Master Thesis

From Nine to One: Combining ICF Functioning Level Classifiers in the A-PROOF Project

Urtė Jakubauskaitė

Supervisor	Piek Vossen & Edwin Geleijn
2^{nd} reader	Sophie Arnoult

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics
(Text Mining - Language & AI)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date	29th December, 2025
Student number	2870111
Word count	17,662

Abstract

This thesis investigates whether the language used in medical notes to describe patient functioning is sufficiently consistent to support a single, generalizable classification model across multiple ICF categories. Building on the A-PROOF project, which automates the annotation of functioning levels, the study explores whether nine category-specific classifiers can be replaced by a single, more efficient model.

Three modeling approaches are developed and evaluated: the original setup with nine separate classifiers, a unified model trained across all categories, and a unified model augmented with category-specific encodings. These models are trained on the original A-PROOF dataset as well as on combined datasets that incorporate generated data, including data representing previously unseen categories. Model performance is assessed using Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, complemented by two targeted error analyses.

The results show that a single generalizable model is feasible. The model incorporating category encodings achieves the best overall performance, maintaining or improving accuracy across most functioning categories. Performance is reduced for *Attention functions (ATT)*, where limited data availability and imbalanced level distributions remain challenging, as well as for *Walking (FAC)* and *Exercise tolerance functions (INS)* due to scale differences and difficulties in annotating these categories. Overall, most categories benefit from larger and more diverse training data, even when this includes unseen categories. This effect is particularly evident in Model 3, where the mixed-category dataset improves performance for all categories except *Energy levels (ENR)*. Predictions tend to be conservative, favoring mid-scale functioning levels. These findings highlight level imbalance as a key limitation and emphasize the importance of larger, balanced datasets for improving predictive accuracy across ICF categories.

Declaration of Authorship

I, author, declare that this thesis, titled *From Nine to One: Combining ICF Functioning Level Classifiers in the A-PROOF Project* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 29th December, 2025

Signed: .

Acknowledgments

I would like to express my sincere gratitude to everyone who contributed to the success of this thesis.

First and foremost, I would like to thank my supervisors, Piek Vossen and Edwin Geleijn. I am deeply grateful not only for their guidance and support throughout this process but also for the unique opportunity to work on a topic that truly mattered to me. I greatly appreciate the chance to contribute to the field of AI in Health and to gain hands-on experience working with real medical data.

I would also like to thank Jesse Aarden, Marike van der Leeden, Sabina van der Veen, Carel Meskers, and the aforementioned Edwin Geleijn for the time and care they dedicated to validating my dataset. Their expertise, feedback, and attention to detail were invaluable and made this work stronger.

I would also like to thank all the teachers and staff of the Computational Linguistics & Text Mining Lab whom I had the pleasure of learning from during this program, namely: Antske Fokkens, Hennie van der Vliet, Pia Sommerauer, Luís de Passos Morgado da Costa, Lucia Donatelli, Isa Maks, Sophie Arnoult, and Stella Verkijk. Thank you for the knowledge you shared and for inspiring me through your passion and dedication.

Furthermore, I would like to thank the members of the Institute for Logic, Language and Computation at the University of Amsterdam. I began my academic journey there as a student in the MSc Logic program, and I am thankful for their support when I decided to pursue a second Master's degree. A special thank you goes to Raquel G. Alhama, Paul Dekker and Tanja Kassenaar for their encouragement and kind words along the way.

Finally, I want to thank my parents, Daiva and Remigijus, and my sister, Agnė, for making me feel their unwavering support despite the distance between us. My last thank you goes to my partner, Adel, whose care, encouragement, and practical support made it possible for me to stay focused and confident throughout this journey.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Artificial Intelligence (AI) is increasingly transforming many aspects of our lives, including healthcare. With recent advances in one of its subfields, Natural Language Processing (NLP), AI systems are now capable of interpreting unstructured textual data in Electronic Health Records (EHRs) (Centers for Medicare & Medicaid Services, 2024), unlocking new opportunities for patient care and clinical decision-making. One area where this has proven particularly valuable is in monitoring and assessing patient functioning, for instance, in post-COVID-19 rehabilitation.

This thesis focuses on optimizing the classification pipeline developed within the A-PROOF project: *Automated Prediction of post-COVID-19 RecOvery Of Functioning* (2022). The project aims to assess and predict patients' functional status based on textual clinical notes extracted from EHRs. These notes, provided by Amsterdam UMC, contain descriptions of patients' performance across multiple domains, such as walking, attention, respiration, and emotional state.

1.1 Thesis Aim and Research Question

This research builds upon the A-PROOF project by attempting to reduce the number of classifiers required for predicting functioning levels, while ensuring that relevant functioning categories, namely, *Energy level (ENR)*, *Attention functions (ATT)*, *Emotional functions (STM)*, *Respiration functions (ADM)*, *Exercise tolerance functions (INS)*, *Weight maintenance functions (MBW)*, *Walking (FAC)*, *Eating (ETN)*, and *Work and employment (BER)*, are accurately identified within medical notes.

Currently, the system uses a two-step classification process. First, a multi-label text classifier identifies which (if any) International Classification of Functioning, Disability, and Health (ICF) (World Health Organization, 2025) category applies to each sentence. Second, for each of the nine functioning categories, a separate regression classifier determines the specific level of functioning. These sentence-level predictions are then aggregated to produce a note-level classification. However, this pipeline is inefficient due to the large number of models required.

This thesis aims to make the system more compact and efficient by replacing the nine individual regression classifiers with a single unified regression classifier, while keeping the existing category classifier unchanged. Reducing the number of classifiers would decrease the system's computational load and improve its practicality for deployment in clinical environments, particularly in resource-constrained settings. Moreover, a unified model would be more generalizable, increasing its potential to predict functioning levels for new categories in the future.

A key challenge lies in the linguistic variability of medical notes. Medical language is highly specialized, often characterized by non-standard grammar and terminology. For example, the Dutch sentence *De stemming imponeert normofoor, met een normaal modulerend affect* (translated: *Mood impresses normophore, with normal modulating affect*) is ungrammatical in general Dutch; however, it is considered acceptable in medical texts. This example illustrates the complexity and domain-specific phrasing that make automatic interpretation difficult. Moreover, expressions of functioning levels may differ across categories, further complicating the classification task.

A central question, therefore, is whether regression-level classifiers can effectively learn from level annotations across different functioning categories, or whether category-specific variation in language introduces ambiguity. This leads to the following research question:

> *Is the language used in medical notes to describe different functioning levels consistent enough to allow for the development of a single generalizable classification model across all functioning categories?*

1.2 Thesis Contributions

This thesis makes the following contributions:

- Proposes a more efficient classification pipeline by reducing the number of regression models from nine to one, thereby improving scalability and computational efficiency;

- Provides a comparative evaluation of the multi-model and single-model approaches, assessing the trade-offs in performance and generalization;

- Compares the performance of identical model architectures trained on different datasets, ranging from the original nine-category dataset to two expanded combined datasets, to determine the extent to which dataset composition influences model robustness and cross-category generalization;

- Provides a systematic analysis of cross-category linguistic variation, including two detailed error analyses examining scale differences (0–4 vs. 0–5) and the role of intensifiers, minimizers, and negations.

Chapter 2

Background and Related Work

This chapter provides the necessary background and contextualizes the work presented in this thesis. It begins by introducing key concepts in artificial intelligence (AI) and natural language processing (NLP) as applied to the healthcare domain. Next, it discusses the International Classification of Functioning, Disability and Health (ICF), which provides a standardized framework for assessing patient functioning, and the role of electronic health records (EHRs), which serve as the primary data source for this study. The chapter also introduces pre-trained language models, such as BERT and RoBERTa, as well as domain-adapted models, including MedRoBERTa.nl, which forms the basis for the modeling approaches used in this research.

Following this, the chapter addresses system optimization, focusing on replacing the original nine-model architecture with a single regression classifier capable of handling all functioning categories simultaneously while maintaining prediction accuracy. Finally, the chapter concludes with a presentation of the A-PROOF project, which serves as the foundation for this thesis.

2.1 Artificial Intelligence in Healthcare

In recent years, artificial intelligence (AI) has become one of the fastest-growing industries in the world (Elliott, 2025). AI is increasingly integrated into our daily lives: while some people use it as a practical tool for everyday tasks, such as meal planning, others rely on AI-powered coaches to help create personalized running plans.

However, AI does more than just simplify everyday life; it also drives progress in other fields, such as healthcare. Although many people still distrust the use of AI in medical contexts due to concerns about data privacy, ethics, and bias (James, 2023), its benefits are hard to ignore. AI enables higher accuracy in cancer detection (Cancer Research Institute, 2025), assists in patient monitoring (Dubey and Tiwari, 2023), and significantly improves workflow efficiency, which ultimately saves time for healthcare professionals.

2.2 Natural Language Processing in Healthcare

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on enabling computers to understand, interpret, and generate natural, or human, language. In healthcare, NLP plays a crucial role: it saves time of medical workers by dealing with huge amounts of unstructured textual data, for example, transforming clinical notes into structured information.

Increasing availability of digital health data as well as fast advances in machine learning techniques makes the field of NLP grow significantly. Therefore, currently, NLP systems already heavily assist in various tasks including clinical documentation, information extraction, diagnosis prediction, and patient outcome analysis (Nelson et al., 2025). For example, NLP algorithms can automatically identify symptoms, medications, and treatment plans mentioned in clinical narratives, reducing manual effort and improving data quality.

Moreover, NLP supports decision-making in clinical settings by extracting relevant insights from clinical texts that are otherwise difficult to analyze at scale. However, challenges remain due to the complexity and variability of medical language, which often includes abbreviations, jargon, and non-standard grammar (Saxena, 2024).

In this thesis, NLP techniques are applied to analyze medical notes from electronic health records (see Section 2.4.1), aiming to classify patient functioning levels according to the ICF.

2.3 International Classification of Functioning, Disability and Health (ICF)

The International Classification of Functioning, Disability and Health (ICF) (World Health Organization, 2025) is a framework developed by the World Health Organization and formally adopted in 2001. It provides a unified and scientifically grounded system for describing health, functioning, and disability across clinical, research, and policy contexts.

The ICF can be applied at multiple levels. At the individual level, it supports the assessment of a person's functioning, assists in planning personalised interventions, and facilitates communication among healthcare professionals. At the population level, the ICF informs service planning, evaluation of healthcare quality, and analysis of the effectiveness and cost-efficiency of interventions. At the societal level, it contributes to policy development, eligibility assessments for social support, and broader public health monitoring.

Overall, the ICF offers a consistent international standard for understanding and documenting functioning and disability, making it highly relevant for studies aiming to extract functioning-related information from clinical text, such as the work presented in this thesis.

2.4 Medical Data

One of the central challenges in NLP is domain adaptation (Laparra et al., 2020). Models trained on general-purpose corpora often fail to capture the linguistic characteristics of specialised domains, resulting in reduced performance and unreliable predictions. The medical domain strongly exemplifies this challenge, as clinical texts differ substantially from general-purpose language and therefore require specialised approaches to data preparation and model development.

In the present study, the data consist of medical notes written by healthcare professionals. These notes describe observations, functioning levels, symptoms, and other clinically relevant information.

2.4.1 Electronic Health Records (EHRs) and Medical Notes

Electronic Health Records (EHRs) Centers for Medicare & Medicaid Services (2024) are systems used to electronically store patient information collected over time. These records contain both structured data, such as laboratory results, and unstructured data. A substantial amount of clinically relevant information appears in the latter, particularly in discharge summaries, progress notes, and other free-text medical notes.

Importantly, these medical notes follow a writing style distinct from general text: they are concise, rely heavily on domain-specific abbreviations, and are often grammatically irregular (Bagheri et al., 2023). Moreover, to save time, clinicians frequently omit function words.

This thesis, like the broader A-PROOF project (Section 2.7), uses EHRs to identify relevant functioning categories and the corresponding levels of difficulty described in patient notes.

2.4.2 Challenges in Medical Text Processing

Processing data in the medical domain brings several unique challenges for NLP research. First, the prevalence of specialised terminology and acronyms in clinical text increases data sparsity and limits the effectiveness of models trained on general-domain corpora (Moon et al., 2015). NLP systems therefore require sufficient training data to learn these expressions effectively.

Second, medical notes are often ungrammatical and fragmented, which makes them challenging for models that are typically trained on large amounts of well-formed text (Wu and Liu, 2011). Third, this domain frequently suffers from severe data imbalance (De Angeli et al., 2022). Collecting enough training data is already difficult, and on top of that, only a small portion of clinical notes contains information relevant to the task. A-PROOF is one such example: when processing EHRs, researchers found that only around 5% of the notes include information about the relevant ICF categories, further reducing the size of the usable training dataset.

Finally, a major challenge in medical NLP is the limited availability of data and the strict privacy requirements associated with its use. Clinical notes contain sensitive personal information and are therefore subject to stringent regulations (Conduah et al., 2025). As a result, access to such data is typically restricted to controlled settings, and obtaining approval can be a lengthy and complex process. Moreover, anonymisation procedures, while necessary, may remove contextual details that are valuable for downstream NLP tasks.

2.5 Use of Pre-trained Language Models: Transformer Models

This study uses Transformer-based language models to classify functioning categories and their corresponding levels in clinical notes. Transformer models have become the standard in NLP because they handle long-range dependencies and contextual relationships between tokens exceptionally well (Khurana et al., 2022). In this study, the relevant Transformer models are BERT (Devlin et al., 2019), its later version

RoBERTa (Liu et al., 2019), and its domain-adapted variant, MedRoBERTa.nl (Verkijk and Vossen, 2025).

When such models are pre-trained on large amounts of text, they learn rich linguistic representations. This substantially reduces the amount of task-specific data required for fine-tuning (Liu et al., 2024) - an important advantage in domains where data are difficult to access or inherently limited, such as medical text.

2.5.1 BERT and RoBERTa

BERT, which stands for *Bidirectional Encoder Representations from Transformers* (Devlin et al., 2019), is a pre-trained language model trained on large general-domain corpora. Because it is bidirectional, its architecture allows it to consider both left and right context simultaneously. This, combined with BERT's ability to learn from large amounts of unlabeled text, makes it a particularly powerful model.

Importantly, BERT must be fine-tuned on a specific task to ensure that it captures task-relevant patterns and domain-specific details. Fine-tuning BERT is relatively straightforward, as it typically requires no major changes to the model architecture, except for the addition of a specific output layer on top of the pre-trained encoder.

RoBERTa, which stands for *Robustly Optimized BERT Approach* (Liu et al., 2019), is a later model built on the same architecture as BERT but trained more extensively and with several optimizations. These include training on larger datasets, removing the next-sentence prediction objective, and using dynamic masking. As a result, RoBERTa generally achieves stronger performance than the original BERT model.

2.5.2 Domain-adapted Models

Domain-specific language models have consistently been shown to outperform models that have not undergone domain adaptation (Kim et al., 2022). This observation is supported by multiple studies. Two main strategies have proven effective: extending pre-training with domain-specific data, as in BioBERT (Lee et al., 2019) and ClinicalBERT (Huang et al., 2019), and pre-training from scratch on domain-specific data, as in PubMedBERT (Gu et al., 2020) and SciBERT (Beltagy et al., 2019). The choice between these approaches often depends on the specific downstream task (Chalkidis et al., 2020).

2.5.3 MedRoBERTa.nl

In the Dutch medical context, the only existing large language model is MedRoBERTa.nl (Verkijk and Vossen, 2025), which was trained from scratch on nearly 10 million EHR notes. It is also one of only two publicly accessible models worldwide pre-trained on EHRs, the other being GatorTron (Yang et al., 2022).

MedRoBERTa.nl is a domain-adapted version of RoBERTa (Verkijk and Vossen, 2025), created by pre-training the RoBERTa architecture on Dutch electronic health records. Because it is trained on clinical text, MedRoBERTa.nl is particularly well-suited to handle medical terminology, abbreviations, and the stylistic characteristics commonly found in medical notes. MedRoBERTa.nl has demonstrated clear improvements over general Dutch models, such as BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020), in zero-shot similarity judgment tasks and medical text classification after fine-tuning.

2.6 Optimization of Medical NLP Systems

When developing Medical NLP systems, it is essential to focus not only on accuracy but also on practical applicability, as the ultimate goal is to implement such systems in real clinical settings (Ansar et al., 2024). The existing setup, nine separate regression classifiers, one for each functioning category, requires substantial computational resources, making it both time-consuming and difficult to deploy in environments where resources are limited. For this reason, the present study aims to optimize the system by replacing the nine-model architecture with a single regression classifier capable of handling all categories simultaneously.

The main challenge in this optimization is maintaining prediction accuracy. Training one model on all categories at once may reduce its sensitivity to category-specific linguistic patterns. Beyond this, the study must also address the issue of data imbalance, which is common in medical datasets. Some functioning categories, such as *Respiration functions (ADM)*, appear frequently, while others, such as *Attention functions (ATT)*, are much more rare. When categories are combined into a unified model, there is a risk that the classifier becomes biased toward the dominant classes, paying insufficient attention to the underrepresented ones simply because they appear less often in the training data.

Optimizing Medical NLP models therefore requires carefully balancing efficiency and predictive reliability, ensuring that the system remains computationally lightweight while still capturing the nuanced patterns necessary for accurate predictions across all functioning categories.

2.7 The A-PROOF Project

An example of AI implementation in healthcare is the A-PROOF project, which stands for *Automated Prediction of post-COVID-19 RecOvery Of Functioning.* The project aims to develop technology that can automatically monitor the functioning of individuals who require healthcare support. It is based on the dataset created by Meskers et al. (2022) and the system developed by Kim et al. (2022) for detecting functioning categories and their corresponding levels of difficulty in Dutch EHR texts.

The system proposed by Kim et al. (Kim et al., 2022) follows a two-step modelling approach. In the first step, ICF category classification is performed using MedRoBERTa.nl, a domain-adapted version of RoBERTa trained on Dutch medical data (see Section 2.5.3), which is fine-tuned for multilabel text classification. This step determines which ICF functioning categories are mentioned in a given clinical note. In the second step, level classification is carried out using text regression models that predict the level of difficulty associated with each identified category. Separate regression classifiers are fine-tuned for each ICF category, resulting in a total of nine category-specific regression models.

Using this two-stage architecture, the model achieved F1-scores above 80% for the main ICF categories and produced level predictions with average errors of less than one point on a five-point Likert scale, demonstrating strong performance in extracting ICF-based functioning information from clinical text.

The present research builds directly on this project by investigating whether the original modelling approach can be made more computationally efficient without compromising predictive accuracy by training a single regression classifier for assigning the level to any category.

Chapter 3

Methodology

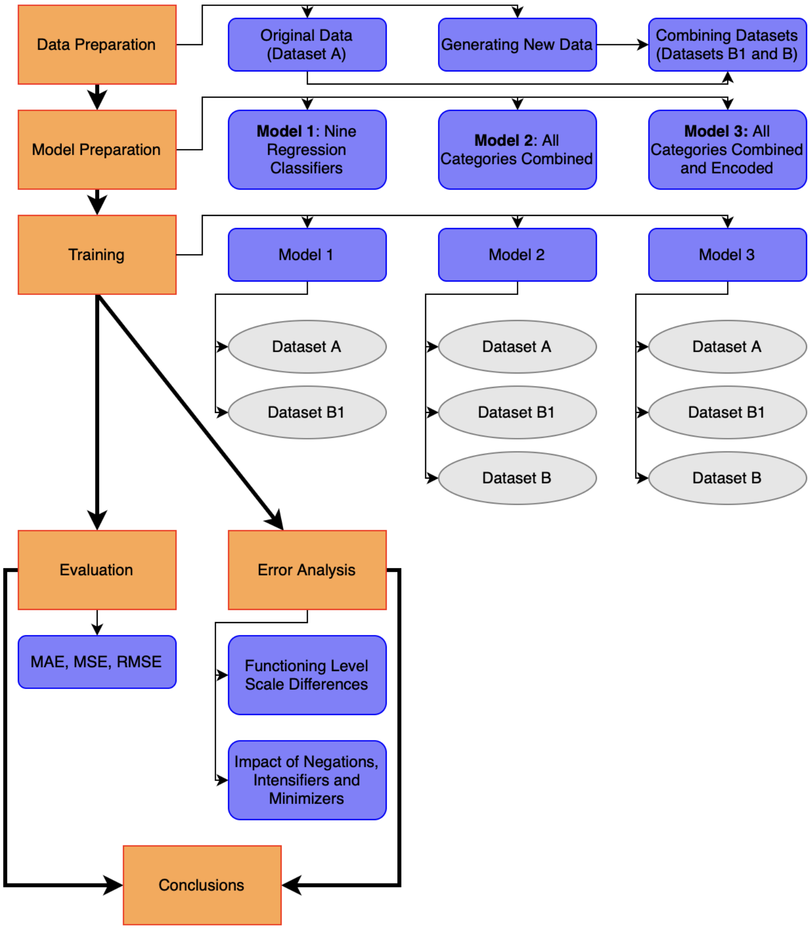
Data Preparation
Original Data (Dataset A)
Generating New Data
Combining Datasets (Datasets B1 and B)
Model Preparation
Model 1: Nine Regression Classifiers
Model 2: All Categories Combined
Model 3: All Categories Combined and Encoded
Training
Model 1
Model 2
Model 3
Dataset A
Dataset B1
Dataset A
Dataset B1
Dataset B
Dataset A
Dataset B1
Dataset B
Evaluation
MAE, MSE, RMSE
Error Analysis
Functioning Level Scale Differences
Impact of Negations, Intensifiers and Minimizers
Conclusions

Figure 3.1: Flowchart illustrating the methodological steps taken to answer the research question.

