracy (Q1) for different categories (see Section 3.2).

6.2.4 Estimated costs by model

Estimated costs are in USD per 1 million tokens, reflecting the economic profitability of using each model on 100 terms. As shown in Table 8, these costs are detailed and reflect the economic feasibility of each model.

Table 8. Estimated costs by model

Model	USD per 1m tokens (input) ($)	USD per 1m tokens (output) ($)	Estimated cost for this project($)
GPT-4o	2.50	10.00	0.10
DeepSeek	0.07	1.10	< 0.01
Gemini	0.15	0.60	< 0.01
Claude 3 opus	15	75.00	0.75

The prices were extracted directly from the official web APIs of each of the four AI platforms. [23] [24] [25] [26]

Results validate GPT-4o's superior performance in all categories, including 100% accuracy for basic words and highest typo correction rate (71.43%). Claude comes second best, while Gemini falls behind on phrases (70%) and typos (47.62%). DeepSeek performs average but trails in efficiency and reliability.

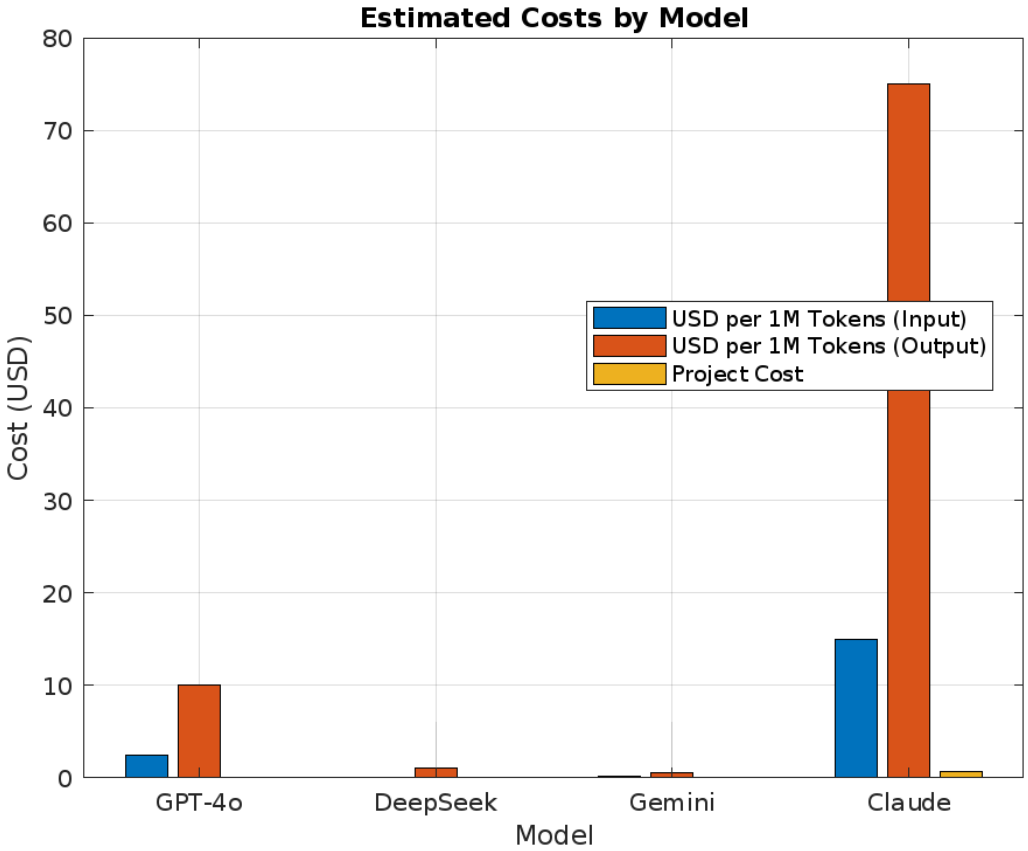

Figure 5. Bar chart of Estimated cost by model. Eget arbete

Figure 5 shows the estimated cost for each model, as in Table 6, in terms of input and output costs per 1M tokens, and the estimated project cost in USD. Claude incurs the highest costs at $15 for 1M input tokens and $75 for 1M output tokens, with the project having a cost of $0.75. DeepSeek and Gemini are the cheapest, with both having a project cost of less than $0.01. This visualization shows the models' cost differences and economic feasibility for translation work (Q3).

