Master Thesis

From 9 to 17 Categories: Weakly Supervised Sentence-Level ICF Classification in Dutch Rehabilitation Notes with GPT-4 Labeling and MedRoBERTa Fine-Tuning

Shutao Chen

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics (Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Piek Th.J.M. Vossen
2nd reader: Pia Sommerauer

Submitted: August 28, 2025

Abstract

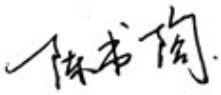
This thesis presents a sentence-level natural language processing (NLP) system for automatically extracting ICF (International Classification of Functioning, Disability and Health) codes from Dutch rehabilitation notes. Building on an existing 10-category classifier, we expand it to cover 18 ICF categories, including newly introduced contextual and psychosocial codes. To overcome limited labeled data, we engage GPT-4 for weak supervision, generating training labels to augment human annotated examples. We fine-tune a Dutch clinical language model (MedRoBERTa.nl) on various combinations of expert labels and labels generated by GPT-4.

The system's performance is evaluated at both the sentence level and the note level, and we compare its predictions with GPT-4's own classification outputs. Results show that the fine-tuned model outperforms GPT-4, achieving higher overall accuracy and significantly improving recall on categories that were previously challenging to identify. We also provide a detailed analysis of error patterns, category-specific performance gains, and the effects of training data augmentation. Our findings demonstrate an effective approach to broaden ICF code classification in clinical text, signifying the promise of combining expert knowledge with state-of-the-art language models for rehabilitation documentation analysis.

Declaration of Authorship

I, Shutao Chen, declare that this thesis, titled *From 9 to 17 Categories: Weakly Supervised Sentence-Level ICF Classification in Dutch Rehabilitation Notes with GPT-4 Labeling and MedRoBERTa Fine-Tuning*, 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 at this University.
- Where any part of this thesis has previously been submitted for a degree or qualification at this 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 joint work, I have made clear what was done by others and what I have contributed myself.

Date: 27 August 2025 Signed: 

Acknowledgments

I am deeply grateful to my supervisor, Piek Vossen, whose mentorship accompanied every stage of my experiment. Thank you for arranging extra meetings whenever I was stuck, answering my questions with patience, and offering constructive, actionable advice. Your support, both academically and emotionally, helped me keep momentum and clarity throughout the project.

I also thank Edwin Geleijn, the supervisor at Amsterdam UMC for continuous support whenever I ran into hurdles, whether technical (GPU access, virtual machines, permissions) or coordination and communication, your help kept the experiment moving when it mattered.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Health professionals increasingly recognize the importance of documenting patient functioning along with diagnoses. The World Health Organization's International Classification of Functioning, Disability and Health (ICF) provides a standardized framework to describe how a person's body, activities, participation, and environment interact in daily life. Indeed, scholars argue that functioning should stand with morbidity and mortality as a core health indicator, underscoring the need for routine ICF documentation in clinical practice. However, coding free text medical notes into ICF categories is challenging because much of the relevant information, such as daily activities or psychosocial context, appears only in unstructured narratives. Manual ICF coding is labor intensive and hard to scale, motivating the use of natural language processing (NLP) to automate this process. Early studies showed that transformer-based NLP models can map clinical text to broad ICF domains with high accuracy, but they also emphasized the substantial annotation effort required to cover finer grained categories.

This thesis addresses the challenge of expanding an ICF text classifier to cover a more comprehensive set of categories using minimal manual labeling. In an existing rehabilitation project (Meskers et al., 2022), a sentence-level classifier was originally trained to detect 10 categories of patient functioning (nine ICF codes plus a "None" category for no finding). These categories spanned several body functions and activities (e.g. energy level, attention, walking), but important details about daily life were missing, such as sleep, cognitive executive function, mobility aids, stress handling, sensory deficiency, and family or social context. Leaving out such environmental factors can make care and research less impartial, as key personal factors (for example, whether family support is strong) remain hidden in narrative notes (Newman-Griffis et al., 2022). To bridge this gap, with inspiration from a subset of ICF directed at community-dwelling older patients in primary care (Rink, 2023), we introduce eight additional ICF categories covering these underrepresented aspects (including **B280 Sensations of pain, B134 Sleep functions, B164 Higher-level cognitive functions, B230 Hearing functions, D240 Handling stress, D410 Changing basic body position, D465 Moving around using equipment, and D760 Family relationships**). By incorporating these, the classifier aims to

identify a fuller picture of patient health and context from text.