This chapter provides an overview of the methodological approach used to address the research question, which examines whether the language used to describe different functioning levels in medical notes is consistent enough to support the development of a single generalizable model across all functioning categories. The overall methodology consists of five stages: data preparation, model development, model training, evaluation, and error analysis. Figure 3.1 illustrates the complete workflow.

3.1 Data Preparation

The study is based on three types of datasets. The original A-PROOF data are split into training, development, and test sets. The training split of this original data serves as the primary training resource in this study and is referred to as Dataset A.

In addition to the original data, a generated dataset is constructed using OpenAI. For this dataset, both the functioning categories and the corresponding functioning level annotations are automatically generated based on selected sentences. The purpose of this dataset is to investigate whether increasing the amount of available training data improves prediction performance, particularly for categories with limited coverage in the original data.

Finally, combined datasets are created by merging the original A-PROOF data with the generated sentences. Two variants of the combined data are used. Dataset B1 includes only the nine original functioning categories present in Dataset A, whereas Dataset B contains both the original nine categories and eight newly generated categories. These datasets allow for a systematic comparison of model performance under different data availability conditions.

3.2 Model Development

Three different modeling approaches are developed and compared in this study. All models are based on the same underlying transformer architecture, MedRoBERTa.nl, and differ only in how functioning categories are represented and learned.

The first model corresponds to the baseline approach used in the A-PROOF project and consists of nine separate regression classifiers, with one classifier trained independently for each functioning category.

The second model adopts a unified approach in which all functioning categories are combined into a single training setup. In this configuration, a single regression model is trained to predict functioning levels across categories without explicit information about the category to which a sentence belongs.

The third model extends the unified approach by explicitly encoding category information through special tokens. Each input sentence is preceded by a category-specific token representing the relevant functioning category, such as `[DOM_ADM]` for *Respiration functions (ADM)*. This category-aware design enables the model to capture category-specific patterns while still benefiting from joint training across all categories.

Because MedRoBERTa.nl relies on Byte-Pair Encoding (BPE) for subword tokenization, medical terms and abbreviations are often split into multiple subword units. To prevent misalignment between input tokens and gold labels, the preprocessing pipeline ensures that labels remain correctly aligned with all generated subword tokens. After prediction, a post-processing step merges subword-level predictions back to the word level, resulting in coherent and interpretable outputs.

3.3 Training

The three model types are trained on different combinations of the datasets described above, resulting in a total of eight trained models. The baseline model with nine separate classifiers is trained on the original dataset (Dataset A) and on the combined dataset containing only the original categories (Dataset B1). It is not trained on Dataset B, as no test data are available for the newly generated categories.

The unified model without category encoding is trained on all three datasets: Dataset A, Dataset B1, and Dataset B. The same training setup is used for the unified model with category encoding, allowing for a direct comparison between the two unified approaches under identical data conditions.

3.4 Evaluation

All trained models are evaluated on the same original test set using standard regression metrics. Model performance is assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), enabling a consistent comparison across models and datasets.

3.5 Error Analysis

To gain deeper insight into model behaviour and potential limitations, two complementary error analyses are conducted. First, a functioning level scale analysis examines whether differences in the scales used across functioning categories affect prediction accuracy. Second, a linguistic feature analysis investigates the influence of negations, intensifiers, and minimizers on model predictions, with the aim of identifying systematic error patterns related to linguistic phenomena.

3.6 Conclusions

The results from the evaluation and error analyses are combined to address the research question. This includes assessing cross-category consistency in language use, evaluating the impact of generated data on model performance, and determining whether the inclusion of category-specific tokens improves the effectiveness of a unified modeling approach.

Chapter 4

Data and Annotations

This chapter presents the two training sets, three development sets, and one test set used in the study, and provides detailed descriptions of the processes of data collection, generation, annotation, and validation.

4.1 Clinical Data

This study uses data collected from Electronic Health Records (EHRs) of patients admitted to either Amsterdam University Medical Centre (Amsterdam UMC) or Vrije Universiteit Medical Centre (VUMC). These medical notes consist of various types of Dutch-language texts, such as admission notes, progress reports, and discharge summaries. Importantly, this type of data differs substantially from general-domain Dutch texts: it contains extensive medical jargon, numerous abbreviations, and non-standard or fragmented sentence structures. Therefore, training medical classifiers on real clinical notes is preferable to using general Dutch corpora, as it ensures that the models learn from authentic language used in the clinical context.

4.1.1 ICF Categories

The functioning information in the dataset was annotated according to the *International Classification of Functioning, Disability and Health* (ICF) framework, developed by the World Health Organization (WHO) (World Health Organization, 2025). The ICF provides a standardized structure for describing health and functioning across physical, mental, and social domains.

In this project, the original dataset included nine ICF-based categories that capture key aspects of patients' functional status: *Energy level (ENR)*, *Attention functions (ATT)*, *Emotional functions (STM)*, *Respiration functions (ADM)*, *Exercise tolerance functions (INS)*, *Weight maintenance functions (MBW)*, *Walking (FAC)*, *Eating (ETN)*, and *Work and employment (BER)*.

After extending the dataset with additional annotations generated by OpenAI, eight new categories were introduced: *Moving around using equipment (MAE)*, *Handling stress and other psychological demands (HSP)*, *Higher-level cognitive functions (HLC)*, *Sensations of pain (SOP)*, *Family relationships (FML)*, *Sleep functions (SLP)*, *Changing basic body position (CBP)*, and *Hearing functions (HRN)*.

Each sentence in the dataset may contain one or more of these categories. For every identified category, a functioning level was assigned on a numerical scale (either 0–4 or

ICF code	Abbreviation	Category	Scale
Old categories			
b1300	ENR	Energy level	0–4
b140	ATT	Attention functions	0–4
b152	STM	Emotional functions	0–4
b440	ADM	Respiration functions	0–4
b455	INS	Exercise tolerance functions	0–5
b530	MBW	Weight maintenance functions	0–4
d450	FAC	Walking	0–5
d550	ETN	Eating	0–4
d840–d859	BER	Work and employment	0–4
Newly added categories			
d465	MAE	Moving around using equipment	0–4
d240	HSP	Handling stress and other psychological demands	0–4
b164	HLC	Higher-level cognitive functions	0–4
b280	SOP	Sensations of pain	0–4
d760	FML	Family relationships	0–4
b134	SLP	Sleep functions	0–4
d410	CBP	Changing basic body position	0–4
b230	HRN	Hearing functions	0–4

Table 4.1: Overview of the ICF categories in the project.

0–5) reflecting the degree of limitation or difficulty. A score of 0 represents a complete problem or disability, while 4 (or 5 for some functions) represents no difficulty. Table 4.1 provides an overview of all categories, their abbreviations, and their corresponding functioning scales.

4.2 Data Sources

The experiments in this project were carried out using three datasets. The first dataset consisted solely of the original data (Kim et al., 2022), while the second and third datasets combined the original data with additional AI-annotated medical notes.

4.2.1 Original dataset

The original dataset (Kim et al., 2022) consists of clinical notes from the Amsterdam UMC and VUMC EHRs collected between 2017 and 2020. It includes both COVID-19 and non-COVID-19 patients.

During the data selection process, batches of notes were sampled weekly to ensure balanced representation across functional levels. Three parameters guided the selection:

1. the type of clinical note (for example, progress reports and discharge instructions),

2. keyword-based filtering to find relevant content,

3. the proportion of notes from COVID-19 patients.

Early analyses revealed that certain categories were overrepresented while others were rare. Adjustments to the selection parameters improved the overall balance, although some imbalance remained. The keyword-based filtering did not increase the total number of labeled sentences but slightly improved coverage of less frequent functional categories.

Data Annotation

After data collection, the notes were manually annotated. Of the 6,000 clinical notes collected, approximately 90% (around 5,400) were deemed relevant and further analyzed. These notes contained roughly 286,000 sentences, of which only about 5% included at least one functional category label. The final annotated dataset thus contained approximately 15,000 labeled sentences.

Annotation was performed by six trained (para)medical students who were native speakers of Dutch. To assess inter-annotator agreement (IAA), two metrics were used: the F1-score for category assignment and the Mean Absolute Error (MAE) for level assignment. The results revealed that the annotators found some categories more challenging than others, with category-level F1-scores ranging from 0.34 to 0.78. The most problematic categories were *Exercise tolerance functions (INS)*, *Work and employment (BER)*, and *Eating (ETN)*. The INS category showed overlap with other categories, while the reasons for the lower agreement on BER and ETN were less clear.

However, while the annotators struggled with category assignment, agreement on functioning levels was notably higher. The MAE for all categories was below 1, indicating that once annotators agreed on a category, they typically also agreed on the corresponding level.

4.2.2 AI-Generated Data

Data Choice

In the original study (Kim et al., 2022), some regression models performed worse than others. While part of this variation can be attributed to the linguistic complexity of Dutch clinical texts, another contributing factor was the underrepresentation of certain categories in the dataset, particularly *Attention functions (ATT)* and *Work and employment (BER)*. To address this imbalance and increase the number of sentences per category, additional data were required.

Although human annotation remains the most reliable approach, it is both time- and cost-intensive. Therefore, I generated functioning level annotations for a new set of previously unused medical notes using OpenAI. The dataset used for this purpose was created by my classmate, Shutao Chen, who conducted her thesis research in parallel (Chen, 2025). This dataset was constructed by combining two sources: a dataset of previously collected sentences that had already been manually annotated for the original ICF categories, and a newly extracted dataset retrieved by searching the 2023 Amsterdam UMC notes for medically meaningful keywords related to the 8 new ICF categories. After merging these two datasets into a single pool, she used OpenAI to assign category labels to the merged data.

Filtering Sentences and Generating Labels

While Chen selected the data and generated the corresponding ICF categories for each sentence, I expanded her dataset by prompting OpenAI to generate functioning level annotations for these categories. Her final dataset contained 40,414 sentences; however, most sentences did not contain any relevant ICF categories and were therefore excluded from this project.

A custom Python script was used to automatically annotate Dutch clinical sentences with ICF categories using Azure OpenAI's GPT-4 model. This script reads sentences from a CSV file, constructs structured prompts with pre-defined examples and category definitions, and sends the data to the model in batches of 50. The model's responses are parsed, validated, and saved in JSON format, producing category annotations for each sentence. The few-shot prompt was adapted from an earlier version developed by my classmate Shutao Chen (2025), which was maintained at a temperature of 0 because it yielded the most accurate results. Since I had less freedom to experiment with prompts - particularly because no examples of functioning levels were available for the newly added categories - I extended the existing few-shot prompt rather than testing multiple alternative prompts. This approach allowed for efficient annotation while leveraging a prompt already proven effective in a similar annotation setting.

Two subsets of sentences were prepared for functioning level generation. The first subset included 5,172 sentences annotated with the original ICF categories. The second subset comprised 7,103 sentences annotated with newly added categories. The same level-generation process was applied to both subsets.

Because many sentences contained multiple categories, some appeared in both subsets, leading to partial overlap between the two generated files. For example, sentences that included both old and new categories were annotated more than once. This overlap was resolved during the validation stage, when the data were shuffled and separated into training and development sets, as well as into individual category files. After removing duplicates, the final dataset contained 11,701 unique sentences.

Importantly, I used all available data to generate category labels. Consequently, after combining the datasets, the overall data volume increased for all categories, not only for those that had previously underperformed. Some categories gained a substantial number of new sentences (such as *Sensations of pain (SOP)*), while others received only a few additional examples, such as *Eating (ETN).* Moreover, despite the increased data volume, the categories suffered from level imbalance. A detailed overview of the resulting level distributions can be found in Section 4.3.3.

Another important note concerns the availability of examples for the new categories. Since no previous dataset existed for these newly added categories, I also lacked example sentences to guide the model. When generating labels, OpenAI was provided with two examples per original category, covering different functioning levels. However, no such examples were available for the new categories. As a result, OpenAI likely drew on examples from semantically related categories during generation. This difference should be considered when evaluating the overall quality and consistency of the AI-generated annotations.

Data Validation

Before the generated dataset was used for model training, it underwent validation by medical professionals. Since Chen's dataset (Chen, 2025) already contained OpenAI-

Abbreviation	ICF Category	Sentences	Category Correct (%)	Level Correct (%)
FAC	Walking	15	80.00	83.33
HSP	Handling stress and other psychological demands	10	80.00	100.00
ADM	Respiration functions	8	62.50	80.00
SOP	Sensations of pain	8	100.00	100.00
ENR	Energy level	6	50.00	66.67
BER	Work and employment	6	83.33	80.00
STM	Emotional functions	6	66.67	100.00
HLC	Higher-level cognitive functions	6	83.33	80.00
MBW	Weight maintenance functions	6	50.00	100.00
SLP	Sleep functions	6	100.00	100.00
ETN	Eating	6	66.67	100.00
MAE	Moving around using equipment	6	16.67	0.00
INS	Exercise tolerance functions	5	80.00	100.00
FML	Family relationships	5	100.00	60.00
CBP	Changing basic body position	5	100.00	60.00
ATT	Attention functions	5	60.00	100.00
HRN	Hearing functions	5	40.00	100.00
Average			**73.57**	**82.10**

Table 4.2: Performance per ICF Category during data validation (sorted by number of sentences).

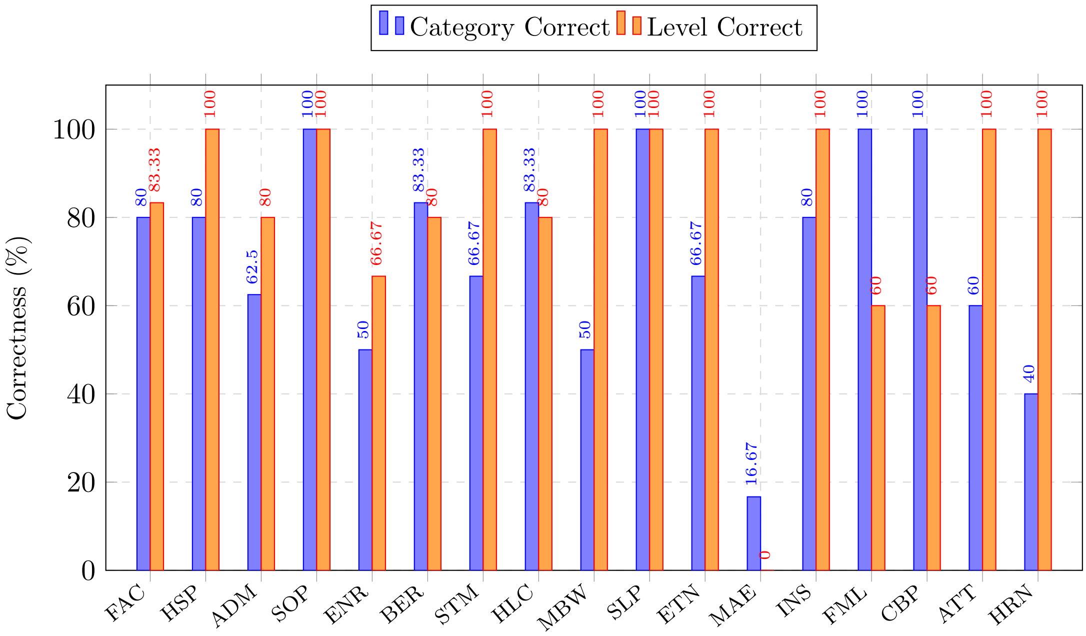

Figure 4.1: Category and Level Correctness per ICF Category (sorted by number of sentences).

generated category labels, not all sentences were correct. In fact, 61.5% of the OpenAI-generated labels matched the expert-validated labels in the validation set. Consequently, some of the level annotations generated on top of this data were also erroneous.

For validation, I randomly selected a subset of 86 sentences from both the old- and new-category datasets, ensuring that each category was represented at least five times. Because some sentences contained multiple categories, the validation set comprised 114 annotated category instances. For example, *Walking* (FAC) appeared 15 times, and *Handling stress and other psychological demands* (HSP) appeared 10 times. The number of sentences and the correctness per category are shown in Table 4.2.

The validation was performed by four Dutch-speaking medical professionals who were actively involved in the project, combining medical expertise with a deep understanding of its goals and methodology. Two annotators reviewed 29 sentences (33.7%), one reviewed 61 sentences (70.9%), and one annotated the full validation dataset (100%). The 29 sentences annotated by all four were used to calculate the Inter-Annotator Agreement (IAA) score, while the detailed analysis was based on the annotations of the professional who reviewed the entire set.

Among the 86 validated sentences, the annotators marked 11 (12.8%) as *"also"*, indicating that OpenAI had missed one or more relevant categories. Sentences in which the model incorrectly identified all categories and levels were marked as *"none"*. The annotators corrected both types of errors and could optionally include comments to justify their decisions.

Based on the full set of annotations, 73.57% of the 114 category sentences were found to contain correct categories, and of these, 82.10% also had correct functioning levels. Table 4.2 and Figure 4.1 summarize these results, showing that correctness varied substantially across categories. For instance, *Sensations of pain (SOP)* achieved 100% correctness for both category and level, while *Moving around using equipment (MAE)* showed the weakest performance, with only 16.67% category correctness (one out of six) and 0% level correctness.

Inter-Annotator Agreement (IAA) Score

To assess consistency between annotators, I calculated Fleiss' Kappa scores for both categories and levels. The IAA for categories was 0.68, indicating substantial agreement, while the score for levels was 0.59, suggesting moderate agreement. The lower score for levels is expected, as only 12 sentences met the criterion of all annotators agreeing on the correct category. This limited subset means that the level-related score should be interpreted as exploratory rather than definitive.

It is worth noting that the original study by Kim (2022) used the F1-score for categories and the Mean Absolute Error (MAE) for levels. Therefore, I also calculated the MAE for levels in my study, as this was its main focus. The resulting score was 0.26, indicating fairly good agreement among the annotators. However, this measure, as well as the result of the Fleiss' Kappa, should be interpreted with caution.

Data Shuffling

After validation, I merged the datasets for the old and new categories, shuffled the data, and separated it into individual category files. As discussed in Section 4.2.2, overlapping sentences were removed. One category, *Handling stress and other psychological demands (HSP)*, contained a single annotation error where the category code was incorrectly assigned as *b240* instead of *d240*. Because no level was generated for this sentence, it was excluded from further analysis.

Each category file was subsequently split into training and development subsets using a 90/10 ratio. Although the nine original regression models could not yet be trained on the new categories due to the absence of test data, this preparation step was performed to support future work within the A-PROOF project. The final distribution of sentences per category in the AI-generated dataset is presented in Figure 4.2. In addition, an overview of the level distributions for each category is provided in Section 4.3.3.

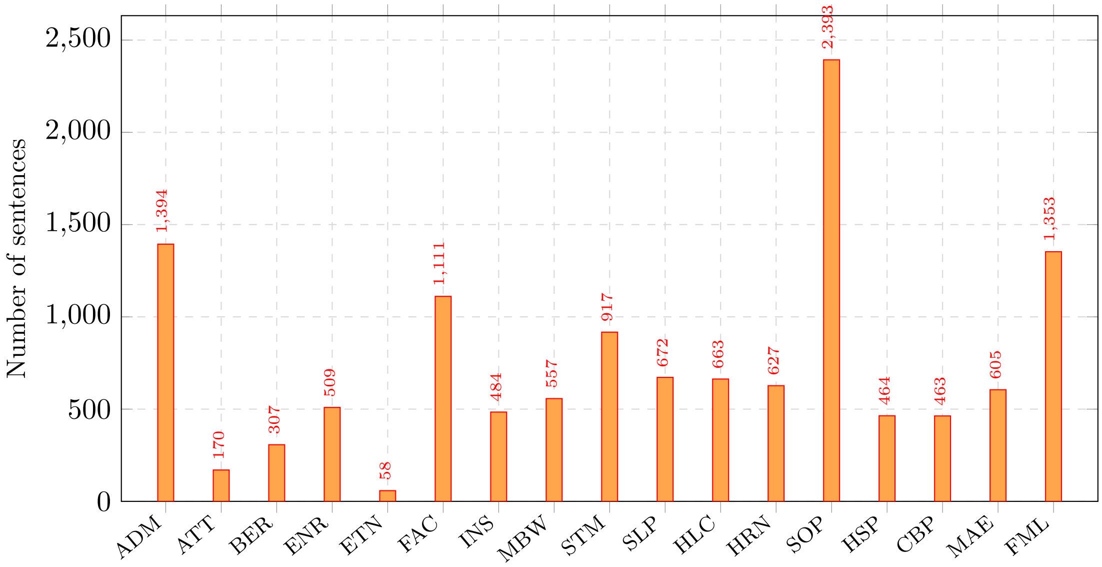

Figure 4.2: Number of sentences per ICF category generated with OpenAI (later split into training and development sets).