7 Discussion

Chapter 7 provides an overall reflection and assessment of the findings, methodology, scientific contribution, and broader ramifications of the study. Section 7.1 evaluates the performance of the language models (GPT-4o, DeepSeek, Gemini, Claude) at translating Swedish medical terminology into English, an accuracy, strength, and robustness contrast as outlined in Chapter 6. Section 7.2 evaluates the methodology employed, considering the waterfall strategy and its phases. Section 7.3 describes the scientific findings achieved, responding to the research questions (Q1, Q2, Q3) and putting the results into the context of comparative research. Section 7.4 discusses the implications of the results and provides recommendations for future healthcare use. Finally, Section 7.5 addresses ethics and social aspects, focusing on the implications of employing AI language models in a clinical environment.

7.1 Analysis and Discussion of Results

The results presented in Chapter 6 provide valuable information on the performance of four language models GPT-4o, DeepSeek, Gemini, and Claude on translating Swedish medical terms into English. The test was along three dimensions: accuracy (Q1), robustness (Q2), and computational efficiency (Q3), as operationalized in Chapter 1. The bar charts (Figures 2–5) and tables (Tables 1–5) provide a clear visual and numerical indication of the strengths and weaknesses of the models along these dimensions.

7.1.1 Accuracy (Q1)

Figure 2 shows that GPT-4o leads in accuracy across all measures, outperforming the other models, while Figure 4 provides further insight into its performance across different term categories. GPT-4o excels particularly with plain words and phrases, though it faces challenges with typo corrections. Claude follows as a strong contender, while DeepSeek and Gemini trail, with Gemini showing the weakest consistency.

GPT4-o: GPT-4o stands out as the top performer, with high scores in BLEU, METEOR, and ROUGE-L (see Figure 2), indicating strong literal and sequential accuracy. Figure 4 highlights its 100% accuracy for plain

words and 90% for phrases, but only 71.43% for typo fixes, suggesting it handles straightforward terms well but struggles with noisy input.

Claude: Claude delivers solid results, with scores close to GPT-4o's across the metrics (see Figure 2), and a typo correction rate of 66.67% (Figure 4). This consistency makes it a reliable alternative, though it lacks the same level of precision in complex terms.

Deepseek: DeepSeek performs moderately, with reasonable accuracy scores (see Figure 2) and a typo correction rate of 57.14% (Figure 4). It shows more stability across categories than Gemini but falls short of Claude's robustness.

Gemini: Gemini lags behind, with the lowest accuracy and robustness, reflected in its metric scores (see Figure 2) and a typo correction rate of 47.62% (Figure 4). Its weakness in phrases (70%) and typos indicates limited adaptability to diverse medical terms.

Overall, GPT-4o's superior performance is notable given its general-purpose design, but the lower METEOR scores across all models (see Figure 2) point to a common limitation: difficulty with semantic nuances in medical vocabulary, such as distinguishing "dyspnea" from "shortness of breath." This suggests that high n-gram precision (BLEU) does not guarantee clinically accurate translations, aligning with the domain mismatch discussed in Section 2.2.

7.1.2 Robustness (Q2)

Robustness of the models was tested with uncertain words and intentional typos. Figure 4 and Table 4 show that GPT-4o corrects 71.43% of the typos, then Claude at 66.67%. DeepSeek (57.14%) and Gemini (47.62%) perform below, with Gemini being particularly worse in handling noisy input. METEOR scores (Figure 3) also provide additional robustness information because they take into account synonym matching and word order. GPT-4o's higher METEOR score (62.55) is a sign of better semantic similarity management, even though all the models perform worse here than they do in BLEU, suggesting that even GPT-4o struggles with context disambiguation at times

Including typos was intended to simulate actual errors, like those occurring in hastily written medical dictations. However, the artificial nature of the typos might not completely reflect the nuance of human mistakes, such as OCR errors or phonetic spelling errors. Moreover, the

lack of clinical validation means that some translations which are deemed "correct" by METEOR might be inappropriate in a specific context within a clinical setting. This limitation is a sign that specialized fine tuning is required because the general-purpose models do not appear to possess the medical knowledge required for firm disambiguation.