Expanding the label set presents a data bottleneck, there were no expert-labeled examples for the new categories in the original training set. To overcome this, we adopt a large language model (LLM) assisted labeling strategy. Recent transformer LLMs (such as GPT-4o) can understand clinical text and assign plausible labels with little hand crafted rule design. In this work, we use GPT-4o to perform weak supervision for the new ICF categories, generating synthetic annotations on a large collection of Dutch rehabilitation notes. A carefully designed prompt with definitions and examples for each category was used in a few-shot manner, providing an automated annotation of thousands of sentences. These GPT derived labels were then combined with the original manually labeled data to create an augmented training set that includes all 18 categories (the original 9 + 8 new categories, plus "None"). We fine-tuned a Dutch medical language model (MedRoBERTa.nl) on this blended dataset to build the extended classifier. The underlying hypothesis was that the synthetic labeling would introduce enough varied examples of the new classes to train an accurate multi-label classifier, without requiring a large manually labeled corpus.

The contributions of this thesis are as follows. First, we use a practical method to extend a clinical text classifier to cover additional ICF concepts using minimal expert effort, by employing a powerful LLM as an annotator. Second, we show that incorporating LLM-labeled synthetic data can increase performance on new categories without sacrificing accuracy on original categories. Third, we provide an empirical comparison between the fine-tuned model and the GPT-4o classifier itself, along with an error analysis for the fine-tuned model. Overall, this work contributes to a more comprehensive and scalable ICF classification pipeline for Dutch clinical text. It illustrates how modern NLP techniques can enrich electronic health records with structured functioning information, potentially improving holistic patient assessment.

Chapter 2

Background and Related Work

2.1 The ICF Framework and Functional Status in Text

The World Health Organization's International Classification of Functioning, Disability and Health provides a standardized taxonomy for describing health and disability in terms of functioning (WHO, 2001). It encompasses domains of body functions and structures, activities, participation, and contextual factors. In recent years, clinicians and researchers have emphasized that a patient's functioning should be tracked alongside morbidity and mortality as a key health indicator, supporting the case for routine ICF-based documentation (Płaszewski & Płaszewski, 2025; Tan et al., 2025). Much of this information appears only in unstructured clinical notes rather than structured fields, making it difficult to exploit at scale. Manual coding of narrative notes into ICF categories is labor-intensive, which motivates the use of NLP to automate ICF extraction.

2.2 Early and Neural NLP Approaches

Initial efforts to identify functioning information in text used rule-based methods and simple machine learning, demonstrating that mapping clinical text to ICF codes is feasible but not easily scalable. The advent of deep learning, especially transformer models, greatly improved accuracy. For example, a BERT-based classifier by Newman-Griffis et al. (2021) could assign broad ICF domains (e.g. mobility, cognition) to U.S. disability benefit narratives with near-clinician accuracy. In Dutch rehabilitation notes, transformer models similarly achieved F1 scores around 0.70 for categories like Walking and Emotional functions, confirming that text-encoded functional status can be learned reliably (Meskers et al., 2022). As these successes accumulated, researchers turned their attention to practical challenges such as data privacy, cross-site generalization, and limited training data. For instance, Fu et al. (2024) introduced

FedFSA, a federated learning approach that trains a multi-institutional ICF classifier without sharing patient data, resulting in performance comparable to single hospital models.

2.3 Exploiting Unlabeled Data and Domain Adaptation

Subsequent studies showed that the use of unlabeled clinical texts or the adaptation to specific domains can further improve ICF classification. Nieminen et al. (2025) fine-tuned a self-supervised model on just 151 rehabilitation documents and still achieved macro F1 above 0.80 for sentence-level ICF labels, suggesting that even small specialty corpora can be highly informative (Newman-Griffis & Fosler-Lussier, 2021). Functional status information has been extracted in other clinical domains as well, for example, transformer-based systems have identified heart failure NYHA classes from cardiology notes with high AUROC (Adejumo et al., 2024). A recent systematic review of 37 systems reported a median macro F1 of 0.77 for automated detection of activities of daily living, despite diversity in targets and datasets (Wieland-Jorna et al., 2024). These findings underscore that functional health data are both clinically valuable and are largely text-bound in settings.