4.3 Training Data

To train the models, I used two datasets. The first dataset was the original dataset created by Jenia Kim (Kim et al., 2022), hereafter referred to as Dataset A. The second dataset was a combination of the original data and additional medical notes, which included sentences with categories and levels generated by OpenAI (Dataset B). This combined dataset was further divided into two parts: one containing only the categories present in the original dataset (Dataset B1), and another containing the categories that did not appear in Kim's dataset (Dataset B2). The following sections describe each dataset in detail.

4.3.1 Dataset A: Original Training Data

The original training dataset, used to train the nine level classifiers by Kim et al. (2022), contained 13,572 sentences. Table 4.3 shows the descriptive statistics per ICF category. The category *ADM* (*b550 Respiration functions*) had the highest number of sentences (5,233 sentences) representing 38.56% of the entire dataset. Several categories also contained more than 1,000 sentences, including *ETN* (*d550 Eating*), *STM* (*b152 Emotional functions*), *INS* (*b455 Exercise tolerance functions*), *FAC* (*d450 Walking*), and *ENR* (*b1300 Energy level*). In contrast, categories such as *ATT* (*b140 Attention functions*) and *BER* (*d840–d859 Work and employment*) were underrepresented, with fewer than 300 examples each.

Despite this imbalance, sentence lengths across categories were relatively similar, with an average ranging from 11.81 to 14.63 words and a median between 9 and 12 words. All categories included very short (one word) and longer, more complex sentences (up to 107 words). This consistency in sentence length distribution suggests that differences in model performance across categories were not likely caused by differences

ICF Category	Number of Sentences	Average Length	Median Length	Min. Length	Max. Length
ADM	5,233	11.81	9	1	97
ATT	251	14.63	12	1	107
BER	216	13.56	12	1	52
ENR	1,005	12.81	10	1	107
ETN	2,491	12.37	10	1	98
FAC	1,086	13.16	12	1	67
INS	1,104	13.31	11	1	68
MBW	766	13.19	10	1	93
STM	1,420	14.57	12	1	107
Total	**13,572**				

Table 4.3: Descriptive statistics per ICF category for Dataset A (original training data). Sentence length is measured in words.

in text length.

4.3.2 Dataset B: Combined Training Data

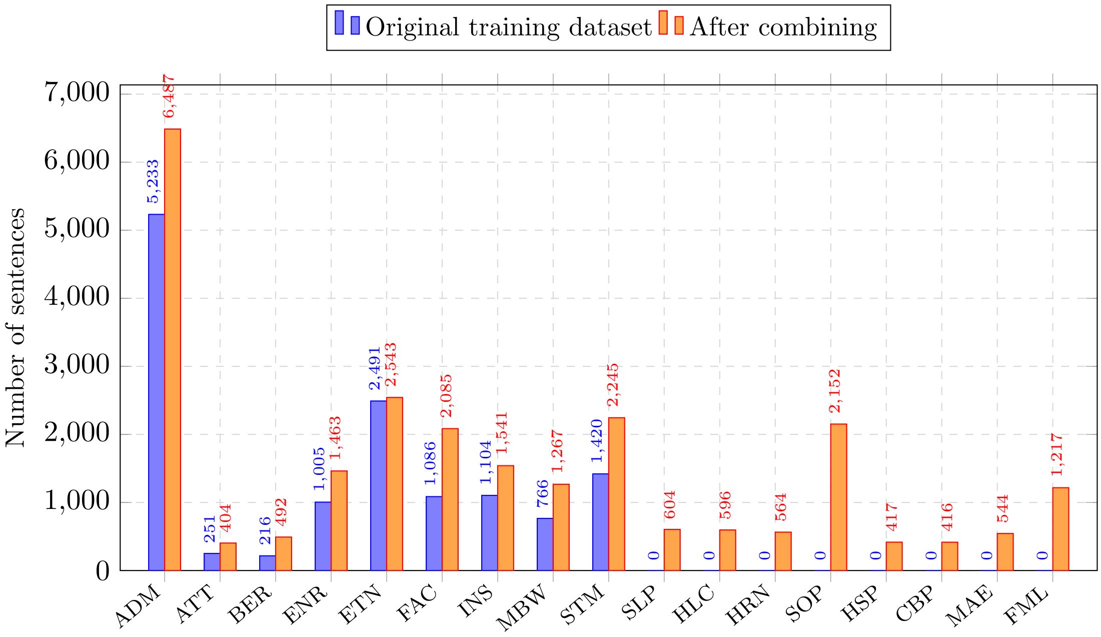

Figure 4.3: Number of sentences per ICF category before and after combining training datasets.

The combined dataset (Dataset B) was created by merging the original training data (Dataset A) with AI-generated data. After merging, the dataset was divided into two subsets: Dataset B1, containing the original nine ICF categories, and Dataset B2, containing eight newly defined categories introduced during data generation.

As shown in Figure 4.3, the first nine categories on the left (those with both blue and orange bars) represent Dataset B1, while the remaining eight correspond to Dataset B2. The categories that grew the most after combination were ADM, FAC, SOP, and FML, each gaining over 1,000 new sentences. In contrast, categories such as ATT, BER, and ETN grew only slightly, indicating that these functional domains are less frequently mentioned in clinical notes. For instance, *Walking (FAC)* frequently co-occurs with other ICF categories, such as *Exercise tolerance functions (INS)*, which partly explains its higher representation.

Dataset B1: Combined Training Data (Old Categories)

Dataset B1 contained the original nine ICF categories from Dataset A, expanded with AI-generated sentences. This dataset was used to re-train the original nine level classifiers. In addition, it was used in two alternative modeling approaches: (1) training a single model on all categories simultaneously, and (2) first encoding each category separately and then training a combined model using these encodings. The inclusion of AI-generated data allowed me to investigate whether adding synthetic examples could improve the performance and generalization of the classifiers on the original test data.

Dataset B2: Combined Training Data (New Categories)

Dataset B2 consisted of the newly introduced ICF categories that were not present in the original dataset. Because no test data were available for the functioning levels of these new categories, Model 1 (5.3), when trained on these categories, could not be evaluated. However, the new categories were included in the training process for the second and third modeling approaches described above. Incorporating the new categories in training allowed an evaluation of whether learning additional functional domains could improve the models' performance in predicting the original nine categories on the original test dataset.

4.3.3 Level Distribution

While the number of sentences per category in the training data was imbalanced, an equally important factor to consider is the distribution of levels within each category.

Old Categories

Figures A.1 to A.18 in Appendix A show the level distribution for each category. For each category, two graphs are presented: the left graph, with orange columns, represents the level distribution in the original training dataset (Dataset A), while the right graph, in blue, represents the level distribution in the combined dataset (Dataset B). Importantly, both graphs maintain the same values on the y-axis to ensure an accurate visual comparison.

Inspection of the graphs reveals that the level distribution in most categories is highly imbalanced. Several categories, namely ATT, ENR, FAC, INS, and STM, show a pronounced peak in the middle levels, indicating that most sentences describe mild limitations rather than complete disability or no difficulties. In contrast, BER has the majority of sentences indicating no problem at all (95 in Dataset A and 189 in Dataset

B), while MBW is relatively balanced, except for level 0 (complete disability), which has 14 sentences in Dataset A and 17 in Dataset B. The best-balanced category is ADM, which also has substantially more examples in the datasets compared to other categories.

New Categories

The new categories did not appear in the original dataset (Dataset A). Therefore, Figures A.19 to A.26 in Appendix A show the level distributions only for Dataset B, with all graphs displayed in blue. Since the datasets are not compared side by side in this case, the y-axis values differ depending on the number of sentences in each category.

A closer inspection reveals that the level distribution for the new categories is also heavily skewed. While CBP, HLC, HRN, HSP, MAE, and SOP peak around the middle levels, FML is more represented in levels 2–4. In contrast, SLP, despite having very few sentences with level 0 (only four sentences), is considerably less skewed compared to the other categories.

4.4 Development Data

This section presents three development sets: the original one and two combined versions.

4.4.1 Dataset C: Original Development Data

ICF Category	Number of Sentences	Average Length	Median Length	Min. Length	Max. Length
ADM	440	12.02	9	1	59
ATT	23	14.48	11	4	37
BER	29	17.10	12	4	94
ENR	107	13.69	12	2	40
ETN	236	11.89	10	1	64
FAC	124	15.51	11	1	117
INS	132	13.04	11.5	1	53
MBW	98	11.40	9	1	41
STM	148	15.26	12	2	117
Total	**1,337**				

Table 4.4: Descriptive statistics per ICF category for Dataset C (original development data). Sentence length is measured in words.

The original development dataset (Dataset C) was considerably smaller than the training set but maintained similar category distributions. ADM was the most represented category with 440 sentences. Other categories with more than 100 sentences included ETN (236), STM (148), INS (132), FAC (124), and ENR (107). ATT (23) and BER (29) were the least represented, reflecting the distribution patterns observed in the training data (Table 4.4).

Similar to the original training data, the average and median sentence lengths across categories were relatively stable, ranging from 11.4 to 17.1 words on average and between 9 and 12 words for the median. All categories included both very short sentences (from one to four words) and longer, more complex sentences, with some extending up to 117 words.

4.4.2 Dataset D: Combined Development Data

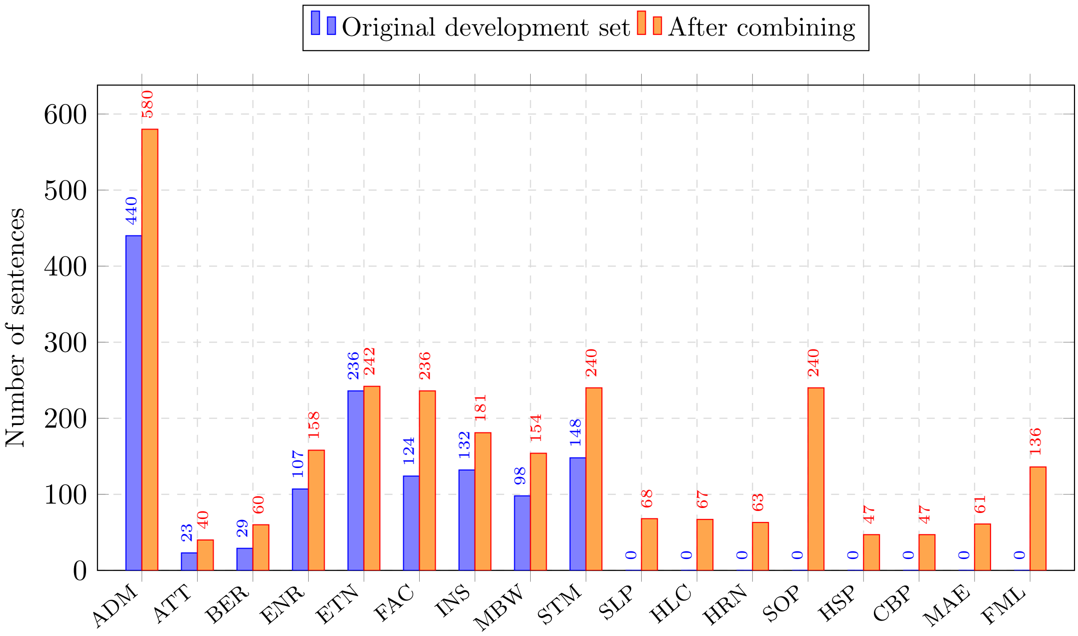

Figure 4.4: Number of sentences per ICF category before and after combining development datasets.

Dataset D was created by merging the original development set with AI-generated data, analogous to Dataset B in the training data. After merging, the dataset was split into two subsets: Dataset D1, containing the original nine ICF categories, and Dataset D2, containing eight newly introduced categories.

Figure 4.4 shows the number of sentences per category before and after combining the datasets. Categories ADM, FAC, SOP, and FML experienced the largest increases, whereas some categories, such as ETN, gained only a few additional sentences (6). This pattern reflects the distribution trends observed in the combined training data, where some categories are naturally more prevalent in clinical notes.

Dataset D1: Combined Development Data (Old Categories)

Dataset D1 contained the original nine ICF categories from Dataset C, expanded with AI-generated sentences. This subset can be primarily used for evaluating model performance during training. It allows assessment of whether adding synthetic data improves predictions for the original categories, both for re-trained classifiers and for alternative modeling approaches, such as the all-categories model and the encoded-category model.

Dataset D2: Combined Development Data (New Categories)

Dataset D2 included the eight newly introduced categories that were absent in the original development set. While these categories cannot be used to directly evaluate the original classifiers due to the lack of corresponding test data, they can be incorporated during training of the alternative modeling approaches. Using these new categories in development data help monitor whether expanding the model's coverage of functional domains improves performance on the original nine categories.

4.5 Test Data

All three models in this study, including the runs with different datasets, were evaluated on the same original test set to ensure comparing the results afterwards.

4.5.1 Dataset E: Original Test Data

ICF Category	Number of Sentences	Average Length	Median Length	Min. Length	Max. Length
ADM	421	11.62	10	1	60
ATT	32	16.25	14	3	41
BER	26	12.35	11	3	33
ENR	100	11.99	10.5	2	35
ETN	183	12.39	10	1	67
FAC	139	14.45	11	1	78
INS	136	13.33	11	2	47
MBW	60	13.27	12.5	1	45
STM	155	13.64	12	1	50
Total	**1,252**				

Table 4.5: Descriptive statistics per ICF category for Dataset E (original test data). Sentence length is measured in words.

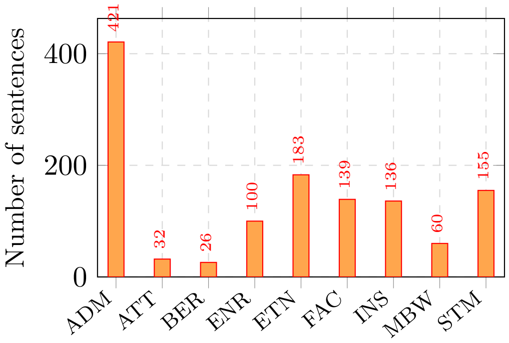

Figure 4.5: Number of sentences in Dataset E (Test Data).

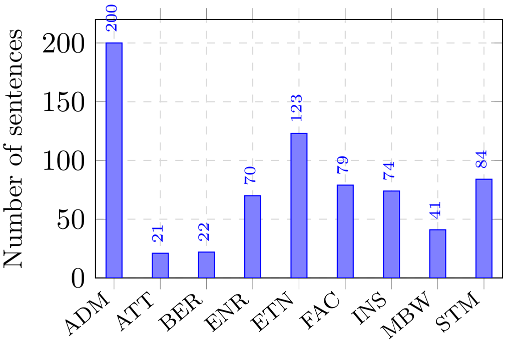

Figure 4.6: Number of notes in Dataset E (Test Data).

Similar to the original training and development datasets, the average sentence length ranged between 11.62 and 16.25 words per category. The shortest sentences contained between one and three words, while the longest sentences ranged from 33 to 78 words. The full statistics can be found in Table 4.5. The distribution of sentences and notes per category is shown in Figures 4.5 and 4.6.

Chapter 5

Models

This chapter presents the models used in this study, and provides details on their training loops and general training environment.

5.1 Data Preparation

When preparing the original datasets (Kim et al., 2022), referred to in this study as Datasets A, C, and E, the researchers first anonymized each note and then split the notes into sentences using `spaCy`[1]. Consequently, no additional preprocessing was required for this study. The same applies to the OpenAI-generated data: these notes were also anonymized and sentence-segmented prior to use. All remaining data preparation steps for these datasets are described in detail in Section 4.2.2.

5.2 Models Overview

Model	Dataset A (Original)	Dataset B1 (Old Categories)	Dataset B (All Categories)	Total Models
Model 1 (Original Nine Regression Classifiers)	✓	✓		2
Model 2 (Combined Categories)	✓	✓	✓	3
Model 3 (Combined & Encoded Categories)	✓	✓	✓	3
Total Models per Dataset	3	3	2	**8**

Table 5.1: Models and datasets used in this thesis.

In this project, three different model setups were used to predict functioning levels from Dutch clinical notes. Model 1 was trained on two datasets, while Models 2 and 3 were trained on three datasets: the original dataset and two combined datasets, as described in Section 4 (Table 5.1). Using multiple datasets allows for a comparison of model performance on the original data versus enriched, combined datasets. The three models are:

1. **Model 1: Original Nine Regression Classifiers** The system developed by Kim et al. (2022), serving as a benchmark,

2. **Model 2: Combined Categories** A variant where functioning categories were combined to reduce the number of independent classifiers,

[1] `https://spacy.io`

3. **Model 3: Combined & Encoded Categories** An extended version of Model 2 that includes encoded categories to facilitate learning across similar domains.

Each model follows a similar training procedure, adapted to the specifics of the dataset and category configuration. They are described in detail in the sections below.

5.3 Model 1: Original Nine Regression Classifiers

The first model setup in this project was based on the original system developed by Kim et al. (2022). The initial code was obtained from the GitHub repository[2]. While file paths were adapted to fit the current project structure, all other parts of the implementation, including the hyperparameter configuration, were preserved to ensure reproducibility. Starting with the original model allowed me to verify that this setup could re-produce the results reported by Kim et al. (2022), providing a reliable benchmark for later experiments with combined datasets.

The system consists of nine independent regression classifiers, each trained to predict the level of functioning within a specific domain: *Respiration functions (ADM)*, *Attention functions (ATT)*, *Work and employment (BER)*, *Energy level (ENR)*, *Eating (ETN)*, *Walking (FAC)*, *Exercise tolerance functions (INS)*, *Weight maintenance functions (MBW)*, and *Emotional functions (STM)*. Each domain-specific model was fine-tuned using the `SimpleTransformers` library and configured for regression (`num_labels = 1`), producing continuous predictions corresponding to functioning levels. The models were initialized with the pre-trained `MedRoBERTa.nl` weights from Hugging Face [3].

5.3.1 Training Loop

For each domain, the code executes the following steps:

1. Load the domain-specific training and evaluation datasets (`train.pkl` and `dev.pkl`),
2. Convert the datasets to `pandas DataFrame` objects if necessary,
3. Retrieve model arguments from the JSON configuration file,
4. Initialize the Transformer model with `num_labels = 1`,
5. Fine-tune the model on the training data using `SimpleTransformers`,
6. Save the fine-tuned model to the specified output directory.

5.4 Model 2: Combined Categories

The second model setup used in this project combines all functioning domains into a single regression model, predicting the functioning level across domains simultaneously. The initial code was adapted from the same repository as Model 1. While the overall structure of the training script remained the same, modifications were made to combine the domain-specific datasets into a single dataset for joint training. This approach was designed to explore whether multi-domain training could improve, or at least keep the performance stable by using less classifiers at the same time.

[2] `https://github.com/cltl/a-proof-zonmw/tree/main/clf_levels`

[3] `https://huggingface.co/CLTL/MedRoBERTa.nl`

5.4.1 Training Loop

The script executes the following steps for Model 2:

1. Load all domain-specific training and evaluation datasets (`train.pkl` and `dev.pkl`),
2. Convert datasets to `pandas DataFrame` objects if necessary and add a `domain` column,
3. Concatenate all domain datasets into a single combined dataset and shuffle the rows,
4. Save the combined training and evaluation datasets in `.pkl` format,
5. Retrieve model arguments from the JSON configuration file,
6. Initialize the Transformer model with `num_labels = 1`,
7. Fine-tune the model on the combined training data using `SimpleTransformers`,
8. Save the fine-tuned model to the specified output directory.

5.5 Model 3: Combined & Encoded Categories

The third model extends the combined-domain approach by explicitly encoding domain information in the input text using special tokens. Each domain was mapped to a unique token (for example, `[DOMAIN_ADM]` for *Respiration functions (ADM)*), which was pre-pended to each sentence. This approach allows the model to learn domain-specific patterns while training on a single combined dataset. The initial code was adapted from the same repository as Model 1 and 2.

5.5.1 Training Loop

The script executes the following steps for Model 3:

1. Load all domain-specific training and evaluation datasets (`train.pkl` and `dev.pkl`),
2. Convert datasets to `pandas DataFrame` objects if necessary and add a `domain` column,
3. Concatenate all domain datasets into a single combined dataset and shuffle the rows,
4. Pre-pend the corresponding domain token to each sentence in both training and evaluation sets,
5. Save the processed datasets in `.pkl` format,
6. Retrieve model arguments from the JSON configuration file,
7. Initialize the Transformer model with `num_labels=1`, update its tokenizer with the special tokens, and resize the model embeddings if new tokens are added,
8. Fine-tune the model on the token-encoded training data using `SimpleTransformers`,
9. Save the fine-tuned model to the specified output directory.

5.6 Training Environment and Procedure

All models were trained on a high-performance computing (HPC) cluster using the SLURM workload manager. Each model configuration (whether domain-specific, combined, or token-encoded) was submitted as an independent SLURM job, with allocated GPU resources (NVIDIA H100), memory, and runtime limits. Standard output and error logs were redirected to dedicated files to facilitate monitoring and later inspection.