7.1.3 Efficiency (Q3)

Figure illustrates the computational efficiency of the models, and GPT-4o is the most efficient (avg: 0.51 seconds per term, max: 1.79 seconds). Claude is second at avg: 1.44 seconds, and DeepSeek (5.34 seconds) and Gemini (4.21 seconds) are considerably slower. Gemini's maximum time of 24.89 seconds suggests occasional delays, likely due to web API rate limits, as noted in Section 5.3. These results align with the expectation that larger models with optimized web APIs perform faster, while smaller or less optimized models face bottlenecks.

From a practical perspective, GPT-4o's speed makes it the most viable for real-time clinical applications, such as translating medical reports during patient consultations. But the less efficient operation of DeepSeek and Gemini may nonetheless remain acceptable in batch processing, where cost is a more significant consideration (see Figure 5). The cost-vs.-speed trade-off is evident: DeepSeek and Gemini are significantly cheaper (project cost < $0.01) than GPT-4o ($0.10) and Claude ($0.75) as reflected in Table 5. This is reflective that, while GPT-4o is the best performer, lower-budget applications might be more likely to favor DeepSeek since it is less efficient but at lower cost.

7.1.4 General Observations

The outcome indicates a clear ranking: GPT-4o ranks the highest in terms of accuracy, strength, and efficiency, followed by Claude, DeepSeek, and Gemini. However, the lower METEOR scores across all models indicate a shared deficiency in handling semantic nuances, an issue of prime importance in medical use cases where accuracy comes first.

The bar graphs (Figures 2–5) proved most helpful in representing these differences such that trends were more evident. While GPT-4o's potential is encouraging, the lack of clinical validation and the general-purpose

nature of the models requires that their outputs be employed cautiously in real-world medical settings.

7.2 Project/Work Method Discussion

The methodology, as explained in Section 3.2, was a waterfall with five steps: Theory, Pre-study, Implementation, Measurement, and Evaluation. This section explains if this was an optimal way to go, what measures were utilized, and to what degree each stage accomplished its purpose to fulfill the goals of the project.

7.2.1 Phase 1: Theory

The Theory phase accomplished establishing the scientific foundation by exploring medical NLP, translation metrics (BLEU, METEOR, ROUGE-L), and Swedish-specific problems (Section 3.2.1). The literature review informed the model and metric choice to maintain compatibility with current NLP standards. However, the limited availability of Swedish medical corpora meant some domain-specific nuances (such as clinical synonyms) were not addressed, as discussed in Section 2.2. This constraint affected the Pre-study phase, in which the dataset had to be derived from general sources like SNOMED CT Browser and Folkets Lexikon.

7.2.2 Phase 2: Pre-study

Pre-study phase was able to develop the dataset (100.csv) with 100 Swedish medical terms, organized in phrases, abbreviations, and typos.

However, one of the limitations was having only one reference translation per term that can be incorrect compared to having multiple clinically acceptable translations possible in medical environments (such as andnöd as shortness of breath or dyspnea). While METEOR mitigated this by capturing synonyms, BLEU's strict exact-match criterion likely underestimated accuracy for terms with valid alternatives, as seen in Gemini's lower scores (BLEU: 85.00). This suggests that a multi-reference dataset could improve metric reliability, though it was beyond the scope of this study due to time constraints.

7.2.3 Phase 3: Implementation

Implementation phase (Section 3.2.3) effectively transformed the 100 terms with API-based models, with batching for rate limit handling. Web API usage was a practical choice, enabling rapid experimentation without local model training (Section 4.2). However web API downtime and rate limits occasionally disrupted processing, particularly with Gemini, with an exhibited highest response time of 24.89 seconds (Figure 4). Manual merging of CSVs was also dangerous in providing errors, which were not found in end results.