2.4 Persistent Challenges - Imbalance and Privacy

Despite improved models, two key challenges persist in ICF text classification. First, class imbalance and limited coverage of certain categories hinder model learning. Some important functioning aspects appear very infrequently (or not at all) in available training data, causing models to struggle with underrepresented classes (Newman-Griffis et al., 2021). Second, data fragmentation and privacy restrictions limit data sharing between hospitals. Methods like federated learning have emerged to address the latter by training joint models without pooling raw data. FedFSA combines a rule-based extractor for ADL mentions with a federated BERT classifier for impairment status; across four institutions, federated training outperformed non-federated baselines for impaired ADL extraction and classification, although performance varied by ADL category (Fu et al., 2024). Complementary sequence labeling work in Mobility has also shown that ensembles over CRF/RNN/BERT components can push exact match NER F1 toward the mid-80s on physical therapy notes, reinforcing that functioning information is consistently learnable from clinical narratives (Thieu et al., 2021).

Meanwhile, addressing label sparsity has inspired data-efficient learning strategies. Active learning, for example, has contributed to nearly 70% reduction in annotation effort for rare

clinical labels, while maintaining moderate performance (e.g. F1 0.56 at sentence level) by prioritizing informative samples (Weissenbacher et al., 2024). Annotator-centric active learning further improved representation of minority classes by optimally selecting expert annotators (van der Meer et al., 2024). In parallel, weak supervision approaches have shown that carefully designed automatic labeling pipelines or pre-trained model annotators can rival fully supervised baselines until large gold standard datasets become available. Zhang et al. (2025), for instance, demonstrated that programmatic labeling functions can approach manual labeling performance on tasks with high amount of categories using only a fraction of manually labeled data. A recent systematic review of functional-status extraction underscores these trends: deep-learning systems (often transformers) now dominate, but heterogeneity in targets, datasets, and metrics complicates pooled performance comparisons and system selection (Wieland-Jorna et al., 2024). Taken together, the literature indicates a clear trend that modern transformer NLP enables accurate ICF coding, and a combination of privacy-aware training and data-efficient annotation can alleviate the data bottlenecks that limit broader deployment.

2.5 Expanding ICF Coverage - Gaps and New Categories

One notable gap in previous ICF extraction efforts is the management of contextual factors and other underaddressed aspects of functioning. Clinical notes often contain important details about a patient's daily life, such as family support or use of assistive devices, which are not identified in structured records. Newman-Griffis et al. (2022) warn that leaving out these "environmental helpers and hurdles" from documentation can lead to less equitable care and research. To bridge this gap, our work extends the label scheme to include eight additional ICF categories covering such underrepresented aspects. In clinical practice, it is common to focus on a subset of ICF codes most relevant to a given setting.

The WHO's ICF Core Sets provide examples of concise category selections developed via expert consensus for specific contexts (e.g. stroke rehabilitation). The WHO ICF Research Branch promote Core Sets—shortlists of ICF categories selected through literature review, multi-round expert consensus, and empirical validation (e.g. the methodology used to develop Core Sets for low back pain and stroke) (Cieza et al., 2004; Geyh et al., 2004). Widely used generic selections include the ICF Generic-7 and Generic-30, and a pragmatic ICF Rehabilitation Set supports routine reporting across services (ICF Research Branch, n.d.; Prodinger et al., 2018). Notably, these rehabilitation selections already cover three of the nine domains used in the original A-PROOF pipeline: b1300 Energy level (under b130 Energy and drive functions), b152 Emotional functions, and d450 Walking; supporting their continued inclusion

(ICF Research Branch, n.d.; Prodinger et al., 2018).

Guided by these core sets and related literature, we identified eight domains that are consistently highlighted as important but were missing from the original pipeline. For example, Pain (b280) is a core item in many musculoskeletal rehabilitation sets and strongly influences functional outcomes (Cieza et al., 2004; Hernández-Lázaro et al., 2023). Sleep functions (b134) appear in dedicated core sets for sleep disorders and affect fatigue, mood, and daily performance (Stucki et al., 2008). Family relationships (d760) also emerge as significant in rehabilitation research, as family support impacts discharge planning and recovery (Stallinga et al., 2021; Chung et al., 2014; Hakbijl et al., 2023). Other added domains include Higher-level cognitive functions (b164), which addresses executive function deficits common in neurological cases, nursing instruments such as NOSCA target executive functions because of their impact on independence and safety (Geyh et al., 2004; Persoon et al., 2011); Hearing functions (b230), relevant in geriatric and general rehab settings (Grill et al., 2007; ICF Rehabilitation Set, n.d.); Handling stress (d240), which pertains to psychological coping; Changing basic body position (d410), a frequent limitation mentioned in stroke or low back pain studies (Geyh et al., 2004; Cieza et al., 2004); and Moving around using equipment (d465), covering device-assisted mobility essential for rehab pathways (Geyh et al., 2004). Each of these has documented clinical importance in ICF based assessments or core sets, lending support to their inclusion. By incorporating these eight categories, in addition to the original nine, we assemble a more comprehensive label inventory of 17 ICF categories (plus a general "None" for irrelevant content). This extended set aims to capture a fuller picture of patient functioning from text, including the environmental and personal factors often overlooked.