Training was managed via Python scripts that handled data loading, model initialization, and fine-tuning. Each script accepted several command-line arguments. The parameters included:

- `--datapath`: relative path to the dataset directory,
- `--doms`: list of domains to train,
- `--config`: JSON configuration file containing model hyperparameters,
- `--model_type`: type of Transformer model (for example, `roberta`),
- `--model_name`: name or local path of the pre-trained model (`MedRoBERTa.nl`),
- `--clas_unit`: classification unit (`sent` or `note`),
- `--train_on` and `--eval_on`: filenames of the training and evaluation sets.

The training and evaluation datasets were stored in structured directories in `.pkl` format. Each file contained two crucial columns: `text`, holding the sentence or note text, and `labels`, containing numeric values representing functioning levels.

Default hyperparameters were applied consistently across all models: the AdamW optimizer with a learning rate of 4e-5, batch size of 8, and one training epoch.

This setup ensured reproducibility and allowed for systematic monitoring of training across different domains and datasets.

Chapter 6

Results

This chapter outlines the evaluation procedure and the metrics used to assess model performance, and presents the results obtained from training the three models:

1. **Model 1**: the original nine regression models;
2. **Model 2**: the combined categories model;
3. **Model 3**: the combined and encoded categories model.

Each model was trained and evaluated on two datasets: **Dataset A** (Original Training Data) and Dataset B (Combined Training Data). Importantly, Dataset B was used in two ways: as **Dataset B** (All categories) and as **Dataset B1** (Only Old Categories). All models were evaluated at both the **sentence** and **note levels**.

6.1 Evaluation

The fine-tuned regression models were evaluated using Python scripts. Each script runs the model on the same test set (Dataset E), calculates standard regression metrics (Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE)) and saves the model outputs as well as the incorrectly predicted instances for further analysis. The scripts can be easily customized through command-line parameters. It is possible specify the data path, the domains to evaluate, the type of pre-trained model, the model directory, the classification unit (sentence or note), and the evaluation dataset. For combined-domain models, the scripts also pre-pend domain-specific tokens to the input text, allowing the models to use domain information when making predictions. This ensures a consistent evaluation setup whether models are trained on individual domains or on a combined dataset.

6.1.1 Evaluation Metrics

The fine-tuned regression models were evaluated using three standard metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)[1]. All three metrics measure the difference between predicted values $\hat{y}_i$ and true values y_i over n instances, but they differ in how they penalize errors.

[1] `https://scikit-learn.org/stable/modules/model_evaluation.html`

Mean Absolute Error (MAE)

MAE calculates the average magnitude of the errors, ignoring their direction. It is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where n is the total number of observations, y_i is the true value for the i^{th} observation, and $\hat{y}_i$ is the predicted value. Lower MAE values indicate better model performance.

Mean Squared Error (MSE)

MSE computes the average of the squared differences between predicted and true values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here, n, y_i, and $\hat{y}_i$ have the same meaning as in MAE. By squaring the errors, MSE penalizes larger deviations more heavily than smaller ones. It is sensitive to outliers, which can strongly influence the evaluation.

Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and expresses the error in the same units as the original values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Again, n is the total number of observations, y_i is the true value, and $\hat{y}_i$ is the predicted value. RMSE combines the benefits of MSE's sensitivity to larger errors with interpretability in the original scale of the data. Lower RMSE values indicate more accurate predictions.

6.2 Sentence-level predictions

This section presents sentence-level evaluation results obtained from the three models across the different datasets.

6.2.1 Models 1: Original Nine Regression Classifiers

Dataset A: Replicating the Original Results

	Current Model			Original Model		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.50	0.57	0.75	0.48	0.55	0.74
ATT	0.97	1.32	1.15	0.99	1.35	1.16
BER	1.60	3.57	1.89	1.56	3.06	1.75
ENR	0.52	0.51	0.72	0.48	0.49	0.70
ETN	0.56	0.58	0.76	0.59	0.65	0.81
FAC	0.63	0.77	0.88	0.70	0.91	0.95
INS	0.71	0.81	0.90	0.69	0.80	0.89
MBW	0.74	0.83	0.91	0.81	0.83	0.91
STM	0.67	0.83	0.91	0.76	1.03	1.01

Table 6.1: Comparison of evaluation results for the current regression models and the original models.

Model 1 was first trained on the original Dataset A to replicate the baseline results. As shown in Table 6.1, the outcomes closely matched the original ones, confirming that the training conditions and hyperparameters were consistent with the original setup.

Dataset B1: Combined Training Data - Old categories

	Model 1: Original Dataset A			Model 1: Combined Dataset B1		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.50	0.57	0.75	0.48	0.54	0.74
ATT	0.97	1.32	1.15	0.86	1.06	1.03
BER	1.60	3.57	1.89	1.50	3.02	1.74
ENR	0.52	0.51	0.72	0.48	0.48	0.69
ETN	0.56	0.58	0.76	0.60	0.65	0.81
FAC	0.63	0.77	0.88	0.59	0.66	0.81
INS	0.71	0.81	0.90	0.63	0.69	0.83
MBW	0.74	0.83	0.91	0.87	1.11	1.05
STM	0.67	0.83	0.91	0.72	0.92	0.96

Table 6.2: Comparison of evaluation results for Model 1 with the original and combined datasets. Orange highlighting indicates better performance.

Due to the lack of test data for the new categories, the nine regression classifiers were trained only on Dataset B1 (which contained the original categories).

The results show consistent but modest improvements in most categories when trained on the combined dataset. While the differences in MAE, MSE and RMSE are small, six out of nine categories improved, suggesting that including AI-generated data had an overall positive effect (Table 6.2). For instance, *Attention functions (ATT)* improved from MAE 0.97 to 0.86, MSE 1.32 to 1.06, and RMSE 1.15 to 1.03. Similarly, *Work and employment (BER)* improved from MAE 1.60 to 1.50, MSE 3.57 to 3.02, and RMSE 1.89 to 1.74.

Conversely, categories such as *Eating (ETN)*, *Weight maintenance functions (MBW)*, and *Emotional functions (STM)* performed slightly worse after data combination. MBW showed the largest decrease, with MAE increasing from 0.74 to 0.87. These results suggest that while AI-generated data generally helps, its effect varies across categories - particularly those sensitive to erroneous data.

6.2.2 Model 2: Combined Categories

	Model 2: Original Dataset A			Model 2: Combined Dataset B1 (Only Old Categories)			Model 2: Combined Dataset B (All Categories)		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.54	0.62	0.79	0.53	0.63	0.79	0.52	0.62	0.78
ATT	0.64	0.72	0.85	0.68	0.79	0.89	0.64	0.71	0.84
BER	1.24	2.34	1.53	1.03	2.07	1.44	1.02	2.02	1.42
ENR	0.55	0.58	0.76	0.52	0.52	0.72	0.55	0.60	0.78
ETN	0.46	0.51	0.68	0.50	0.46	0.68	0.52	0.51	0.71
FAC	0.75	0.97	0.99	0.73	0.99	1.00	0.69	0.90	0.95
INS	0.95	1.42	1.19	0.93	1.31	1.15	0.96	1.42	1.19
MBW	0.65	0.72	0.85	0.66	0.70	0.84	0.72	0.82	0.91
STM	0.66	0.84	0.92	0.63	0.76	0.87	0.60	0.71	0.84

Table 6.3: Comparison of evaluation results for Model 2: original, combined (old categories), and combined (all categories). Orange highlighting indicates better performance.

Model 2 (Combined Categories) was trained on Dataset A, Dataset B1, and Dataset B.

Overall, results varied across categories. For most categories - *Respiration functions (ADM)*, *Attention function (ATT)*, *Energy levels (ENR)*, *Weight maintenance functions (MBW)*, and *Emotional functions (STM)* - performance remained relatively stable across datasets. However, *Work and employment (BER)* benefited significantly from the combined datasets (B1 and B), suggesting that this category is less dependent on category-specific vocabulary and gains from a larger dataset, even with some noisy sentences.

6.2.3 Model 3: Combined & Encoded Categories

	Model 3: Original Dataset A			Model 3: Combined Dataset B1 (Only Old Categories)			Model 3: Combined Dataset B (All Categories)		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.50	0.61	0.78	0.51	0.56	0.75	0.50	0.56	0.75
ATT	0.70	0.92	0.96	0.67	0.84	0.92	0.58	0.61	0.78
BER	1.12	2.42	1.55	1.09	2.52	1.59	1.06	2.13	1.46
ENR	0.50	0.52	0.72	0.44	0.36	0.63	0.48	0.44	0.66
ETN	0.49	0.46	0.68	0.51	0.51	0.72	0.48	0.45	0.67
FAC	0.69	0.90	0.95	0.66	0.83	0.91	0.60	0.75	0.87
INS	0.74	0.94	0.97	0.83	1.03	1.01	0.69	0.80	0.89
MBW	0.66	0.69	0.83	0.67	0.72	0.85	0.65	0.65	0.81
STM	0.63	0.76	0.87	0.63	0.78	0.88	0.59	0.66	0.81

Table 6.4: Comparison of evaluation results for Model 3: original, combined (old categories), and combined (all categories). Orange highlighting indicated better performance.

Model 3 (Combined and Encoded Categories) shows clear performance differences across the three dataset variants (Table 6.4). In contrast to the original dataset and the combined dataset containing only old categories, the combined dataset including all categories consistently yields the lowest error values for almost all ICF categories. Substantial improvements are observed for *Attention functions (ATT)*, *Work and employment (BER)*, *Walking (FAC)*, and *Exercise tolerance functions (INS)*, while smaller but consistent gains are also present for most remaining categories. Only *Energy level (ENR)* shows slightly better results with the combined dataset containing only old categories. Overall, these results indicate that, when categories are encoded, adding sentences from both old and newly introduced categories generally improves model performance, suggesting that the model benefits from increased data diversity.

6.2.4 Summary

To summarize, the results clearly differ across categories (Table 6.5). Some categories, namely *Respiration functions (ADM)*, *Walking (FAC)*, and *Exercise tolerance functions (INS)*, performed better with the original model. However, the performance of Model 3 for ADM was comparable and not drastically lower. Model 2 achieved the best results for *Attention functions (ATT)*, *Work and employment (BER)*, and *Eating (ETN)*. Although the results for BER and ETN were close to those of the other models, ATT showed a substantial improvement under Model 2. This finding suggests that ATT does not rely heavily on category-specific vocabulary and benefits from being trained on a larger, cross-category dataset.

The remaining three categories - *Energy levels (ENR)*, *Weight maintenance functions (MBW)*, and *Emotional functions (STM)* - showed improvements when trained with Model 3. This may indicate that Model 3 is generally more robust, as it either substantially improves performance or maintains results comparable to the other models (with the exception of ATT, FAC, and INS).

	Model 1: Dataset A			Model 2: Dataset A			Model 3: Dataset A		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.50	0.57	0.75	0.54	0.62	0.79	0.50	0.61	0.78
ATT	0.97	1.32	1.15	0.64	0.72	0.85	0.70	0.92	0.96
BER	1.60	3.57	1.89	1.24	2.34	1.53	1.12	2.42	1.55
ENR	0.52	0.51	0.72	0.55	0.58	0.76	0.50	0.52	0.72
ETN	0.56	0.58	0.76	0.46	0.51	0.68	0.49	0.46	0.68
FAC	0.63	0.77	0.88	0.75	0.97	0.99	0.69	0.90	0.95
INS	0.71	0.81	0.90	0.95	1.42	1.19	0.74	0.94	0.97
MBW	0.74	0.83	0.91	0.65	0.72	0.85	0.66	0.69	0.83
STM	0.67	0.83	0.91	0.66	0.84	0.92	0.63	0.76	0.87

Table 6.5: Comparison of evaluation results for Models 1, 2 & 3 using the original dataset. Orange highlighting indicates better performance.

When examining each category individually, *Respiration functions (ADM)* showed minimal variation across datasets and models. As this category is the best represented in the datasets, this finding suggests that it already contains sufficient training data and does not benefit from additional sentences. Moreover, since its performance remains stable even when trained on combined datasets, it is likely that ADM does not rely heavily on category-specific vocabulary.

Another category, *Attention functions (ATT)*, significantly benefited from the combined datasets and models trained across categories. For example, in the original Model 1 trained on Dataset A, its MAE was 0.97, MSE 1.32, and RMSE 1.15. However, when trained using Model 3 on Dataset B, its performance improved markedly to MAE 0.58, MSE 0.61, and RMSE 0.78. This result suggests that ATT is not strongly category-specific and benefits greatly from larger training datasets. This is unsurprising given that, in the original Dataset A, ATT contained only 251 training instances.

A similar pattern was observed for *Work and employment (BER)*, where results improved when using combined datasets as well as Models 2 and 3. While its original results were MAE 1.60, MSE 3.57, and RMSE 1.89, the best performance was achieved with Model 3 trained on Dataset B (MAE 1.06, MSE 2.13, and RMSE 1.46). This indicates that BER is not highly sensitive to vocabulary from other categories and benefits from larger training sets, which is again expected given its small original size of 216 sentences.

The category *Energy levels (ENR)* performed similarly across datasets and models, with the exception of Model 3 trained on Dataset B1, where its MAE decreased from the original 0.52 to 0.44, MSE from 0.51 to 0.36, and RMSE from 0.72 to 0.63. This suggests that it does benefit from larger datasets, especially when the categories are not encoded.

Eating (ETN) produced relatively stable results across datasets but improved further with Model 3 trained on Dataset B1. Its MAE decreased from 0.56 to 0.48, MSE from 0.58 to 0.45, and RMSE from 0.76 to 0.67.

Walking (FAC) and *Exercise tolerance functions (INS)* proved to be challenging categories. As also noted in the original paper by Jenia Kim et al. (2022), their performance patterns remain inconsistent. For all three models, both performed better with the combined datasets, most often with Dataset B, with an exception of INS for Model

2, where Dataset B1 yielded better results. While this suggests that these categories benefit form larger datasets, they achieved the highest results using the original Model 1, where all categories where trained separately. One possible reason for this is that these categories use a different rating scale (0–5) compared to most other categories, which are rated 0–4. Consequently, they do not benefit as much from unified model approaches. Future studies may train these categories using Model 1, potentially with more professionally annotated data, as these functions appear improve when trained with more as well cross-category data. However, another possible approach would be to invesitgate their scales more attentively as this could solve the issue.

Weight maintenance functions (MBW) seem to benefit from larger datasets, as performance improved with Dataset B1 under Model 2 and with Dataset B under Model 3. However, for Model 1, results were better with the original Dataset A. This suggests that while MBW gains from more data, it is also sensitive to noisy or erroneous sentences.

Finally, *Emotional functions (STM)* showed a pattern similar to MBW. While Datasets B1 and B yielded good results under Model 2, the best performance was achieved with the combined dataset (all categories) B using Model 3. This suggests that STM, too, may benefit from category-specific and cleanly annotated data.

6.3 Note-level predictions

This section presents note-level evaluation results obtained from the three models across different datasets.

6.3.1 Model 1: Original Nine Regression Classifiers

	Model 1: Original Dataset A			Model 1: Combined Dataset B		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.40	0.36	0.60	0.39	0.35	0.59
ATT	1.02	1.05	1.22	0.89	1.16	1.08
BER	1.50	3.25	1.80	1.43	2.82	1.68
ENR	0.46	0.44	0.66	0.44	0.42	0.65
ETN	0.49	0.41	0.64	0.55	0.51	0.71
FAC	0.58	0.67	0.82	0.56	0.64	0.80
INS	0.58	0.56	0.75	0.56	0.59	0.77
MBW	0.62	0.58	0.76	0.79	0.91	0.95
STM	0.57	0.68	0.82	0.61	0.73	0.85

Table 6.6: Comparison of evaluation results for Model 1 with the original and combined datasets. Orange highlighting indicates better performance.

When evaluating Model 1 on Dataset A and Dataset B1, most categories show similar results (Table 6.6). However, the *Attention functions (ATT)* category benefits noticeably from the combined dataset: MAE decreased from 1.02 to 0.89 and RMSE from 1.22 to 1.08, while MSE slightly increased from 1.05 to 1.16, suggesting smaller but more consistent errors. *Work and employment (BER)* also improved, with MSE dropping from 3.25 to 2.82. Conversely, *Weight maintenance functions (MBW)* performed better on Dataset A; combining datasets increased MAE from 0.62 to 0.79, MSE from 0.58 to 0.91, and RMSE from 0.76 to 0.95, indicating sensitivity to dataset composition.

6.3.2 Model 2: Combined Categories

For Model 2, most categories remain stable across datasets (Table 6.7). *Work and employment (BER)* benefits from the combined dataset: MAE decreased from 1.15 to 0.87, MSE from 2.13 to 1.67, and RMSE from 1.46 to 1.29. Similarly, *Walking (FAC)* improved with combined datasets, with MAE dropping from 0.73 to 0.63, MSE from 1.00 to 0.78, and RMSE from 1.00 to 0.88, indicating the advantage of additional cross-category data.

6.3.3 Model 3: Combined & Encoded Categories

Evaluation of Model 3 (Table 6.8) across datasets reveals that training on combined datasets, especially the one including all categories, generally improved performance. The one exception was the category *Energy levels (ENR)*, for which the best-performing dataset was Dataset B1 (the combined dataset containing only old categories). This finding highlights the advantage of additional training data for most categories and

	Model 2: Original Dataset A			Model 2: Combined Dataset B1 (Only Old Categories)			Model 2: Combined Dataset B (All Categories)		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.42	0.40	0.63	0.42	0.41	0.64	0.41	0.39	0.63
ATT	0.61	0.65	0.81	0.61	0.62	0.79	0.61	0.61	0.78
BER	1.15	2.13	1.46	0.96	1.86	1.36	0.87	1.67	1.29
ENR	0.46	0.44	0.66	0.43	0.38	0.61	0.47	0.47	0.68
ETN	0.44	0.32	0.56	0.44	0.37	0.61	0.47	0.39	0.62
FAC	0.73	1.00	1.00	0.70	0.97	0.98	0.63	0.78	0.88
INS	0.90	1.24	1.11	0.87	1.08	1.04	0.90	1.20	1.10
MBW	0.63	0.63	0.79	0.60	0.59	0.77	0.65	0.67	0.82
STM	0.59	0.70	0.84	0.55	0.65	0.81	0.53	0.60	0.78

Table 6.7: Comparison of evaluation results for Model 2: original, combined (old categories), and combined (all categories). Orange highlighting indicates better performance.

	Model 3: Original Dataset A			Model 3: Combined Dataset B1 (Old Categories)			Model 3: Combined Dataset B (All Categories)		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.41	0.39	0.62	0.42	0.38	0.61	0.39	0.37	0.61
ATT	0.63	0.78	0.88	0.61	0.70	0.84	0.57	0.56	0.75
BER	1.01	2.06	1.44	0.99	2.15	1.47	0.93	1.76	1.33
ENR	0.41	0.39	0.63	0.37	0.31	0.55	0.41	0.35	0.59
ETN	0.44	0.35	0.59	0.44	0.40	0.63	0.41	0.32	0.57
FAC	0.65	0.85	0.92	0.64	0.87	0.93	0.59	0.76	0.87
INS	0.68	0.76	0.87	0.78	0.89	0.94	0.62	0.63	0.79
MBW	0.58	0.55	0.74	0.59	0.56	0.75	0.56	0.50	0.71
STM	0.55	0.66	0.81	0.55	0.64	0.80	0.52	0.57	0.76

Table 6.8: Comparison of evaluation results for Model 3: original, combined (old categories), and combined (all categories). Orange highlighting indicates better performance.

shows that such data is particularly useful when it is category-coded, even if it does not include the target categories.

6.3.4 Summary

To conclude, note-level evaluation results clearly differ across categories (Table 6.9). Similar to sentence-level evaluation, some categories, namely *Walking (FAC)* and *Exercise tolerance functions (INS)*, performed better with the original model. Model 2 achieved the best results for *Attention functions (ATT)*, *Eating (ETN)*, and *Emotional functions (STM)*. For ETN, the results across Models 1 and 2 were very close, making

	Model 1: Original Dataset A			Model 2: Original Dataset A			Model 3: Original Dataset A		
ICF Category	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
ADM	0.40	0.36	0.60	0.42	0.40	0.63	0.41	0.39	0.62
ATT	1.02	1.05	1.22	0.61	0.65	0.81	0.63	0.78	0.88
BER	1.50	3.25	1.80	1.15	2.13	1.46	1.01	2.06	1.44
ENR	0.46	0.44	0.66	0.46	0.44	0.66	0.41	0.39	0.63
ETN	0.49	0.41	0.64	0.44	0.32	0.56	0.44	0.35	0.59
FAC	0.58	0.67	0.82	0.73	1.00	1.00	0.65	0.85	0.92
INS	0.58	0.56	0.75	0.90	1.24	1.11	0.68	0.76	0.87
MBW	0.62	0.58	0.76	0.63	0.63	0.79	0.58	0.55	0.74
STM	0.57	0.68	0.82	0.59	0.70	0.84	0.55	0.66	0.81

Table 6.9: Comparison of evaluation results for Models 1, 2 & 3 using the original dataset. Orange highlighting indicates better performance.

this difference not significant. For STM, the results across all three models were very similar. These findings suggest that ATT does not rely heavily on category-specific vocabulary and benefits from training on a larger, cross-category dataset.