7.2.4 Phase 4: Measurement

Measurement phase (Section 3.2.4) accurately computed BLEU, METEOR, and ROUGE-L metrics as well as response times using Python libraries (NLTK, rouge_score, time module). Results were written to CSVs to facilitate analysis. Automatic computation of metrics was a benefit that facilitated scalability as well as objectivity. However, the lag in calculation of ROUGE-L scores (discovered in earlier drafts of Chapter 6) pushed back the analysis, and the lack of clinical validation limited the reliability of the scores within a healthcare environment.

7.2.5 Phase 5: Evaluation

The Evaluation phase (Section 3.2.5) was highly successful in decomposing the results and displaying them in bar graphs (Figures 2–5). The visualizations allowed for simple comparison of model performance across dimensions, setting GPT-4o as the better one and Gemini as the worse one. Error analysis served for the detection of specific failures, but the lack of clinical testing meant that possible errors might have been missed or underestimated.

7.2.6 Overall Method Reflection

Waterfall model offered a linear and systematic evaluation, where each phase built sequentially upon the deliverable of the prior phase. Linearity was beneficial for a time-limited thesis project because it prevented scope creep and ensured all the research questions were addressed. The inability to iterate and improve the dataset or the metrics halfway through the project (such as making typos more realistic) was the sole limitation of the waterfall model. The choice of automated scores

(BLEU, METEOR, ROUGE-L) was practical for scalability reasons at the cost of clinical accuracy because human clinician assessment would have provided more nuanced information (Section 4.2). Methodology generally achieved the goals of the project, but future studies can do better by adopting a more iterative process and clinical validation.

7.3 Scientific Discussion

This study aimed to evaluate to what extent Swedish medical language can be handled by language models, in terms of accuracy (Q1), robustness (Q2), and efficiency (Q3). The results provide valuable scientific contribution, particularly in the poorly researched area of Swedish medical NLP, and to the broader debate on the use of general-purpose models for specialized domains.

7.3.1 Research Questions

Q1: To what extent can English-Swedish medical terms and phrases be translated accurately? The study demonstrates that it is possible to achieve high accuracy for general-purpose models, with GPT-4o achieving 92.75 BLEU, 62.55 METEOR, and 86.66 ROUGE-L (Table 1). The use of a single reference translation in 100.csv, however, may have biased BLEU scores downward for words with multiple valid translations (such as andnöd). METEOR's higher sensitivity to synonyms addressed this to some degree, but the limitation suggests the value of multi-reference datasets for medical NLP research in the future

Q2: How robust are language models when translating ambiguous medical terms in a specific context? The robustness results are mixed. GPT-4o corrects 71.43% of the typos (Table 4). METEOR scores (Figure 4) also support that even the best-performing models fail at semantic disambiguation for ambiguous terms like andnöd. This is consistent with the domain mismatch in Section 2.2, where general-purpose models do not have the medical know-how for robust disambiguation.

Q3: What is the computational efficiency of the models when they conduct translation tasks? Efficiency varies widely, with GPT-4o being the most efficient (0.51 seconds per term) and Gemini being the least efficient (4.21 seconds average, 24.89 seconds maximum) (Figure 3). The speed-cost trade-off (Figure 5) is intriguing: DeepSeek and Gemini are cheaper but slower, whereas GPT-4o and Claude are more costly but

faster. This means that efficiency must be balanced against economic feasibility in real-world applications.

7.3.2 Scientific Contributions and Limitations

This study contributes to the field by providing us with a point of reference for Swedish medical translation, which heretofore has been little researched. Using a locally developed dataset (100.csv) with diverse types of terms (phrases, acronyms, misspellings) and the comprehensive evaluation framework (Table 1 in Section 3.4) offer an replicable process that can be adopted by future studies. The findings confirm that general-purpose models will be accurate on simple terms but struggle with semantic nuance and robustness, providing an argument for domain-specific fine-tuning.