2.6 Dutch Clinical NLP and MedRoBERTa

Dutch clinical-text processing has accelerated in recent years, motivated by privacy constraints around Electronic Health Records and the availability of de-identified Dutch corpora. Two resources are particularly relevant: the Erasmus MC Dutch Clinical Corpus (DCC) used for context/negation studies on GP letters, specialist letters, radiology reports and discharge notes (van Es et al., 2023), and the Dutch ADE corpus of ICU clinical notes annotated for adverse drug events (Murphy et al., 2025). Early Dutch systems typically fine-tuned general-purpose encoders such as BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020), which improved over multilingual baselines but were not pre-trained on the characteristic register of Dutch EHR notes.

Verkijk and Vossen introduced MedRoBERTa.nl as the first Transformer-based language model

trained specifically on Dutch hospital notes. Pre-training from scratch on around 13 GB of EHR text yielded a model that outperformed general Dutch encoders on an intrinsic medical similarity task and, after fine-tuning, on classifying sentences about patients' mobility levels; notably, the model already surpassed general baselines after only 10k pre-training steps (Verkijk & Vossen, 2021). A subsequent report details how the model was prepared for public release via a dedicated anonymization procedure and provides broader intrinsic and extrinsic evaluations, enabling compliant reuse in Dutch healthcare research (Verkijk & Vossen, 2025). Muizelaar et al. further pre-trained MedRoBERTa.nl on 148k HagaZiekenhuis notes ("MedRoBERTa.nl-HAGA"). The resulting model achieved strong performance on lifestyle status extraction, including F1 0.93 for smoking and 0.77 for drug use; it outperformed other Dutch BERT/RoBERTa baselines in the same setting (Muizelaar et al., 2024).

2.7 Use LLMs for Weak Supervision

Expanding the label set naturally worsens the data scarcity problem since initially, no expert-labeled examples were available for the new categories. Modern transformer-based LLMs (e.g. GPT-4o) possess extensive medical and general knowledge and can perform classification tasks with minimal prompting. A recent systematic review of LLMs on real EHR tasks reports that most studies rely on zero or few-shot prompting, with mixed gains from prompt engineering, and that many production settings prefer fine-tuning smaller open models rather than serving proprietary LLMs directly (Du, 2024). For sentence-level classification specifically, zero-shot LLMs can perform strongly. In heart-failure symptom classification over synthetic note snippets, ChatGPT-4 showed near-perfect performance, with sensitivity to prompt design and temperature (Workman et al., 2024).

Several studies show that LLM-derived labels can successfully train a smaller student model. Hsu et al. (2025) prompt Llama-2 to create weak labels and then train a BERT classifier that, with only a small gold set, outperforms purely supervised baselines while avoiding the high runtime cost of serving an LLM. Frameworks now operationalise LLM assistance under clinician oversight. EHRmonize evaluates multiple LLMs for medication abstraction and reports 60–68% reductions in annotation time using GPT-4o with few-shot prompts, with outputs reviewed and corrected by clinicians (Matos et al., 2024). CLEAR retrieves relevant sentences for a target variable and feeds them to an LLM; this improves extraction F1 while using fewer tokens, and downstream fine-tuned BERT models trained on these curated labels can surpass the trainer LLM on some variables (Lopez et al., 2025). In burn rehabilitation, ChatGPT-4o showed moderate agreement with experts at the code level and almost perfect agreement on item perspective during ICF linking, supporting the use of LLMs as assistive tools rather than

replacements (Gül et al., 2025) These studies demonstrate practical pipelines that pair LLM assistance with compact models.