The remaining three categories, *Work and employment (BER)*, *Energy levels (ENR)*, and *Weight maintenance functions (MBW)*, showed improvements with Model 3. Among them, only BER exhibited substantial gains. This indicates that Model 3 is generally more robust, as it either significantly improves performance or maintains results comparable to other models (with the exception of ATT, FAC, and INS).

Looking at each category individually, most categories showed stable results across different datasets and models: *Respiration functions (ADM)*, *Energy levels (ENR)*, *Eating (ETN)*, *Weight maintenance functions (MBW)*, and *Emotional functions (STM)*.

Attention functions (ATT) improved considerably when using Model 2, but the best results were obtained with Model 3 on Dataset B (the combined dataset with all categories). Here, MAE decreased from 1.02 to 0.57, MSE from 1.05 to 0.56, and RMSE from 1.22 to 0.75. These results indicate that ATT is less sensitive to category-specific attributes and benefits from larger datasets.

Similarly, *Work and employment (BER)* achieved its best performance with Model 2 on Dataset B. MAE dropped from 1.50 to 0.87, MSE from 3.25 to 1.67, and RMSE from 1.80 to 1.29. This suggests that BER, like ATT, benefits from larger datasets and is not strongly dependent on category-specific features.

Finally, *Walking (FAC)* and *Exercise tolerance functions (INS)* performed best with Model 1. While FAC benefited from the combined datasets, especially Dataset B, INS performed similarly across different models and datasets, with the exception of Model 3 on Dataset B1, where the model's performance drastically decreased. These results indicate that these categories are sensitive and require carefully and professionally annotated data, as they tend to achieve higher performance when trained separately.

Chapter 7

Error Analysis

This chapter presents an extensive error analysis, divided into two parts. The first part examines the functioning category *Walking (FAC)* and investigates how differences in its annotation scale (0–5 instead of 0–4 for most other categories) influenced the performance of the models. The second part analyzes the use of negations, intensifiers, and minimizers, focusing on how these linguistic elements affect the assignment of functioning levels to individual sentences across the different models.

7.1 Analysis 1: Functioning Level Scale Differences

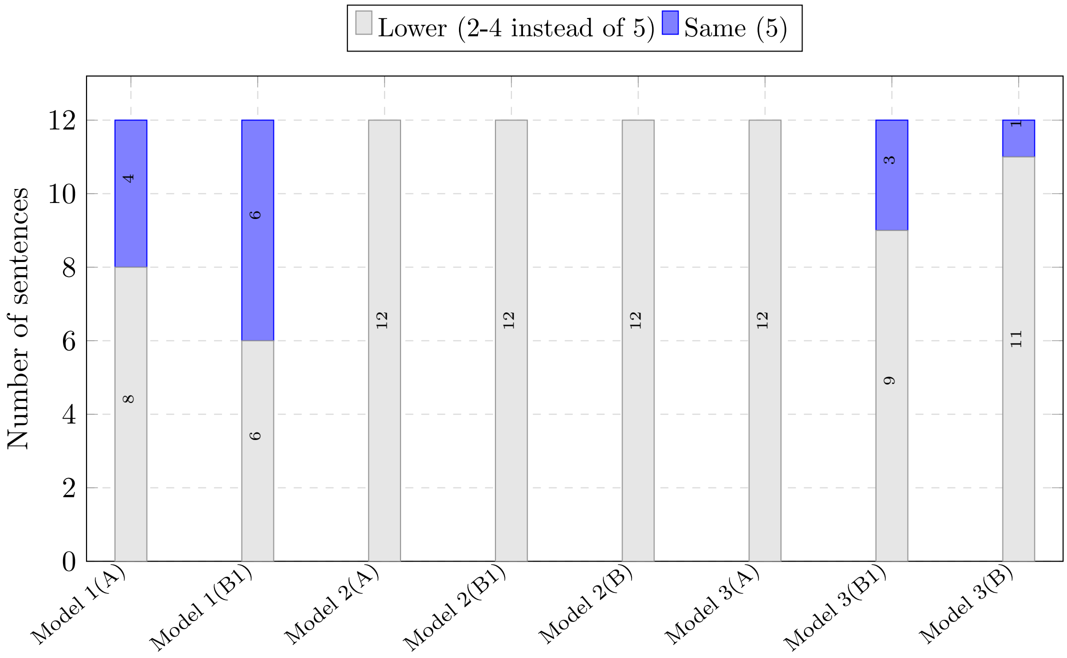

Figure 7.1: Model predictions for category *Walking (FAC)* with gold label 5.

Out of nine original and eight newly added functioning categories, the majority assign functioning levels on a 0–4 scale, where 0 indicates complete disability with respect to the category and 4 indicates no limitations. However, two of the seventeen categories, *Walking (FAC)* and *Exercise tolerance functions (INS)*, use a 0–5 scale. As

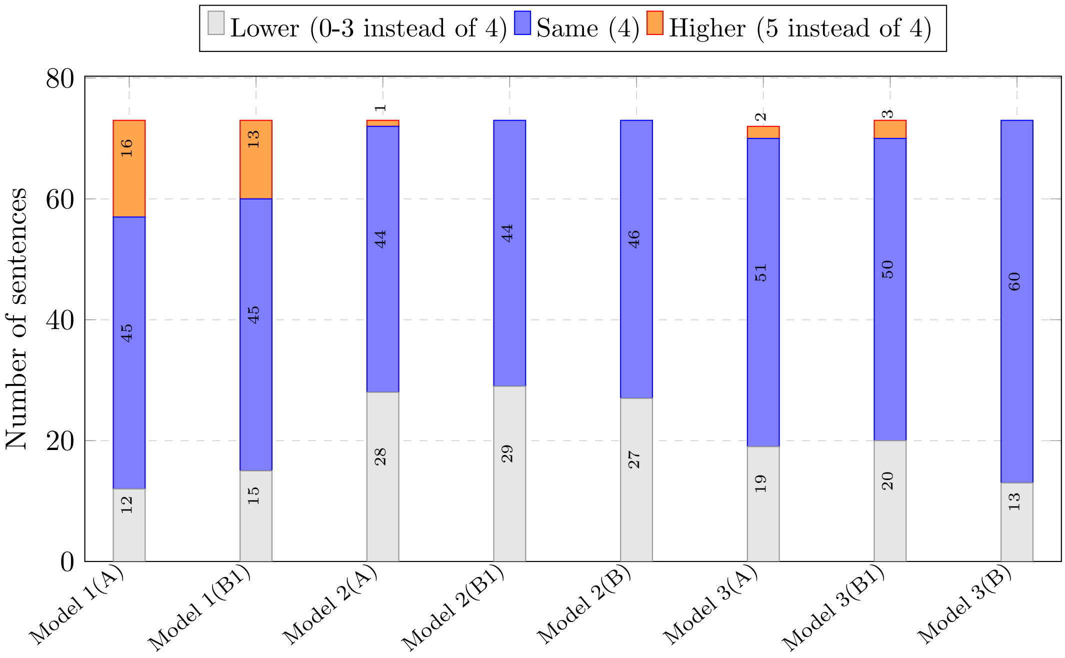

Figure 7.2: Model predictions for category *Walking (FAC)* with gold label 4.

in the other categories, 0 represents complete disability, but in these cases 5 represents full functioning.

Although these scale differences do not pose difficulties during manual annotation, because the annotators are well trained and provided with clear guidelines, they may introduce challenges for machine learning models. This is particularly relevant when a model is trained on combined datasets without explicit category encoding, as is the case for Model 2 in this thesis (Section 5.4).

The hypothesis that differences in functioning level scales negatively affect model performance is supported by both the sentence-level and note-level results. The two categories with the 0–5 scale (*Walking (FAC)* and *Exercise tolerance functions (INS)*) behaved markedly differently from the others. Both categories achieved their best performance when trained using Model 1, which consists of separate regression models for each category (Section 5.3), and showed improvement when trained on mixed-category data for Model 3, when the categories were combined and encoded (Section 5.5). As discussed in Section 6.2.4, this indicates that FAC and INS are highly category-specific and should not be trained using combined datasets or models if the categories are not explicitly encoded.

To examine this hypothesis more closely, I conducted a detailed analysis of the *Walking (FAC)* category. I manually compared the predictions of each model–dataset combination (eight experiments in total) with the gold labels in the test set, which contains 139 sentences. Particular attention was given to sentences with a gold label of 4 (73 instances) and those with a gold label of 5 (12 instances). The goal was to determine how the models behaved for these cases, and specifically whether they systematically mapped the unique FAC scale (0–5) onto the more common 0–4 scale used by other categories.

Model 2 (the model trained on all categories combined but without encoding category identity) consistently predicted 4 instead of 5 for all instances with the gold label 5 (Figure 7.1). This indicates that the model failed to recognise that FAC uses a different scale and instead defaulted to the scale shared by the majority of categories.

For sentences with the gold label 4, Model 2 predicted either the correct level or a lower level, with only one instance misclassified as 5 when trained on the original dataset (Figure 7.2). This further demonstrates that the combined-category model ignores category-specific scale differences. Moreover, when additional training data was added, the model relied even more heavily on the dominant patterns in the data, further diminishing its ability to account for the unique scale of FAC.

In conclusion, this analysis supports the hypothesis that the distinct functioning level scales used for *Walking (FAC)* and *Exercise tolerance functions (INS)* contribute to their divergent behaviour compared to the other categories. These categories do not benefit from exposure to mixed-category training data, as the models tend to override their specific scales in favour of the majority scale. This suggests that these categories may need to be trained separately. Alternatively, the scales could be reconsidered and potentially re-annotated by medical experts. For instance, two adjacent levels could be merged, or the scale could be adjusted from five to four points. Such decisions could be guided by the level distribution in the training and test data. For example, if level 0 occurs very rarely, all level 1 instances could be shifted down and merged with level 0. Implementing these adjustments could help align the scales with the overall dataset while preserving clinical interpretability, thereby improving model performance for these categories.

7.2 Analysis 2: The Impact of Negations, Intensifiers, and Minimizers

In this part, I analyse three linguistic elements - negations, intensifiers, and minimizers - and discuss their impact on the assignment of functioning levels across the three models: Model 1 (5.3), Model 2 (5.4), and Model 3 (5.5). To do so, I select a subset of training examples containing at least one of these linguistic elements and construct an expanded dataset in which each sentence is paired with a modified version that no longer contains the target element. In a few cases, it was necessary to add opposing elements or antonyms to ensure that the modified sentences remained grammatically and semantically correct. For example, a sentence containing a negation such as *Kan geen lange gesprekken voeren.* (translated: *Cannot have long conversations.*) was modified to *Kan lange gesprekken voeren.* (translated: *Can have long conversations.*). For intensifiers, a phrase like *veel, bezorgen per scooter* (translated: *many, delivery by scooter*) was altered to *soms, bezorgen per scooter* (translated: *sometimes, delivery by scooter*). For minimizers, a sentence such as *POB gaat wel over, hoewel patiënt er misselijk en wat benauwd blijft.* (translated: *Chest pain does improve, although the patient remains nauseous and somewhat short of breath.*) was changed to *POB gaat niet over, terwijl patiënt er misselijk en wat benauwd blijft.* (translated: *Chest pain does not improve, while the patient remains nauseous and somewhat short of breath.*).

After assigning functioning levels to the modified sentences, I re-train the three models on the remaining training data and evaluate them using test sets that include both the original and modified sentences. Subsequently, I perform a detailed analysis for each category, examining whether the functioning level changed between original and modified sentences and comparing model performance across sentence pairs and across models. This tests the sensitivity of the models for such overt linguistic markers.

Below, I describe each step in detail. The subsections outline the selection of sentences, the construction of the datasets, and the procedure for generating modified sentences and assigning them functioning levels. This is followed by a subsection describing the training and evaluation process. The section concludes with three subsections: the first discusses the effect of negation, the second examines intensifiers and minimizers, and the third provides a summary of both.

Importantly, given the very small size of the test sets, the results presented below should be interpreted as exploratory rather than definitive.

7.2.1 Filtering Data and Constructing Sentence Pairs

To begin with, my supervisor Piek Vossen, together with two members of CLTL, namely Isa Maks and Hennie van der Vliet, created a list of Dutch words belonging to several linguistic categories. These included negations (for example, *niet*, *geen*), minimizers (for example, *beetje*, *amper*), intensifiers (for example, *erg*, *enorm*), approximators (for example, *bijna*, *nagenoeg*), degree modifiers (for example, *enigszins*, *gemiddeld*), as well as adjectives used to express functioning levels (for example, *goed*, *slecht*). I then expanded this list by adding all possible inflections and removing duplicates.

Importantly, I chose to conduct the error analysis on a subset of the training dataset rather than on the development or test sets. The main reason for this choice was that both the development and test sets were very small for most categories and therefore insufficient for a meaningful analysis. For example, the *Attention function (ATT)* category contained only 32 sentences in the test set, *Work and employment (BER)* had 26,

ICF Category	Original Sentences	Retained Sentences
ADM	5,233	1,587
ATT	251	128
BER	216	102
ENR	1,005	565
ETN	2,491	936
FAC	1,086	411
INS	1,104	462
MBW	766	263
STM	1,420	740
Total	**13,572**	**5,194**

Table 7.1: Number of sentences per ICF category in Dataset A (original training data) before and after removing various linguistic elements.

and *Weight maintenance function (MBW)* had 60 sentences. After filtering the training dataset (retaining only those sentences containing at least one linguistic element from the Dutch word list introduced above) the size of the dataset decreased from 13,572 to 5,194 sentences (Table 7.1). This showed even more clearly that the smaller development and test sets would not contain enough examples of several linguistic categories, especially the less frequent ones. Therefore, I opted for the more elaborate approach: conducting the error analysis on the training dataset and later re-training the model on a reduced training set from which the selected examples were removed.

First, I filtered the training set using the Dutch word list from the previous paragraph, retaining only sentences containing at least one target element, which resulted in 5,154 sentences (Table 7.1). Focusing on negations, intensifiers, and minimizers, I created a dedicated word list for each and re-filtered the dataset separately for each category. Each time, the code selected only the relevant sentences and then randomly sampled a fixed number per functioning level category (12 for negations and 6 for each minimizers and intensifiers). As a result, the final document for negations contained 108 sentences, while the documents for minimizers and intensifiers contained 54 sentences each, corresponding to the nine functioning-level categories.

Finally, my supervisor, Piek Vossen, a native Dutch speaker, manually reviewed all sentences in the three documents and created versions in which the target linguistic element had been removed, occasionally replacing it with an opposite marker to maintain grammatical and semantic correctness.

7.2.2 New Sentences and Functioning Levels

Edwin Geleijn, a trained medical professional that co-developed the annotation guidelines, reviewed the new sentences from each pair and assigned functioning levels to them. Importantly, during this task, he did not have access to the original sentences or their functioning levels. In addition, while assigning functioning levels, the annotator discarded five sentences (one each from the ATT, ENR, and ETN categories, and two from the INS category) because they did not contain information relevant to the respective functioning categories.

7.2.3 Re-training the Models and Evaluation

ICF Category	**Original Dataset A**	**New Dataset A**
ADM	5,233	5209
ATT	251	229
BER	216	194
ENR	1,005	982
ETN	2,491	2467
FAC	1,086	1062
INS	1,104	1080
MBW	766	743
STM	1,420	1396
Total	**13,572**	**13,362**

Table 7.2: Number of sentences per ICF category in Dataset A (training data), before and after removing the sentences used as test data.

After assigning functioning levels to the new sentences, I prepared the datasets for re-training and evaluation. First, I removed from the original training dataset all examples that I had selected for the error analysis. This resulted in 24 sentences being removed from each category, with the exception of the ATT and BER categories, from which 22 sentences were removed, and the ENR and MBW categories, from which 23 sentences were removed (Table 7.2. These differences occurred because some sentences appeared multiple times across the documents, for example, occurring in both the intensifiers and negations sets.

Next, I split the three test datasets (negations, minimizers, and intensifiers) into nine documents each, corresponding to the nine functioning-level categories. I then re-trained the three models used in this study and evaluated them on each of the datasets containing the different linguistic elements.

7.2.4 Negations

In this subsection, I present the error analysis results for one particular linguistic element: negations. Tables 7.3–7.11 show the evaluation outcomes. Each row corresponds to a different sentence; the first three columns display model performance on the original sentences, and the last three columns show performance on the modified sentences. The middle columns indicate the gold labels - original levels for the original sentences and new levels for the modified ones.

The tables are also color-coded to facilitate interpretation: blue cells indicate that the model predicted the same level as the gold label; orange cells mark predictions higher than the gold label; and grey cells represent predictions lower than the gold label. While it is also important to consider whether functioning levels changed from original to modified sentences, the color coding helps visualize some patterns.

Respiration functions (ADM)

	Original Sentences			ADM Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	4	4	4	2	3	3	2
2	2	2	1	1.5	1	1	2	1
3	4	4	4	4	2	3	3	3
4	4	4	4	4	1	1	1	2
5	4	4	4	4	2	3	3	3
6	4	4	4	4	2	2	2	2
7	0	0	0	0	0	0	0	0
8	4	4	4	4	2	3	2	2
9	4	4	4	4	2	2	2	2
10	4	4	4	4	1	1	2	2
11	4	4	4	3	2	2	2	2
12	4	4	3	4	2	3	2	2

Table 7.3: Negations: Model predictions for the category *Respiration functions (ADM)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the ADM category, most sentence pairs showed a difference of around two functioning levels between the original and modified sentences - for example, sentence 1 changed from level 4 to level 2. However, one sentence (sentence 7) did not show any change between the original and modified versions. As visible in the table, most model predictions for the original sentences were correct. However, for the modified sentences, the models showed a tendency to predict higher functioning levels than the gold labels. For instance, in sentence pair 5, the original level was 4 and all three models predicted it correctly. The modified version had a new level of 2, and although all models lowered their predictions from 4 to 3, this adjustment was insufficient. This pattern suggests that the models do respond to negations, but the effect of negation is not always strong enough for the models to fully adjust their predictions.

Attention function (ATT)

	Original Sentences			ATT Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	4	4	1	1	3	3	4
2	3	2	2	2	3	3	2	3
3	3	2	2	2	3	3	2	2
4	3	2	2	1	4	3	2	2
5	3	2	2	2	4	3	4	4
6	3	2	2	2	4	3	3	3
7	3	1	1	3	1	3	1	1
8	3	4	4	4	2	3	3	3
9	3	2	2	2	4	3	4	4
10	3	2	2	0	4	3	3	4
11	3	2	2	3	4	3	4	4
12	3	4	4	4	NA	NA	NA	NA

Table 7.4: Negations: Model predictions for the category *Attention functions (ATT)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the ATT category, sentence 1 did not change its gold label in the new sentence. Moreover, sentence 12 was marked as not relevant by the annotator. For some sentences, such as 3 and 7, the model predictions remained the same for both the original and new sentences, despite differences in gold levels. In other cases, for example sentence 6, Models 2 and 3 predicted higher levels for the new sentence (3 instead of 2), but this increase was still insufficient to match the gold label of 4. Overall, Model 1 consistently predicted the same levels for both original and new sentences (level 3), whereas Models 2 and 3 performed better, likely due to the small size of the dataset (also discussed in Section 6.2.4). Consequently, this category benefited from the additional training data. These results suggest that the models do pay attention to negations, but sometimes do not assign sufficient importance to them.

Work and employment (BER)

For the BER category, sentences 1 and 6 did not change their functioning levels in the new sentences. It is worth noting Model 1, whose performance did not change at all between the original and new sentences. This is likely due to the very small training dataset as well as the general difficulty of predicting accurate levels for this category. Indeed, this category showed the worst performance across all models with respect to MAE, MSE, and RMSE, indicating that when the model misassigned a level, the magnitude of the error was higher than for other categories (see Section 6.2.4). While Models 2 and 3 performed slightly better, they tended to overpredict for original sentences and underpredict for new sentences, suggesting that the models attempted to remain conservative by avoiding extreme predictions. Consequently, although the models did pay attention to negations, this effect was insufficient and was overshadowed by the limitations imposed by the small training dataset.

	Original Sentences			BER Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	2	3	4	4	3	2	2
2	3	2	2	0	3	3	2	3
3	3	3	3	2	4	3	3	3
4	3	2	2	0	4	3	3	4
5	3	1	2	0	4	3	3	3
6	3	4	4	4	4	3	4	4
7	3	1	1	0	4	3	3	3
8	3	3	4	2	4	3	4	4
9	2	1	2	2	4	2	3	3
10	3	1	1	2	3	3	3	4
11	3	1	1	0	4	3	3	3
12	3	4	4	4	0	3	2	2

Table 7.5: Negations: Model predictions for the category *Work and employment (BER)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

Energy levels (ENR)

	Original Sentences			ENR Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	2	1	2	2	3	2	2	2
2	2	2	2	2	4	2	3	3
3	2	1	2	0	4	2	4	4
4	2	1	1	4	0	2	1	1
5	2	3	2	2	4	2	4	4
6	2	4	3	2	NA	NA	NA	NA
7	2	2	2	4	2	2	2	2
8	2	2	2	1	3	2	3	3
9	2	2	2	1	4	2	3	4
10	2	1	2	4	0	2	1	1
11	2	2	2	1	4	2	3	4
12	2	2	2	1	4	2	4	4

Table 7.6: Negations: Model predictions for the category *Energy levels (ENR)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the ENR category, all sentences changed their functioning levels in the new sentences, except for sentence 6, which the annotator marked as irrelevant. Once again, Model 1 predicted the same levels for both original and new sentences, indicating a conservative approach. Models 2 and 3 adjusted their predictions for the new sentences, but in most cases, the adjustments were insufficient. For example, in sentence pair 4, the original level was 0, but the models tended to overpredict toward the middle of the scale. The new level was 4, but only Model 3 assigned it accurately, while Models 1 and 2 remained more conservative with predictions of 3. These results suggest that Model 1 pays little attention to negations, whereas Models 2 and 3 are more sensitive but often still too conservative.