However, the work is not without its constraints. The lack of clinical validation means that the "correct" translations based on automated metrics are not necessarily clinically appropriate (Section 7.1). The utilization of general-purpose models with no fine-tuning limits their applicability to real-world medical settings, where mistakes can prove to be hazardous. The artificial typos may not actually represent actual-world mistakes, and the limited data set (100 terms) may not actually represent the richness of medical jargon.

7.4 Consequence Analysis

7.4.1 Scientific Impact

This study advances the state of knowledge related to language models in Swedish medical NLP, a new but significant area of research. The findings highlight the potential of general-purpose models like GPT-4o for medical translation tasks, particularly for simple words, but also underscore the necessity of domain-specific fine-tuning to boost robustness and semantic accuracy. By providing a baseline and setting important bounds (such as lower METEOR scores, non-achievement of typo correction), the work opens doors to subsequent research on fine-tuning models for medical use

7.4.2 Consequence

The results have implications for medical use in healthcare, such as medical report translation or aid in clinical decision support. GPT-4o's efficiency and accuracy are sufficient for such work, but its semantic

limitations (for example, METEOR: 62.55) require that clinicians manually validate its results to avoid errors. An example of mistranslating andnöd as shortness of breath instead of dyspnea may lead to miscommunication in the clinical context, and may alter patient care. The cost analysis (Figure 5) also picks out economic considerations: while GPT-4o performs well, cheaper models like DeepSeek can be more convenient for large-scale applications even though they are slower.

7.4.3 Recommendations

To medical professionals or organizations developing AI-driven medical translation software, I recommend the following:

Fine-Tuning: Invest in fine-tuning models like GPT-4o or Claude on Swedish medical corpora for semantic strength and accuracy. This can involve corpora like those developed by the Swedish Medical Language Data Lab [8].

Clinical Validation: Add human-in-the-loop validation by clinicians to ensure that translations are clinically appropriate, addressing the shortcoming of automatic metrics.

Hybrid Approach: Take the strengths of fast, accurate models (such as GPT-4o) and use lower-cost models (such as DeepSeek) in a hybrid model, using the former for real-time and the latter for batch work.

Increase Dataset: Develop larger, more diverse datasets with real-world errors (such as OCR errors, phonetic misspellings) to better test robustness and generalizability.

Alternative Models: Consider specialized models such as MedPaLM 2, specifically created by Google for clinical environments, that may improve translation precision for medical terminology. Nevertheless, MedPaLM 2 is not available in Sweden and necessitates a selection process for its use, reducing its applicability to this study. Its presence might have set a benchmark for domain-specific performance, as it was trained on clinical data.

7.5 Ethical and Societal Discussion

7.5.1 Ethical Considerations

Medical translation applications of AI language models also raise different ethical concerns. Firstly, possible mistranslations might facilitate miscommunication within clinical settings that could potentially harming patients. For example, if a mistranslated word in a medical report leads to an erroneous diagnosis or treatment, the consequence might be catastrophic. This creates an overriding ethical obligation to generate accurate and trustworthy translations, particularly for critical use like healthcare.

Secondly, the lack of clinical verification in this study means that the translations were not verified by medical professionals, and their applicability to real practice is dubious. Even though the study was not real and on paper based on no real patient data, the use of such models in practice without verification would be unethical as it can put patient safety at risk.

Third, the use of general-purpose models trained from large corpora (Section 2.2) involves the risk of bias. Such models can reflect biases in their training data, for example, underrepresentation of Swedish medical contexts or English-biased medical domain knowledge. This could lead to skewed performance across languages and disproportionately affect non-English speaking communities like Swedish patients.

7.5.2 Societal Impact

The societal value of this study lies in its potential for the improvement of healthcare accessibility in multilingual settings. The high-quality translation of Swedish medical terms to English (and vice versa) could facilitate communication between medical practitioners and patients in multicultural settings, such as in Sweden, where English-speaking practitioners may need to interpret Swedish documentation. This has the potential for improved patient care, reduced miscommunication, and improved health outcomes.