Eating (ETN)

	Original Sentences			**ETN Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	1	1	1	1	1	1	1	1
2	2	3	3	2	2	3	2	2
3	1	1	1	0	0	1	1	1
4	3	2	2	1	4	3	2	2
5	3	3	4	4	2	3	2	3
6	3	4	4	4	2	3	3	2
7	1	1	1	1	NA	NA	NA	NA
8	1	1	1	1	0	1	1	1
9	1	1	1	0	0	1	1	1
10	1	1	0	0.2857	0	1	0	0
11	3	2	2	2	3	3	2	3
12	1	1	1	0	0	1	1	1

Table 7.7: Negations: Model predictions for the category *Eating (ETN)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the ETN category, six sentence pairs did not change their functioning levels after adding the new sentences. Moreover, in sentence 7, the new sentence was marked as irrelevant by the annotator. Model 1 predicted the same levels for both original and new sentences, with one exception in sentence 2: for the original sentence, it assigned level 2, while the new sentence received level 3. However, this was one of the sentences whose level did not actually change. For this category, all three models remained conservative, avoiding levels 0 or 4, despite such gold labels appearing in the dataset. Importantly, even in cases of drastic level changes, such as sentence 4, where the level increased from 1 to 4, the models did not adjust their predictions. These findings indicate that the models pay very little attention to negations in this category. It is advisable to further investigate this phenomenon by analyzing a larger dataset and examining the context in which these negations appear.

Walking (FAC)

For the FAC category, three out of twelve sentences did not change their functioning levels. Moreover, for five additional sentences, the level changed by only 1, leaving four sentences with more substantial changes. Unlike most other categories, this category uses a 0–5 scale rather than 0–4. In these examples, only Models 1 and 3 paid attention to the scale, as Model 2 never predicted level 5, even though it appeared in the gold labels. This is likely because Model 2 considers all category data together and thus

	Original Sentences			**FAC Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	4	4	4	3	4	3	4
2	4	3	4	4	3	4	3	4
3	4	4	4	4	4	4	3	4
4	4	3	3	4	5	4	4	5
5	3	3	3	3	5	3	4	4
6	4	3	4	4	5	5	4	5
7	4	3	4	0	4	4	4	4
8	4	4	4	4	4	4	4	4
9	3	1	3	0	0	3	1	3
10	4	2	3	1	4	4	4	4
11	4	2	3	0	4	3	3	4
12	4	4	4	3	4	4	3	4

Table 7.8: Negations: Model predictions for the category *Walking (FAC)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

overlooks finer details specific to this category. Across this category, the models generally tended to predict levels toward the middle of the scale, such as 3, even when the gold label was much higher or lower (for example, sentence 11). These results suggest that negations do not play a significant role in this category.

Exercise tolerance functions (INS)

	Original Sentences			**INS Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	4	3	2	3	3	3	2
2	1	1	1	1	1	1	1	1
3	1	1	1	1	2	1	2	1
4	4	2	3	2	3	4	2	3
5	2	2	2	3	4	2	2	1
6	3	4	4	2	3	3	4	3
7	1	1	1	1	2	1	2	1
8	3	4	3	4	3	3	4	3
9	3	2	3	1	2	4	4	4
10	3	2	2	2	2	3	2	2
11	2	4	3	2	2	2	3	2
12	2	2	2	1	3	2	3	2

Table 7.9: Negations: Model predictions for the category *Exercise tolerance functions (INS)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the INS category, three sentence pairs did not change their functioning levels after adding the new sentences. The remaining nine sentence pairs changed their levels by only 1, indicating a lack of more substantial changes. As a result, Model 1 predicted the same levels for all original and new sentences, suggesting it could not detect subtle changes in level assignment. Most model predictions remained unchanged, and when adjustments occurred, they often went in the wrong direction. For example, in sentence 1, the models overpredicted levels 3, 4, and 3 when the gold level was 2, and then predicted 3, 3, and 2 when the gold level increased to 3. These findings indicate that the models pay little attention to negations in this category. Moreover, negations may have less impact here than in other categories, since the levels barely changed for the new sentences. It is therefore recommended to conduct a similar analysis for this category using a larger dataset. Additionally, it would be valuable to investigate the context in which negations appear, as this category may be linguistically specific in a way that reduces the importance of traditional linguistic cues such as negations.

Weight maintenance functions (MBW)

	Original Sentences			**MBW Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	4	4	4	2	3	2	2
2	3	3	3	3	2	3	3	3
3	3	4	4	4	2	3	2	3
4	4	4	4	4	3	4	3	3
5	1	2	2	1	4	1	2	3
6	4	4	4	4	3	4	4	4
7	3	3	3	4	2	3	2	2
8	3	4	4	4	2	3	2	2
9	4	4	4	4	2	4	3	4
10	3	3	3	3	2	3	2	2
11	3	3	3	2	2	3	2	2
12	3	3	4	4	2	2	2	2

Table 7.10: Negations: Model predictions for the category *Weight maintenance functions (MBW)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the MBW category, all gold labels changed after introducing the new sentences. Some sentences changed by only 1, while others experienced larger shifts. Model 1, although it adjusted some predictions for the new sentences compared to the original ones, tended to overpredict: most often, its predictions remained the same as the original level, and when it assigned a lower level, the adjustment was insufficient to match the actual new gold label. In contrast, Models 2 and 3 adjusted better to the new sentences. While some mispredictions occurred, they generally moved in the correct direction. These results indicate that negations are important for this category, and that the models, particularly Models 2 and 3, are more sensitive to them.

	Original Sentences			STM Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	2	2	2	2	2	2	2	2
2	3	4	4	4	2	2	2	2
3	2	1	1	3	4	3	3	4
4	3	3	4	4	3	2	2	2
5	3	2	2	3	4	4	4	4
6	2	2	2	2	4	3	4	4
7	4	4	4	4	2	3	2	3
8	2	2	2	2	2	2	1	2
9	4	4	4	4	2	3	3	3
10	4	4	4	4	2	4	4	4
11	2	2	2	3	3	2	3	3
12	2	1	2	2	4	3	2	3

Table 7.11: Negations: Model predictions for the category *Emotional functions (STM)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

Emotional functions (STM)

For the STM category, three sentence pairs did not change their levels for the new sentences, and three additional pairs changed by only 1. The models performed relatively well for this category. While some mispredictions occurred, primarily because the models were conservative and adjusted predictions by only 1, they generally moved in the correct direction. For example, in a sentence 7 where the gold level changed from 4 to 2, the models predicted 3, 2, and 3, demonstrating a shift toward lower levels, although not enough to match the new gold label exactly. These findings suggest that negations play an important role for this category, and the models are fairly sensitive to them. However, negations are not considered a sufficiently strong cue to cause more substantial changes in the predictions.

7.2.5 Intensifiers and Minimizers

In this subsection, I present the error analysis results for the remaining two linguistic elements: intensifiers and minimizers. The analysis is summarised in 18 tables (two per functioning category), one showing the results for intensifiers and one for minimizers (7.12–7.29). As with the negation results, each row corresponds to a different sentence. The first three columns display the model's predictions for the original sentences, and the last three columns show the predictions for the modified sentences. The middle columns contain the gold labels: the original levels for the original sentences and the updated levels for the modified versions.

The tables are colour-coded to facilitate interpretation. Blue indicates that the model predicted the same level as the gold label; orange marks predictions higher than the gold label; and grey highlights predictions lower than the gold label.

For each linguistic element, there were six sentence pairs (original and modified sentences) used to evaluate the models. Because similar patterns were expected for intensifiers and minimizers within each category, the tables are presented sequentially, and the discussion addresses them together.

Respiration functions (ADM)

	Original Sentences			**ADM Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	4	3	4	0	3	4	3
2	0	0	0	2	1	0	0	0
3	1	1	1	0	0	2	2	1
4	0	0	0	0	0	0	0	0
5	3	3	3	2	3	3	4	3
6	1	1	1	2	0	0	1	1

Table 7.12: Intensifiers: Model predictions for the category *Respiration functions (ADM)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			**ADM Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	3	3	4	3	3	3	3
2	2	2	2	2	2	2	2	2
3	1	0	1	0	0	1	0	1
4	4	4	4	3	3	3	3	2
5	2	3	3	1	1	1	2	2
6	1	2	2	1	1	1	1	1

Table 7.13: Minimizers: Model predictions for the category *Respiration functions (ADM)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For the ADM category, most sentence pairs with intensifiers remained stable: two pairs kept the same gold labels after adding the new sentences, and two changed by only one level. Model 3 behaved most conservatively - it produced identical predictions for both original and new sentences, meaning it did not react to the intensifiers at all. Model 2 changed two predictions: one time incorrectly (for a sentence whose level was not supposed to change) and once in the right direction (sentence 5), adjusting the level from 3 to 4. Although this adjustment was still too high compared to the gold label, it showed that the model was sensitive to the intensified expression. Model 1 reacted the most, changing the predictions for three sentences. One change occurred when the gold label stayed the same, but two changes (sentences 1 and 6) moved in the correct direction. Still, even correct adjustments were small, typically by only one level, making them insufficient to reach the intended gold label.

For minimizers, the ADM category showed a different pattern: only one of six sentence pairs had its gold label changed (sentence 1, from level 4 to 3). All models kept the same predictions for the first three sentences, but for the last three they consistently lowered predictions by one - apart from Model 1 on sentence 6, which was already correct for the original sentence. Interestingly, many of these model adjustments happened even when the gold labels did not change, suggesting that the models were particularly sensitive to minimizers and often interpreted them as relevant cues even when the annotator did not.

Overall, for ADM, both intensifiers and minimizers influenced the models to some extent - except for Model 3 with intensifiers. Minimizers especially seemed to trigger prediction changes, sometimes more strongly than warranted by the gold labels. This suggests that future work should examine the contextual role of minimizers more carefully, as the models appear to treat them as meaningful cues.

Attention function (ATT)

	Original Sentences			ATT Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	3	3	2	2	3	4	3
2	2	1	1	0	3	2	1	1
3	3	2	2	2	3	3	2	2
4	3	3	3	2	4	3	3	4
5	3	2	3	3	3	3	2	3
6	3	2	2	2	3	3	3	4

Table 7.14: Intensifiers: Model predictions for the category *Attention functions (ATT)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For ATT intensifiers, two sentence pairs (1 and 5) retained the same gold labels across original and new versions. Model 1 completely ignored the intensifiers, producing identical predictions for all six sentence pairs. Models 2 and 3 did react in a few cases, but their adjustments were limited. For instance, sentence 2 showed a substantial gold-label shift (from 0 to 3), but none of the models captured it. Model 2 increased its predictions for sentences 1 and 6: the adjustment in sentence 6 aligned with the

	Original Sentences			**ATT Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	4	4	1	1	3	3	4
2	3	2	2	3	2	3	2	2
3	3	3	3	3	2	3	2	2
4	3	2	2	4	4	3	4	3
5	3	2	2	2	2	3	2	2
6	3	3	3	3	1	3	2	2

Table 7.15: Minimizers: Model predictions for the category *Attention functions (ATT)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

gold labels, while sentence 1 should not have changed. Model 3 altered predictions for sentences 4 and 6; for sentence 4 the change allowed it to match the new gold label, but for sentence 6 it overcorrected by two levels. In general, Model 1 ignored intensifiers entirely, while Models 2 and 3 showed some sensitivity but not consistently enough to reach the correct label.

For minimizers, three pairs had unchanged gold labels. Model 1 again kept all its predictions unchanged. In contrast, Models 2 and 3 reacted several times, sometimes correctly and sometimes not. Model 2 made both correct adjustments (for example, sentence 2) and incorrect ones (for example, reducing the level for sentence 1 and raising it for sentence 4, although the gold labels stayed the same). Model 3 produced the most improvement: it correctly matched both original and new gold labels for three sentences and moved closer to the correct direction for sentence 6. It also adjusted sentence 4, even though the label did not change, but again toward the more appropriate direction.

In sum, for ATT, Model 1 consistently ignored intensifiers and minimizers, whereas Models 2 and 3 used them, but generally in a conservative manner. Their changes were often by just one level, limiting their ability to reflect larger gold-label shifts.

Work and employment (BER)

	Original Sentences			**BER Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	2	2	0	0	3	2	2
2	3	3	3	2	4	3	2	2
3	3	4	4	4	0	3	2	3
4	3	3	2	4	2	3	3	3
5	3	3	3	0	0	3	3	3
6	3	4	4	4	4	3	3	3

Table 7.16: Intensifiers: Model predictions for the category *Work and employment (BER)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			BER Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	4	4	4	2	3	3	3
2	3	2	2	1	4	3	3	3
3	3	4	4	4	4	3	3	3
4	3	2	3	4	3	3	3	4
5	3	2	3	4	4	3	2	2
6	3	3	3	4	0	3	2	2

Table 7.17: Minimizers: Model predictions for the category *Work and employment (BER)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

In the BER intensifier data, three gold-label pairs changed, while the rest remained stable. Model 1 made no adjustments between original and new sentences, showing the same insensitivity as in previous categories. Model 2 adjusted predictions three times: twice in the wrong direction (sentences 2 and 6, the latter without a gold-label change) and once correctly but insufficiently (sentence 3, where the gold labels changed dramatically from 4 to 0, but the model only shifted slightly). Model 3 made four changes, with three incorrect (sentences 2, 4, 6) and one correct but still too small (sentence 3). Thus, intensifiers in BER were challenging for the models: Model 1 ignored them, while Models 2 and 3 reacted inconsistently and often incorrectly.

For the BER minimizers, Model 1 again left all predictions unchanged. Model 2 made several adjustments, three in the correct direction (sentences 1, 2, 6) and two incorrect (sentences 3 and 4). All its adjustments were minor, usually a one-level shift, even when larger changes were required. Model 3 showed a very similar pattern: three correct adjustments (same sentences: 1, 2, 6) and three incorrect ones (sentences 3, 4, and 5, the last of which should not have changed at all). As with intensifiers, changes were generally small.

In summary, the BER category shows strong similarity across intensifiers and minimizers. Model 1 consistently ignored the linguistic cues, while Models 2 and 3 attempted to react but were often confused, making incorrect directional adjustments or shifting too little. This suggests that BER might require a closer look at specific expressions and their contexts. It may also be necessary to review the gold labels, since the annotation for this experiment was done by only one annotator and could contain inconsistencies.

Energy levels (ENR)

For intensifiers in the ENR category, one sentence pair did not change in level, and all three models predicted both the original and new labels correctly. For the remaining sentences, Model 1 adjusted predictions in the correct direction twice, while Models 2 and 3 adjusted three times each. Model 1 made conservative changes, typically shifting the level by only one, whereas Models 2 and 3 made larger adjustments - twice increasing the level by three - which allowed them to correctly predict sentences 4 and 6. Overall, all models responded to intensifiers, but Models 2 and 3 took larger, riskier steps, resulting in more accurate predictions.

	Original Sentences			**ENR Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	2	2	2	2	2	2	2	2
2	2	2	2	2	4	2	2	2
3	1	2	1	0	3	1	2	1
4	1	1	1	1	4	2	4	4
5	2	2	1	2	4	2	3	3
6	1	1	1	1	4	2	4	4

Table 7.18: Intensifiers: Model predictions for the category *Energy levels (ENR)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			**ENR Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	1	2	3	3	1	1	1	1
2	2	2	2	3	1	1	1	1
3	2	3	3	2	1	2	2	2
4	2	2	2	1	4	2	3	4
5	1	1	1	0	3	2	1	1
6	2	2	2	2	4	2	3	4

Table 7.19: Minimizers: Model predictions for the category *Energy levels (ENR)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

For minimizers in the ENR category, all new gold labels differed. Model 1 adjusted only once (sentence 5), but the change did not match either the original or new gold label. Model 2 adjusted five times, all in the correct direction, with two matching the new gold label. Model 3 also adjusted five times, all correctly, with four predictions matching the new gold label. This demonstrates that all three models considered minimizers, with Models 2 and 3 clearly outperforming Model 1.

In summary, for ENR, all models attended to intensifiers and minimizers. Model 1 remained conservative, while Models 2 and 3 made larger adjustments, improving accuracy. Model 3 performed best for minimizers.

Eating (ETN)

Regarding intensifiers in the ETN category, two sentence pairs showed no change. Model 1 adjusted its prediction only once (sentence 1), but the shift - from 4 to 3 - was insufficient compared to the gold labels (4 to 2). Model 2 made four adjustments, all in the correct direction, with one prediction (sentence 6) matching the new gold label. Model 3 adjusted three times, also correctly, with the same sentence (6) matching the gold label. Overall, Models 2 and 3 outperformed Model 1, though all remained relatively conservative.

For minimizers, two sentence pairs retained the same original and new gold labels.

	Original Sentences			**ETN Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	2	1	4	2	3	1	1
2	1	1	1	0	0	1	1	1
3	3	2	1	2	4	3	3	3
4	3	2	2	3	3	3	2	2
5	2	2	2	2	4	2	3	3
6	2	3	3	4	2	2	2	2

Table 7.20: Intensifiers: Model predictions for the category *Eating (ETN)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			**ETN Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	3	3	4	4	3	3	3
2	2	2	2	2	4	2	3	3
3	3	2	2	1	4	4	4	4
4	4	4	3	3	3	3	3	2
5	2	1	1	1	0	1	1	1
6	2	2	2	1	4	3	3	3

Table 7.21: Minimizers: Model predictions for the category *Eating (ETN)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

Model 1 made four adjustments, all in the correct direction, with sentences 3 and 4 matching the new labels. Model 2 also adjusted four times, correctly directed, again matching the gold labels for sentences 3 and 4. Model 3 made four adjustments, three correctly directed, matching the new label for sentence 3. These results show that all models considered minimizers, with Model 2 slightly outperforming the others, though no definitive conclusion can be drawn.

In summary, for ETN, all models responded to both intensifiers and minimizers but remained conservative, indicating that these linguistic cues sometimes provide limited information for fully accurate predictions.

Walking (FAC)

In the FAC category, three sentence pairs showed no change in gold labels for intensifiers. Model 1 adjusted only once (sentence 5), but overpredicted both original and new levels. Model 2 made five adjustments, four in the correct direction, with two predictions matching the gold labels; the remaining predictions were underpredicted. Model 3 adjusted three times, with predictions closer to the true values than the other models. Overall, all models responded to intensifiers, with Model 3 performing best.

Regarding minimizers, four sentence pairs retained the same gold labels, while changes in sentences 2 and 6 were minimal (one level). Model 1 made two adjust-

	Original Sentences			**FAC Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	2	3	0	3	4	2	4
2	5	3	3	5	5	5	4	5
3	4	3	4	4	3	4	2	4
4	4	3	4	4	4	4	4	4
5	4	3	3	3	4	5	4	4
6	4	4	4	4	4	4	3	4

Table 7.22: Intensifiers: Model predictions for the category *Walking (FAC)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			**FAC Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	2	4	2	2	4	2	4
2	4	2	4	2	4	5	4	4
3	4	4	5	4	4	4	4	5
4	3	3	4	4	4	3	2	4
5	5	4	5	5	5	4	3	4
6	4	3	4	4	5	4	4	4

Table 7.23: Minimizers: Model predictions for the category *Walking (FAC)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

ments: one correct (sentence 2) and one unnecessary (sentence 5). Model 2 made four adjustments, correctly matching sentences 2 and 6, with two unnecessary changes. Model 3 made a single unnecessary adjustment (sentence 5). These results indicate that all models considered minimizers, but adjustments were often incorrect or unnecessary, suggesting that these cues were less informative and pointing to potential annotation inconsistencies.

In summary, for FAC, Model 3 was superior in handling intensifiers, while all models performed similarly for minimizers. The findings also highlight possible issues with annotation quality, emphasizing the value of involving at least two annotators in future work.

Exercise tolerance functions (INS)

Within the INS category, two new sentences were marked as irrelevant by the annotator, leaving four sentences for evaluation. Among these, one sentence pair did not change in gold labels. Model 1 maintained its original predictions across all sentences. In contrast, Model 2 modified predictions for three out of four sentences, achieving one match with the gold label and adjusting two in the correct direction. Model 3 changed only once; nevertheless, it was the most accurate, matching two gold labels for the original sentences and three for the new sentences. Therefore, although it appeared less

	Original Sentences			**INS Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	2	2	2	1	NA	NA	NA	NA
2	2	2	2	2	NA	NA	NA	NA
3	1	2	1	1	1	1	1	1
4	3	2	3	1	2	3	3	4
5	3	2	2	2	2	3	2	2
6	3	3	3	2	3	3	4	3

Table 7.24: Intensifiers: Model predictions for the category *Exercise tolerance functions (INS)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			**INS Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	4	3	4	3	4	3	3
2	4	4	4	3	2	4	3	2
3	1	3	2	2	1	1	3	2
4	3	3	3	4	3	3	4	4
5	4	4	3	5	3	4	3	3
6	3	4	4	2	2	3	4	3

Table 7.25: Minimizers: Model predictions for the category *Exercise tolerance functions (INS)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

sensitive to intensifiers, Model 3 actually demonstrated greater precision by adjusting only when necessary.

Regarding minimizers in INS, just one sentence pair retained the same gold label. Model 1 kept its predictions unchanged, which allowed it to match two new gold labels. Model 2 made four adjustments, three correctly oriented, with one adjustment matching the gold label. Model 3 changed predictions three times, matching three gold labels but incorrectly adjusting once (sentence 4), the same sentence where Model 2 also erred. These results could point to potential annotation inconsistencies.

Taken together, drawing firm conclusions about intensifiers in INS is challenging due to the limited dataset and two sentences being excluded. The lack of adjustments from Model 1 suggests a lower sensitivity to linguistic cues, while Models 2 and 3 performed slightly better. However, as Section 6.2.4 indicates, Model 1 generally produced the most reliable results for this category.

Weight maintenance functions (MBW)

For intensifiers in MBW, one sentence pair (sentence 3) retained the same gold label. Model 1 modified its prediction once, resulting in a correct classification. Model 2 adjusted four times, producing three matches with the gold label (one adjustment went in the wrong direction) and one prediction closer to the gold label. Model 3 modified

	Original Sentences			MBW Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	2	2	2	2	4	3	4	4
2	3	2	3	4	3	3	3	3
3	2	2	2	3	3	3	3	3
4	3	2	2	3	4	3	3	4
5	2	2	2	1	3	2	2	2
6	2	1	1	1	3	2	1	1

Table 7.26: Intensifiers: Model predictions for the category *Weight maintenance functions (MBW)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			MBW Category		New Sentences		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	4	3	3	3	2	4	2	2
2	3	3	3	2.888889	2	3	3	2
3	3	3	3	3	2	4	3	4
4	1	2	2	0	1	2	2	2
5	3	2	3	3	2	3	3	3
6	3	2	2	2	4	3	4	4

Table 7.27: Minimizers: Model predictions for the category *Weight maintenance functions (MBW)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

predictions three times, leading to four matches while the other two levels remained slightly underpredicted. This suggests that all models considered intensifiers, with Model 3 performing best.

Turning to minimizers in MBW, gold labels changed for every sentence pair, though usually by only one level. Predictions were generally accurate for original sentences, particularly for Model 3, but accuracy decreased for the new sentences. Model 1 changed predictions three times, all incorrectly, dropping its matched gold labels from three to zero. Model 2 also adjusted three times, two in the wrong direction, causing overpredictions in four out of six sentences. Model 3 modified four predictions, one incorrectly, resulting in three matched gold labels for new sentences compared to five for the original. Sentence 3 was particularly problematic, as two models adjusted incorrectly, which may indicate annotation or contextual challenges.

Overall, in MBW, the models attended to linguistic cues but often remained conservative or adjusted incorrectly, limiting the accuracy of predictions.

Emotional functions (STM)

In STM, all sentences changed levels after adding new sentences. Model 1 adjusted four predictions, gaining two additional matched gold labels but losing one previously matched. Minimal adjustments were made for the remaining sentences, even when

	Original Sentences			**STM Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	2	2	2	2	3	3	1	2
2	1	2	2	0	3	3	2	4
3	4	4	4	4	3	4	4	4
4	2	2	2	1	4	3	4	4
5	2	2	2	1	3	3	3	3
6	2	1	2	1	4	2	2	2

Table 7.28: Intensifiers: Model predictions for the category *Emotional functions (STM)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

	Original Sentences			**STM Category**		**New Sentences**		
No.	M1	M2	M3	Original Level	New Level	M1	M2	M3
1	3	2	2	3	4	3	3	4
2	2	2	2	2	2	2	2	2
3	2	2	2	2	4	3	3	4
4	3	3	3	3	4	3	4	4
5	3	3	2	2	4	4	4	4
6	2	1	2	0.5	3	2	1	3

Table 7.29: Minimizers: Model predictions for the category *Emotional functions (STM)* for original and new sentences.
Legend: Blue (prediction = gold), Orange (prediction above gold), Grey (prediction below gold).

larger changes were necessary. Model 2 also modified four predictions, adding two matches but losing three previously correct ones. Model 3 adjusted three predictions, resulting in two accurate new matches but losing two previously correct predictions. Overall, most changes were in the correct direction, showing that the models accounted for intensifiers, though sometimes underestimating their effect.

Concerning minimizers in STM, one sentence pair (sentence 2) retained the same gold label. Predictions were generally conservative, clustered between levels 2 and 3, with models avoiding extremes. While many predictions matched gold labels for both original and new sentences, a tendency to overpredict original sentences and underpredict new ones was observed.

In summary, for both intensifiers and minimizers in STM, the models frequently overestimated original sentence levels and underestimated new sentence levels. This indicates that although the models considered linguistic cues, they did not fully capture their intended emphasis.

7.2.6 Summary

Across all three linguistic elements - negations, intensifiers, and minimizers - the results show that these elements do not always affect the gold label, which often remains unchanged. However, when the gold label does change, these cues do influence the models' predictions, though their overall impact remains limited. In many cases, the models adjusted their predictions in the expected direction when a cue was introduced, yet the shift was usually too small to reach the correct gold label. This conservative behaviour appeared consistently across categories: the models tended to shift their predictions by only one level, even when the gold label changed more substantially.

A clear trend throughout the analysis is that Model 1 was the least sensitive to linguistic cues, with some exceptions, such as its strong performance in the ADM category for intensifiers. It frequently produced identical predictions for the original and modified sentences, suggesting that it relied very little on the presence of negations, intensifiers, or minimizers. Models 2 and 3, by contrast, responded more reliably to these cues and showed greater flexibility. This pattern is consistent with the results in Section 6.2.4, where the two models also outperformed Model 1 more broadly. Even so, as mentioned in the previous paragraph, Models 2 and 3 also displayed conservative tendencies, favouring predictions near the middle of the scale and rarely selecting extreme values.

Several factors likely explain these behaviours. First, the size of the training datasets greatly influenced the models' responsiveness: smaller datasets led to more conservative predictions and weaker reactions to linguistic cues. Second, the distribution of functioning levels in the training data is highly imbalanced (see Section 4.3.3), with many more sentences labelled with intermediate levels such as 2 or 3. This imbalance likely encouraged all three models to favour these levels, even when the linguistic cues suggested otherwise.

While no single category behaved in a drastically different way, negations in the BER, ETN, FAC, and INS categories were more difficult to interpret. This aligns with known limitations of language models in handling negation and antonyms (Hosseini et al., 2021; Sergeeva et al., 2019). Moreover, BER, FAC, and INS were also problematic when analysing intensifiers and minimizers. For future work, it would be beneficial to revisit these categories with larger datasets and to examine more closely the contexts in which the cues appear. This may help clarify why functioning levels shifted more dramatically in some cases than in others, both in annotator ratings and in model predictions.

Overall, although the models did recognise negations, intensifiers, and minimizers, they did not always treat them as sufficiently important to justify larger adjustments in functioning levels. This suggests that the models rely more heavily on broader contextual patterns in the training data than on specific linguistic markers. A larger dataset, together with a more detailed investigation of contextual patterns, would likely provide clearer insight into how these linguistic cues shape model behaviour.

Chapter 8

Discussion

In this chapter, I present the answer to my research question, discuss the limitations encountered during this project, and provide an overview of potential directions for future work.

8.1 Research Question: Results and Interpretation

This thesis, in addition to optimizing the functioning level classifiers and analyzing the performance of three different models, aimed to answer the following research question:

> *Is the language used in medical notes to describe different functioning levels consistent enough to allow for the development of a single generalizable classification model across all functioning categories?*

The results (see Section 6.2.4) revealed that Model 3, trained on all categories together with category encodings, was the most robust model, as it either improved performance or maintained results comparable to the other models, with performance further improving when using a dataset that contained both previously seen and unseen categories. However, there were three exceptions. For the *Attention functions (ATT)* category, Model 2 clearly outperformed Models 1 and 3. This result may have been influenced by the extremely small training dataset for this category, consisting of only 251 sentences. Moreover, for *Walking (FAC)* and *Exercise tolerance functions (INS)*, Model 1 performed best. This likely arised from their different rating scale (0–5) compared to most other categories (0–4), which limits the benefit of additional data from other categories.

Moreover, insights from the two error analyses indicated similar trends across categories, such as models tending to be conservative when predicting functioning levels, but Models 2 and 3 performing better in this regard than Model 1. Categories such as *Work and employment (BER)*, *Eating (ETN)*, *Exercise tolerance functions (INS)*, and *Walking (FAC)*, although not performing drastically differently, were difficult to analyze, suggesting a need for further, more extensive investigation. This outcome was likely heavily influenced by the level imbalance in the training set.

Overall, the findings suggest that, although this conclusion is somewhat limited, **the language used in medical notes is generally consistent enough to support the development of a single, generalizable classification model across all functioning categories**.

Using Model 3 instead of Model 1 can also reduce computational resource requirements without sacrificing accuracy; however, further work is needed to ensure robust implementation across all categories.

8.2 Limitations

During this thesis, several limitations were observed that might have affected the quality of the study. First, the dataset was relatively small. The functioning level classifiers were trained on 13,572 sentences, resulting in a limited amount of training data. Moreover, the dataset was heavily imbalanced: while there were nine ICF categories, more than a third of all training sentences pertained to *Respiration functions (ADM)*. Consequently, some categories, such as *Attention functions (ATT)* and *Work and employment (BER)*, had significantly smaller datasets, with only 251 and 216 sentences, respectively.

In addition, the data was strongly imbalanced regarding functioning levels. Most categories contained many examples with levels in the middle of the scale, such as 2 or 3, while extreme levels (for example, 0 or 4, or 5 for FAC and INS) were underrepresented. These dataset issues made it more difficult to train the models and affected their performance. For example, in the second error analysis (Section 7.2), the results show that the models tended to be conservative, predicting levels near the middle of the scale even when the true functioning level was more extreme.

Another limitation was the quality of the OpenAI-generated dataset. The functioning levels were generated for new sentences, which were also generated by my classmate for some categories. This process introduced not only errors from my generation but also from hers, resulting in numerous erroneous sentences.

Furthermore, as noted by Kim et al. (2022), annotation quality was not always perfect, particularly for more challenging categories such as *Work and employment (BER)*, *Exercise tolerance functions (INS)*, and *Eating (ETN)*. Possible inconsistencies in functioning level annotations may have further affected model performance and the insights from the error analysis.

Additionally, the study focuses on only nine categories, which limits the possibility of making accurate generalizations.

Finally, conducting an error analysis without being a native Dutch speaker also posed a challenge. It was difficult to offer insights into the possible causes of the models' mispredictions without fully understanding the complex medical data. Therefore, the choice of which error analysis to conduct was constrained by my language abilities.

8.3 Future Work

In future work, it is recommended to address the limitations mentioned in the previous section. Increasing the dataset size, with a focus on achieving a more balanced distribution of functioning levels, would help ensure that model performance is not constrained by these limitations. This, in turn, would make it easier to investigate other factors that might explain why the models perform better for some categories and worse for others.

Another important step would be to introduce additional categories and annotate their functioning levels. Expanding the number of categories would make the study more widely applicable and relevant across different clinical settings.

After addressing the aforementioned issues, it would be useful to conduct an even broader and more detailed error analysis. With sufficient time, this could involve analyzing each category using a larger dataset. Rather than focusing only on the linguistic elements examined in this study, such as negations, minimizers, and intensifiers, future work could also investigate other relevant groups, such as level indicators, and then examine each linguistic cue individually, as their impact on model performance may differ.

Moreover, involving a native Dutch speaker in the error analysis - preferably a medical professional familiar with the A-PROOF project - could provide valuable insights into the contexts in which these linguistic elements appear, further improving the understanding of their effects on model predictions.

Chapter 9

Conclusion

This thesis builds on the existing A-PROOF project, which aims to automatically describe functioning levels for different ICF categories. The goal was to investigate whether the language used in medical notes to describe different functioning levels is consistent enough to allow the development of a single, generalizable classification model across multiple functioning categories. Replacing nine individual classifiers (one per category) with a single model would create a smaller, more efficient system that could be more easily applied in real-world settings.

To address this goal, three models were trained: the original model consisting of nine separate regression classifiers; a second model where all categories were trained together; and a third model, trained on all categories with category encodings added at the beginning of each sentence. In addition to differences in model design, these experiments involved different training datasets, which either expanded the original category-specific data or introduced previously unseen functioning categories. Furthermore, two detailed error analyses were conducted to better understand the models' performance and limitations.

The results indicate that a single generalizable model is indeed feasible. Model 3, trained on all categories with category encodings, proved to be the most robust, as it either maintained or improved performance compared to models trained individually or with simpler configurations. One exception was the *Attention functions (ATT)* category, where Model 2 outperformed the other models. This limitation may be due to the small size of the training dataset for this category and the imbalance of functioning levels, as the error analysis did not suggest that ATT differs substantially from the other categories. Moreover, two other exceptions were *Walking (FAC)* and *Exercise tolerance functions (INS)*, for which the original Model 1 yielded the lowest error rates.

Surprisingly, Model 2 also performed reasonably well overall despite the lack of category encodings, demonstrating the benefits of larger datasets for most categories. Moreover, most functioning categories responded well to the combined datasets, even when these included previously unseen categories. This effect was especially evident in Model 3, where all categories except *Energy levels (ENR)* achieved better performance when trained on the largest dataset containing both original and newly generated categories. Overall, these results highlight the usefulness of cross-category training, as well as augmenting the training set with both original and synthetic sentences.

The first error analysis indicated that Models 2 and 3 tended to overlook differences in rating scales across categories, likely because models trained on larger datasets give less attention to smaller details. The second error analysis showed that, overall, the models paid attention to the target linguistic elements, namely, negations, intensifiers, and minimizers, but all exhibited conservative behavior, tending to predict functioning levels near the middle of the scale, likely due to level imbalance. Nevertheless, Models 2 and 3 were more flexible and attentive to different linguistic cues compared to Model 1. Despite this conservatism, the models performed reasonably well across most categories, suggesting that the language used in medical notes is sufficiently consistent to support automated classification.

Several limitations may have affected model performance. The dataset was relatively small and imbalanced across both categories and functioning levels, which constrained the models' ability to learn from underrepresented categories and extreme levels. The quality of the OpenAI-generated sentences and annotations introduced additional noise. Furthermore, the study was limited to nine categories and relied on a non-native Dutch speaker for error analysis, which constrained the depth of linguistic insights in complex medical contexts.

In conclusion, this thesis demonstrates that automated functioning level classification across multiple ICF categories is achievable, and that the language in medical notes is generally consistent enough to support a generalizable model. While certain categories and limitations require further attention, the findings provide a foundation for developing a scalable, clinically relevant tool that can assist healthcare professionals in analyzing patient functioning efficiently and accurately. By improving datasets and annotations, and expanding the range of categories, future work can build on this foundation to develop robust and widely applicable solutions in medical NLP.

Appendix A

Data and Annotations

Figures A.1 to A.26 show the level distributions for each category. For the original categories, two graphs are presented: the left graph displays the level distributions in the original dataset (*Dataset A*), and the right graph displays the level distributions in the new categories (*Dataset B*). For the new categories, there is one graph per category.

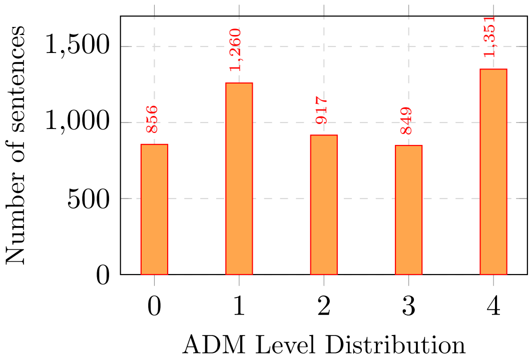

Figure A.1: Distribution of levels for ADM in the original train dataset (Dataset A).

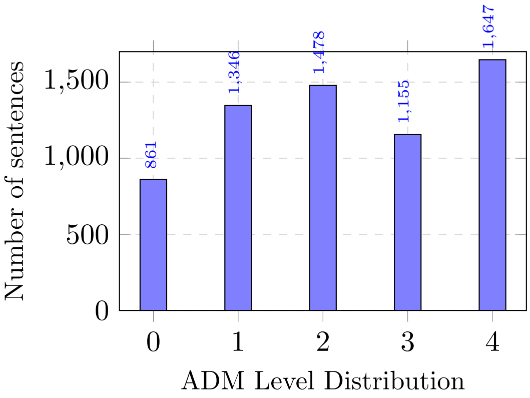

Figure A.2: Distribution of levels for ADM in the combined train dataset (Dataset B).

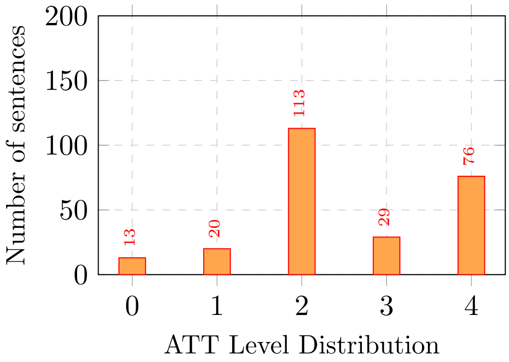

Figure A.3: Distribution of levels for ATT in the original train dataset (Dataset A).

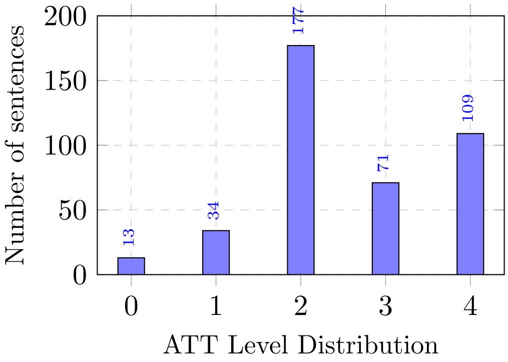

Figure A.4: Distribution of levels for ATT in the original train dataset (Dataset B).

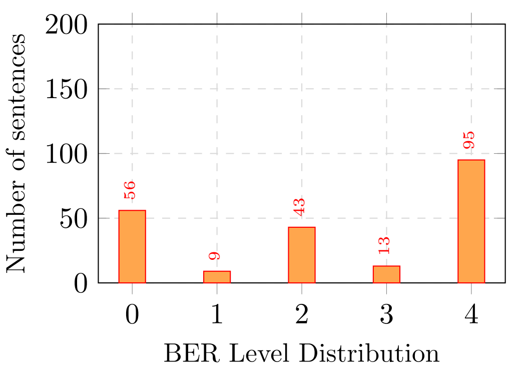

Figure A.5: Distribution of levels for BER in the original train dataset (Dataset A).

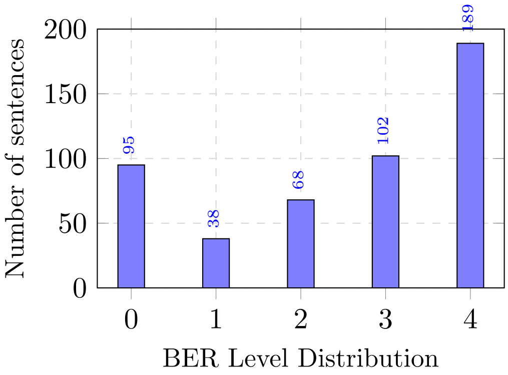

Figure A.6: Distribution of levels for BER in the combined train dataset (Dataset B).

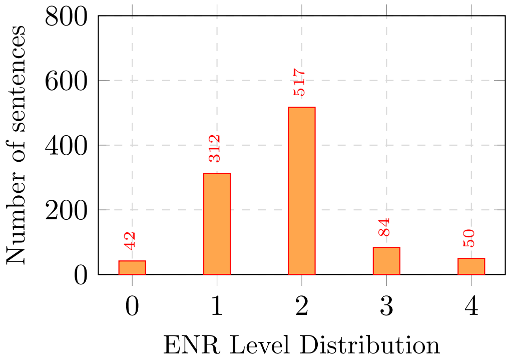

Figure A.7: Distribution of levels for ENR in the original train dataset (Dataset A).

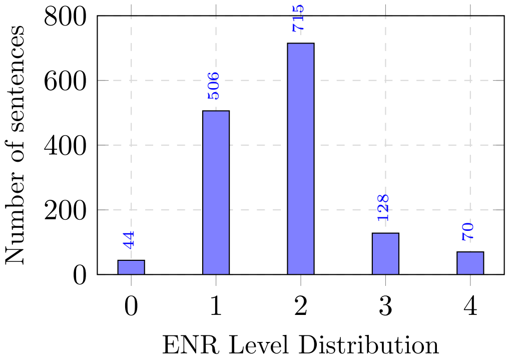

Figure A.8: Distribution of levels for ENR in the combined train dataset (Dataset B).