However, the deployment of such models must be carried out with caution. Mistranslations have the potential to erode clinician and patient trust in AI systems and could hold back the adoption of AI in medicine. In addition, the variations in cost (Figure 5) also raise equity concerns: although GPT-4o offers the best performance, its higher cost ($0.10 per project) may limit accessibility for smaller healthcare organizations,

whereas less expensive models like DeepSeek (< $0.01) are less reliable. This could exacerbate healthcare disparities in favor of well-equipped institutions over under-resourced institutions.

7.5.3 Privacy and Human Involvement

This study didn't involve human subjects or even real patient data, as the data set (100.csv) was generated out of public material (SNOMED CT Browser, Folkets Lexikon). In real-world applications, privacy would be a top priority, though. Medical data is very sensitive, and applying API-based models to converting medical records would put data at risk if adequate measures are not in place. For example, exposing patient data to external web APIs in cleartext would be against privacy regulations like GDPR, which is very strict in the EU, including Sweden.

7.5.4 Responsibilities

As scientists and engineers, it is our responsibility to make sure healthcare AI systems are safe, accurate, and unbiased. This work highlights clinical validation, domain-specialized fine-tuning, and open cost-benefit analysis as critical to inform deployment decisions. We must also advocate for patient privacy-respecting policies and ensure fair access to AI tools, preventing a "digital divide" in healthcare where benefits from advanced technologies only accrue to wealthy institutions.

In short, while this study demonstrates the capabilities of language models for Swedish medical translation, it also identifies the societal and ethical needs of using them. Addressing these challenges through validation, fine-tuning, and protection of privacy will be the critical factor in delivering their benefits and minimizing risks.

8 Conclusions

This thesis examined the performance of four language models—GPT-4o, DeepSeek, Gemini, and Claude—in translating Swedish medical terminology to English, plugging a research gap for low-resource languages in specialized domains. The aim was to contrast the models' accuracy, stability, and computational expense (Q1, Q2, Q3) in an effort to assess their adequacy for processing Swedish medical language, as outlined in the problem statement (Section 1.2). The research method applied a proprietary data set (100.csv) made up of 100 words, automated metrics (BLEU, METEOR, ROUGE-L), and response time measures, depicted by bar charts (Figures 2–5).

The results, as explained in Chapter 6, indicated that GPT-4o was the best in all parameters, achieving a BLEU measure of 92.75, correcting 71.43% typos, and achieving a mean of 0.51 seconds per word. Claude trailed behind with strong performance (BLEU: 89.00, typo correction: 66.67%, average time: 1.44 seconds), whereas Gemini (BLEU: 88.10, typo correction: 57.14%, average time: 5.34 seconds) and DeepSeek (BLEU: 85.00, typo correction: 47.62%, average time: 4.21 seconds) trailed behind. More general METEOR scores (such as 62.55 for GPT-4o), on the other hand, indicated semantic nuance challenges and possibly downward skewed BLEU scores because of the use of a single reference translation in the test set for instances of more than one valid translation (such as andnöd as shortness of breath or dyspnea), as investigated in Chapter 7.

8.1 Answers to Research Questions

Q1: To what extent can English-Swedish medical terms and phrases be translated accurately?

The models trained with good precision, with GPT-4o attaining 92.75 BLEU, which reflects high performance on straightforward words (100% accuracy) and sentences (90%). Precision was low for typos (71.43%), and semantic nuance (METEOR: 62.55) reflects weakness, partially mitigated due to the single-reference dataset restriction.

Q2: How robust are language models while translating ambiguous medical terms in a specific context?

Robustness was moderate, with GPT-4o correcting 71.43% of typos and Claude 66.67%, but failures highlight weaknesses in handling noise and ambiguity. This goal was partially met, limited by the models' general-purpose training.

Q3: What is the computational efficiency of the models when they conduct translation tasks?

Efficiency was inconsistent, with GPT-4o taking 0.51 seconds per term and Gemini 4.21 seconds, as the goal to reflect disparities impacted by web API constraints and model capacity was attained.