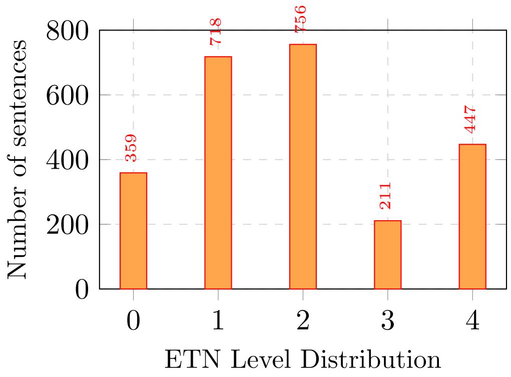

Figure A.9: Distribution of levels for ETN in the original train dataset (Dataset A).

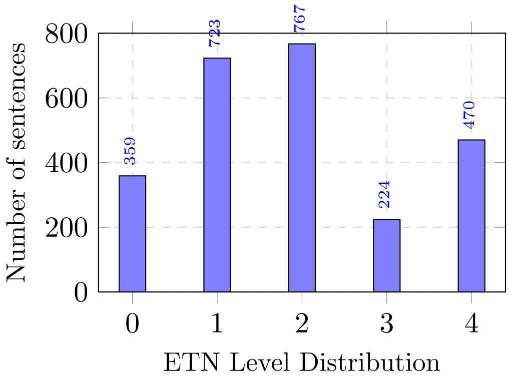

Figure A.10: Distribution of levels for ETN in the combined train dataset (Dataset B).

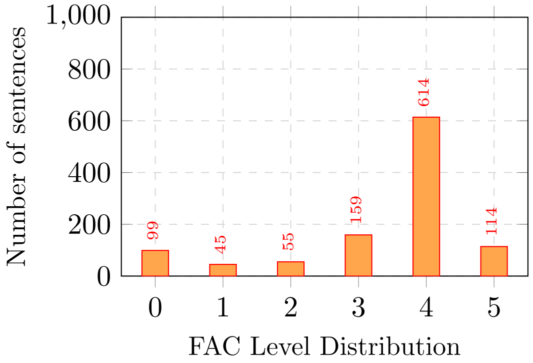

Figure A.11: Distribution of levels for FAC in the original train dataset (Dataset A).

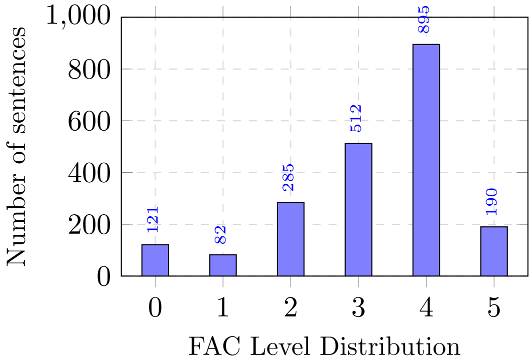

Figure A.12: Distribution of levels for FAC in the combined train dataset (Dataset B).

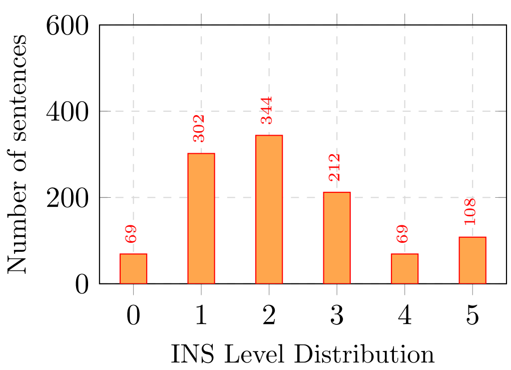

Figure A.13: Distribution of levels for INS in the original train dataset (Dataset A).

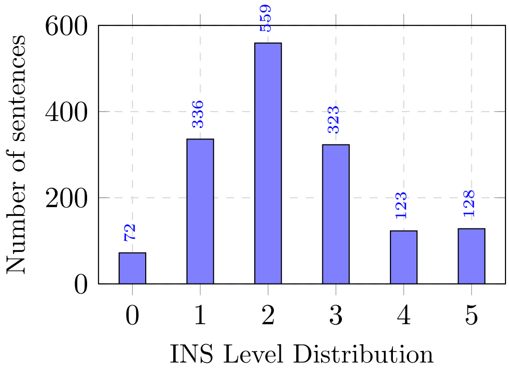

Figure A.14: Distribution of levels for INS in the combined train dataset (Dataset B).

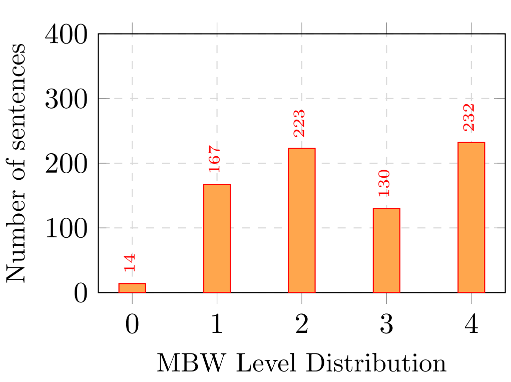

Figure A.15: Distribution of levels for MBW in the original train dataset (Dataset A).

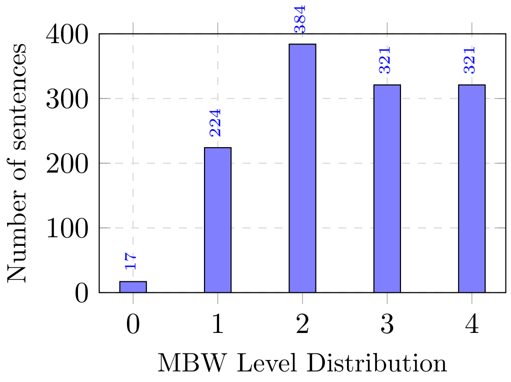

Figure A.16: Distribution of levels for MBW in the combined train dataset (Dataset B).

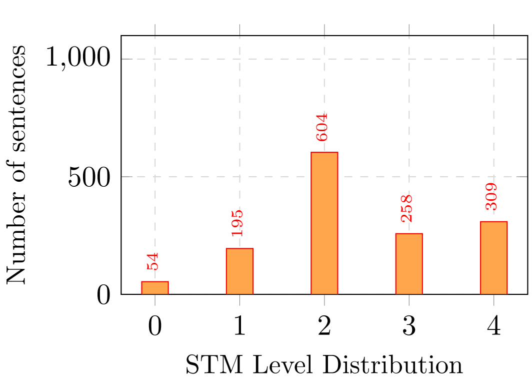

Figure A.17: Distribution of levels for STM in the original train dataset (Dataset A).

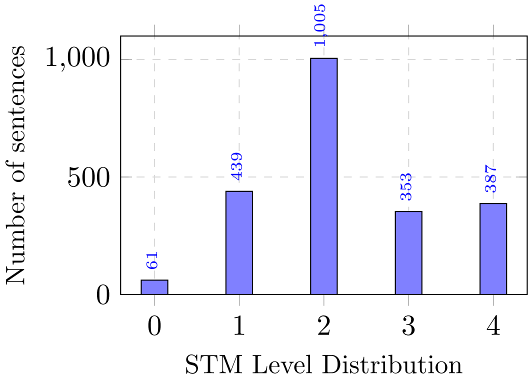

Figure A.18: Distribution of levels for STM in the combined train dataset (Dataset B).

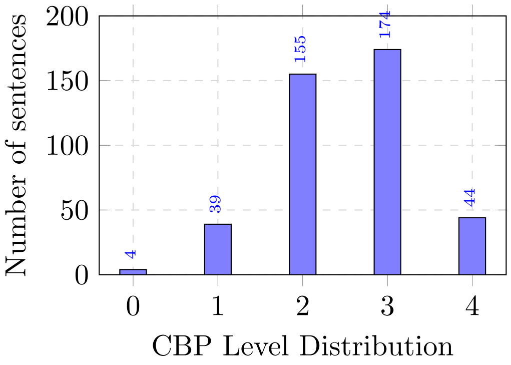

Figure A.19: Distribution of levels for CBP in the combined train dataset (Dataset B).

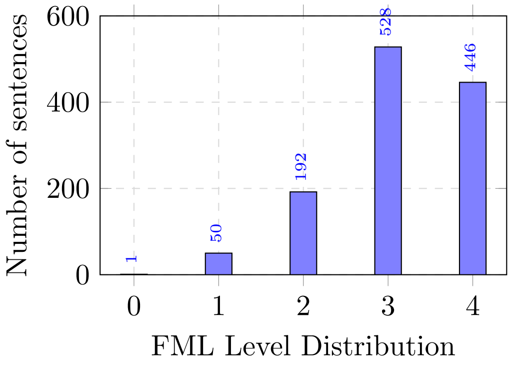

Figure A.20: Distribution of levels for FML in the combined train dataset (Dataset B).

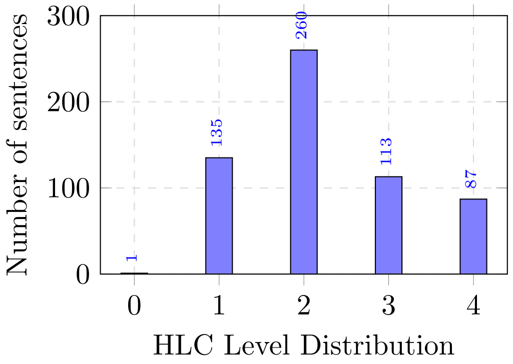

Figure A.21: Distribution of levels for HLC in the combined train dataset (Dataset B).

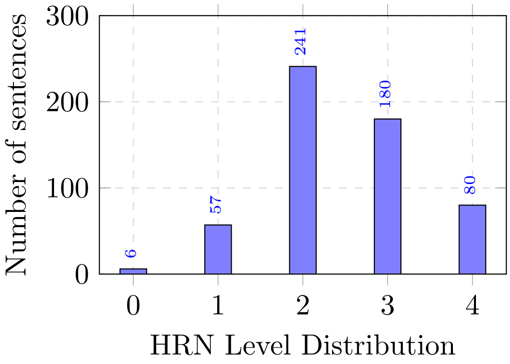

Figure A.22: Distribution of levels for HRN in the combined train dataset (Dataset B).

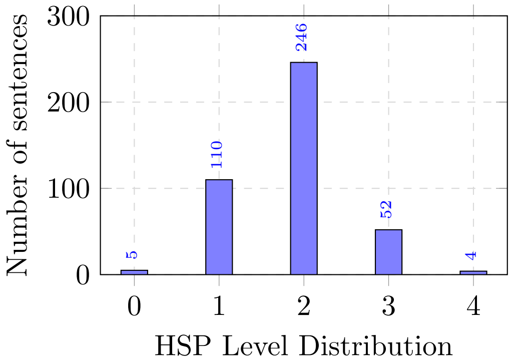

Figure A.23: Distribution of levels for HSP in the combined train dataset (Dataset B).

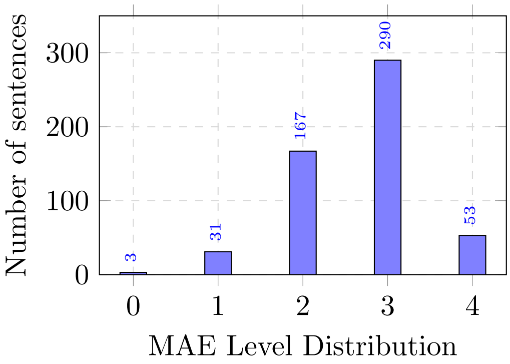

Figure A.24: Distribution of levels for MAE in the combined train dataset (Dataset B).

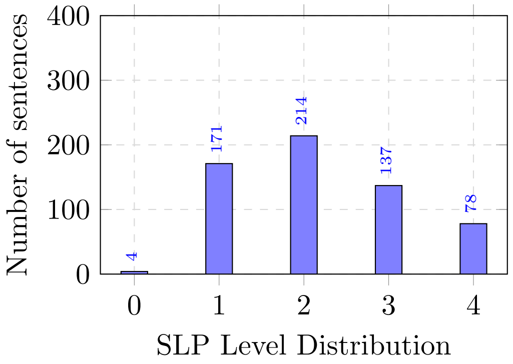

Figure A.25: Distribution of levels for SLP in the combined train dataset (Dataset B).

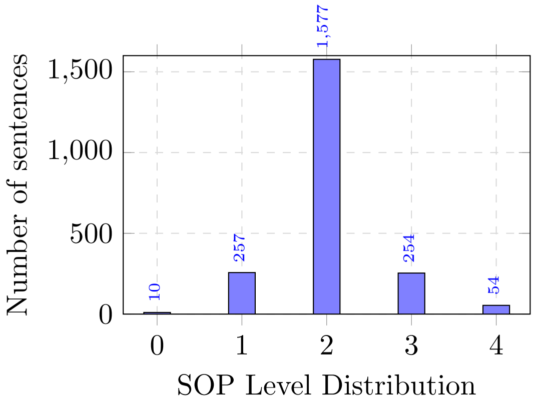

Figure A.26: Distribution of levels for SOP in the combined train dataset (Dataset B).

References

A-PROOF: Automated Prediction of post-COVID-19 RecOvery Of Functioning, 2022. URL https://cltl.github.io/a-proof-project/about. Last accessed on 13th November, 2025.

W. Ansar, S. Goswami, and A. Chakrabarti. A Survey on Transformers in NLP with Focus on Efficiency, 2024. URL https://arxiv.org/abs/2406.16893. Last accessed on 10th December, 2025.

A. Bagheri, A. Giachanou, P. Mosteiro, and S. Verberne. *Natural Language Processing and Text Mining (Turning Unstructured Data into Structured)*, pages 69–93. January 2023. URL https://www.researchgate.net/publication/375328167_Natural_Language_Processing_and_Text_Mining_Turning_Unstructured_Data_into_Structured. Last accessed on 10th December, 2025.

I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1371/. Last accessed on 10th December, 2025.

Cancer Research Institute. AI and Cancer: The Emerging Revolution, 2025. URL https://www.cancerresearch.org/blog/ai-cancer. Last accessed on 9th September, 2025.

Centers for Medicare & Medicaid Services. Electronic Health Records, 2024. URL https://www.cms.gov/priorities/key-initiatives/e-health/records. Last accessed on 9th September, 2025.

I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: The muppets straight out of law school, 2020. URL https://arxiv.org/abs/2010.02559. Last accessed on 10th December, 2025.

S. Chen. From 9 to 17 Categories: Weakly Supervised Sentence-Level ICF Classification in Dutch Rehabilitation Notes with GPT-4 Labeling and MedRoBERTa Fine-Tuning. Master's thesis, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 2025. Last accessed on 3rd November, 2025.

A. K. Conduah, S. Ofoe, and D. Siaw-Marfo. Data privacy in healthcare: Global challenges and solutions. *Digital Health*, 11, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12138216/. Last accessed on 10th December, 2025.

K. De Angeli, S. Gao, I. Danciu, E. B. Durbin, X.-C. Wu, A. Stroup, J. Doherty, S. Schwartz, C. Wiggins, M. Damesyn, L. Coyle, L. Penberthy, G. D. Tourassi, and H.-J. Yoon. Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types. *Journal of Biomedical Informatics*, 125, 2022. URL `https://www.sciencedirect.com/science/article/pii/S1532046421002860`. Last accessed on 10th December, 2025.

W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim. BERTje: A Dutch BERT Model. 2019. URL `https://arxiv.org/abs/1912.09582`. Last accessed on 10th December, 2025.

P. Delobelle, T. Winters, and B. Berendt. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265. Association for Computational Linguistics, November 2020. URL `https://aclanthology.org/2020.findings-emnlp.292/`. Last accessed on 10th December, 2025.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171–4186, 2019. URL `https://aclanthology.org/N19-1423.pdf`. Last accessed on 10th December, 2025.

A. Dubey and A. Tiwari. Artificial intelligence and remote patient monitoring in us healthcare market: a literature review. *Journal of Market Access & Health Policy*, 11(1):2205618, May 2023. URL `https://pubmed.ncbi.nlm.nih.gov/37151736/`. Last accessed on 13th November, 2025.

D. Elliott. Future of Jobs Report 2025: These are the fastest growing and declining jobs, 2025. URL `https://www.weforum.org/stories/2025/01/future-of-jobs-report-2025-the-fastest-growing-and-declining-jobs/`. Last accessed on 9th September, 2025.

Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020. URL `https://arxiv.org/abs/2007.15779`. Last accessed on 10th December, 2025.

A. Hosseini, S. Reddy, D. Bahdanau, R. D. Hjelm, A. Sordoni, and A. Courville. Understanding by Understanding Not: Modeling Negation in Language Models. 2021. URL `https://arxiv.org/abs/2105.03519`. Last accessed on 11th December, 2025.

K. Huang, J. Altosaar, and R. Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission, 2019. URL `https://arxiv.org/abs/1904.05342`. Last accessed on 10th December, 2025.

T. A. James. AI in Health Care: Potential and Progress, 2023. URL `https://learn.hms.harvard.edu/insights/all-insights/how-artificial-intelligence-disrupting-medicine-and-what-it-means-physicians`. Last accessed on 9th September, 2025.

D. Khurana, A. Koli, K. Khatter, and S. Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3): 3713–3744, 2022. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC9281254/`. Last accessed on 10th December, 2025.

J. Kim, S. Verkijk, E. Geleijn, M. van der Leeden, C. Meskers, C. Meskers, S. van der Veen, P. Vossen, and G. Widdershoven. Modeling Dutch Medical Texts for Detecting Functional Categories and Levels of COVID-19 Patients. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 4577–4585, Marseille, France, 2022. European Language Resources Association (ELRA). URL `https://aclanthology.org/2022.lrec-1.488.pdf`. Last accessed on 26th October, 2025.

E. Laparra, S. Bethard, and T. A. Miller. Rethinking domain adaptation for machine learning over clinical language. *JAMIA Open*, 3(2):146–150, April 2020. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC7382626/`. Last accessed on 10th December, 2025.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pretrained biomedical language representation model for biomedical text mining. 2019. URL `https://arxiv.org/abs/1901.08746`. Last accessed on 10th December, 2025.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL `https://arxiv.org/abs/1907.11692`. Last accessed on 10th December, 2025.

Z. Liu et al. Model Balancing Helps Low-data Training and Fine-tuning, 2024. URL `https://arxiv.org/abs/2410.12178`. Last accessed on 10th December, 2025.

C. G. M. Meskers, S. van der Veen, J. Kim, C. J. W. Meskers, Q. T. S. Smit, S. Verkijk, E. Geleijn, G. A. M. Widdershoven, P. T. J. M. Vossen, and M. van der Leeden. Automated recognition of functioning, activity and participation in COVID-19 from electronic patient records by natural language processing: a proof-of-concept. *Annals of Medicine*, 54(1):235–243, 2022. URL `https://www.tandfonline.com/doi/epdf/10.1080/07853890.2021.2025418?needAccess=true`. Last accessed on 10th December, 2025.

S. Moon, B. McInnes, and G. B. Melton. Challenges and Practical Approaches with Word Sense Disambiguation of Acronyms and Abbreviations in the Clinical Domain. *Healthcare Informatics Research*, 21(1):35–42, 2015. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC4330198/`. Last accessed on 10th December, 2025.

J. Nelson, M. Augustine, and S. Matthew. Natural Language Processing (NLP) in Clinical Documentation. February 2025. URL `https://www.researchgate.net/publication/391950535_Natural_Language_Processing_NLP_in_Clinical_Documentation`. Last accessed on 10th December, 2025.

R. R. Saxena. Applications of Natural Language Processing in the Domain of Mental Health. October 2024. URL `https://doi.org/10.36227/techrxiv.173014748.80471770/v1`. Last accessed on 10th December, 2025.

E. Sergeeva, H. Zhu, A. Tahmasebi, and P. Szolovits. Neural Token Representations and Negation and Speculation Scope Detection in Biomedical and General Domain Text. In E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, and F. Rinaldi, editors, *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 178–187. Association for Computational Linguistics, November 2019. URL `https://aclanthology.org/D19-6221/`. Last accessed on 11th December, 2025.

S. Verkijk and P. Vossen. Creating, anonymizing and evaluating the first medical language model pre-trained on Dutch Electronic Health Records: MedRoBERTa.nl. *Artificial Intelligence in Medicine*, 167, 2025. URL `https://www.sciencedirect.com/science/article/pii/S0933365725000831`. Last accessed on 10th December, 2025.

World Health Organization. International classification of functioning, disability and health (icf), 2025. URL `https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health`. Last accessed on 9th September, 2025.

S. Wu and H. Liu. Semantic Characteristics of NLP-Extracted Concepts in Clinical Notes vs. Biomedical Literature. In *AMIA Annual Symposium Proceedings*, pages 1550–1558, 2011. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC3243230/`. Last accessed on 10th December, 2025.

X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. 2022. URL `https://arxiv.org/abs/2203.03540`. Last accessed on 10th December, 2025.