8.1.1 Answer to Problem Statement

The question problem asked to what extent current multilingual and Swedish-expert language models can translate and produce medical content with Swedish-English vocabulary accurately, particularly when ambiguity and specialty words are at issue. Results indicated that the best optimal performance is from GPT-4o but each of the models struggles with semantic accuracy and robustness due to their general-use nature and lack of clinical fine-tuning. The goal was partly fulfilled to provide a baseline but emphasizing that domain-specific applications are required.

8.1.2 Scientific Contribution and Impact

This work provides a new baseline for Swedish medical NLP, evaluating general-purpose models on a hand-annotated corpus with diverse forms of terms (phrases, abbreviations, typos). Results highlight the potential of models like GPT-4o to medical translation with constraints (semantic subtlety, typo repair) that inform future research. The contribution lies in informing healthcare applications, for example, medical report translation, even if clinical validation remains necessary to ensure safety and reliability.

8.2 Future work

The research also provides opportunities for several avenues of further research and development, surpassing its own limitation and building on its findings. Below are specific proposals for continuing this work, including how they could be approached.

8.2.1 Exploration of Specialized Models

One direction that can be explored is adding specialized models like MedPaLM 2, a clinical model developed by Google that would enhance translation quality for Swedish medical terminology. However, its current lack of availability in Sweden and the necessity of a selection process cause issues. If available, I would proceed by collaborating with Google to obtain approval and then transfer MedPaLM 2 to a Swedish medical corpus (for example, SNOMED CT) using transfer learning techniques. This would entail collecting annotated clinical data, training the model in Google Colab with a larger dataset (for example: 500 terms) and subsequently evaluating its performance relative to GPT-4o's using BLEU, METEOR, and clinician-validated scores.

8.2.2 Development of a Multi-Reference Dataset

Future work can also develop a multi-reference dataset having clinically validated variants as a remedy for the restriction of one reference translation. I would accomplish this by working together with Swedish healthcare professionals in annotating 200–300 words using synonyms and context-dependent definitions. The corpus can be used in retraining metric models (such as BLEU with a limited number of references) and improving reliability in accuracy measurement, possibly by having Python scripts perform automatic annotation and test against human sentiments.

8.2.3 Clinical Validation and Real-World Testing

For increased practical usefulness, future work must include clinical evaluation by clinicians. I would do this by recruiting 5–10 Swedish clinicians to review 100 translations from the highest-performing model (GPT-4o), rating their clinical usefulness on a Likert scale (1–5). This could be done via a survey mechanism such as Google Forms, with results analyzed for patterns of error (such as semantic mismatches) and model revision guidance. Real-world evaluation in a healthcare context could then follow, running the model on actual medical reports with ethics approval.

8.2.4 Optimization for Cost and Efficiency

With the cost differences, future development could look to optimize a combined hybrid model offering the quickness of GPT-4o and cost-effectiveness of DeepSeek. I would accomplish this by building a Python pipeline that sends plain words to DeepSeek and tricky words to GPT-

4o, depending on a classifier trained on 100.csv. This would involve evaluating the hybrid system on 200 words, measuring cost, time, and accuracy to trade off economic feasibility with efficiency.

These directions for the future would address the study's limitations, strengthen it scientifically, and pave the way for practical use within healthcare settings, both precise and accessible.

References

[1] Z. M. Zayyanu, "Revolutionising Translation Technology: A Comparative Study of Variant Transformer Models - BERT, GPT, and T5," Computer Science & Engineering: An International Journal, vol. 14, Jun. 2024. https://doi.org/10.5121/cseij.2024.14302.

[2] R. S. Star, "Natural Language Processing (NLP) and its importance in AI - Serenity Star," Serenity Star, Feb. 17, 2025. https://serenitystar.ai/blog/natural-language-processing-nlp

[3] Turkan Ismayilli, "Navigating Complexities in Medical Text Translation: Challenges, Strategies, and Solutions," vol. 1, no. 2, Dec. 2024, https://doi.org/10.69760/aghel.01024080.

[4] H. Al Shamsi, A. G. Almutairi, S. Al Mashrafi, and T. Al Kalbani, "Implications of Language Barriers for healthcare: A Systematic Review," Oman Medical Journal, vol. 35, Apr. 2020, https://doi.org/10.5001/omj.2020.40.

[5] M. Zappatore and G. Ruggieri, "Adopting machine translation in the healthcare sector: A methodological multi-criteria review," Computer Speech & Language, vol. 84, p. 101582, Mar. 2024, https://doi.org/10.1016/j.csl.2023.101582

[6] C. Lin and C.-F. Kuo, "Roles and Potential of Large Language Models in Healthcare: A Comprehensive Review," Biomedical Journal, p. 100868, Apr. 2025, https://doi.org/10.1016/j.bj.2025.100868.

[7] J. Vrdoljak, Z. Boban, M. Vilović, M. Kumrić, and J. Božić, "A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration," Healthcare, vol. 13, no. 6, p. 603, Mar. 2025, https://doi.org/10.3390/healthcare13060603.

[8] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: applications, advances, and challenges,"

Artificial Intelligence Review, vol. 57, no. 11, Sep. 2024, https://doi.org/10.1007/s10462-024-10921-0.

[9] Aadit Jerfy, O. Selden, and Rajesh Balkrishnan, "The Growing Impact of Natural Language Processing in Healthcare and Public Health," INQUIRY The Journal of Health Care Organization Provision and Financing, vol. 61, Jan.0 https://doi.org/10.1177/00469580241290095.

[10] S. Velupillai et al., "Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances," Journal of biomedical informatics, vol. 88, pp. 11–19, 2018, https://doi.org/10.1016/j.jbi.2018.10.005.

[11] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv*, Jun. 12, 2017. https://arxiv.org/abs/1706.03762

[12] M. Malmsten, L. Börjeson, and C. Haffenden, "Playing with Words at the National Library of Sweden Making a Swedish BERT," *arXiv.org*, 2020. https://arxiv.org/abs/2007.01658

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," Jul. 2002. Available: https://aclanthology.org/P02-1040.pdf

[14] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *ACLWeb*, Jun. 01, 2005. https://aclanthology.org/W05-0909/

[15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Jul. 2004. Available: https://aclanthology.org/W04-1013.pdf

[16] "View of An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score," Mozhgan Ghassemiazghandi, 2024. https://tpls.academypublication.com/index.php/tpls/article/view/7867/6371

[17] J. P. Biyong, B. Wang, T. Lyons, and Nevado-Holgado, Alejo J, "Information Extraction from Swedish Medical Prescriptions with Sig-Transformer Encoder," *arXiv.org*, 2020. https://arxiv.org/abs/2010.04897

[18] “Swedish Medical Language Data Lab,” *AI Sweden*, 2025. https://www.ai.se/en/project/swedish-medical-language-data-lab

[19] J. B. Nilsson, “Punctuation restoration in Swedish through fine-tuned KB-BERT,” arXiv.org, 2022. https://arxiv.org/abs/2202.06769

[20] X. Wang, H. Ye, S. Zhang, M. Yang, and X. Wang, “Evaluation of the Performance of Three Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases,” Journal of Medical Systems, vol. 49, no. 1, Feb. 2025, https://doi.org/10.1007/s10916-025-02152-9.

[21] A. Nucci, “What is a Domain-Specific LLM? Examples and Benefits,” Aisera: Best Generative AI Platform For Enterprise, Jan. 11, 2024. https://aisera.com/blog/domain-specific-llm/

[22] “Gemini Developer API Pricing,” *Google AI for Developers*, 2025. https://ai.google.dev/gemini-api/docs/pricing

[23] OpenAI, “OpenAI Platform,” *Openai.com*, 2025. https://platform.openai.com/docs/pricing

[24] DeepSeek, “Models & Pricing | DeepSeek API Docs,” *Deepseek.com*, 2025. https://api-docs.deepseek.com/quick_start/pricing

[25] “Pricing,” *www.anthropic.com*. https://www.anthropic.com/pricing

Appendix A: Data Set and Resources

All datasets, source code, prompts, and additional resources used in this study are provided in GitHub

https://github.com/Juleol2/Thesis_JulioCarvajal