In this project, we employ GPT-4o to generate weak labels for the new ICF categories, essentially using it as an automated annotator for thousands of narrative sentences. By providing definitions and example phrases for each category in a prompt (a few-shot setup), the model can assign likely ICF codes to unlabeled sentences. Previous work has shown that adding such synthetic labels to a small gold standard dataset can significantly improve classifier performance. For instance, Guo et al. (2024) found that augmenting limited clinical training data with GPT-generated annotations boosted downstream accuracy, though caution is needed to manage noise in the LLM's output. In line with these findings, we generate a blended training set containing both original human annotations and new GPT-4o assisted labels. The expectation is that the GPT augmented data will provide the necessary coverage of rare classes to train an effective classifier, while the inclusion of original expert-labeled data anchors the model in clinically validated examples.

We fine-tune a pre-trained Dutch medical language model (MedRoBERTa.nl) on this combined dataset to build the extended multi-label classifier. This approach follows the general trend in the literature: using powerful pre-trained models and synthetic supervision to overcome data limitations. By doing so, our aim is to improve detection of the newly introduced categories without sacrificing performance on the originally well represented categories, effectively testing whether the diversity introduced by LLM-generated data can provide broad gains in a resource-efficient manner.

Given the sentence-level nature of ICF category assignment, the need to run locally with Dutch clinical data, and evidence that in-domain encoders outperform prompt-only LLMs on extraction, I adopt MedRoBERTa.nl as the primary classifier and use a large model only for weak supervision to expand training coverage. This division reflects the best assignment, namely, domain-adapted encoders for high output, privacy-sensitive extraction; LLMs for occasional knowledge-intensive tasks or label augmentation (Bosma et al., 2025; Chen et al., 2025; Menezes et al., 2024; Singhal et al., 2025; Bolton et al., 2024; Timilsina et al., 2025). Overall, the present work builds on established NLP techniques for clinical text and the ICF framework, while putting minimal manual effort for maximal label expansion.

Chapter 3

Problem Definition & Objectives

3.1 Clinical and Technical Gap

The A-PROOF project showed that it is feasible to extract ICF-based information from Dutch clinical notes and to assign note-level functioning signals across nine categories, including 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), Emotional functions (STM). However, even with carefully designed annotation protocols, some categories suffered from scarce training data and low inter-annotator agreement, which translated into weak model recall, especially for Attention (ATT), Respiration (ADM), and Work & employment (BER) at the sentence level. On the other hand, several categories (e.g. Exercise tolerance (INS), Walking (FAC)) achieved stronger note-level F1 when sentence-level outputs were aggregated. These findings underline both the value of note aggregation and the limits imposed by annotation scarcity and label inconsistency in this setting. (Kim et al., 2022)

Clinically, the original nine categories are important but incomplete for a broader rehabilitation picture. ICF Core Set research consistently emphasizes additional functional categories such as pain (b280), sleep (b134), and psychosocial domains such as handling stress (d240), family relationships (d760), and mobility subcomponents like changing basic body position (d410) and moving around using equipment (d465), as significant factors of functioning and participation across conditions. This motivates extending the label space beyond the original A-PROOF scope. (Cieza, 2004; Grill, 2007)

3.2 Problem Statement

I aim to expand ICF category detection in Dutch EHR notes from the nine A-PROOF categories to seventeen clinically motivated categories:

Original 9: B1300 Energy level, B140 Attention functions, B152 Emotional functions, B440 Respiration functions, B455 Exercise tolerance functions, B530 Weight maintenance functions, D450 Walking, D550 Eating, D840–D859 Work & employment.

Added 8: B280 Sensations of pain, B134 Sleep functions, D760 Family relationships, B164 Higher-level cognitive functions, D465 Moving around using equipment, D410 Changing basic body position, B230 Hearing functions, D240 Handling stress and other psychological demands.

In practice, model training must also address a "None" (no-label) class, so experiments compare 18 labels (17 ICF + None). The core problem is to reliably detect these categories at note level under limited human gold labels by using LLM-assisted annotation to augment training data, while ensuring no degradation on the original nine categories.

3.3 Constraints from Previous Work and Available Resources

A-PROOF demonstrated a sentence-to-note pipeline, reported category distribution, and documented annotation challenges and evaluation choices (sentence vs. note level), which I adopt to ensure comparability. Performance limitations in ATT, BER, INS were linked to few examples and low agreement, suggesting that expanding label coverage requires strategies that add volume without sacrificing label quality. (Kim et al., 2022; Kumichev, 2024)

I combined earlier manually labeled notes (10 classes including None) with 2023 AMC and 2023 VUMC notes selected via expert-provided keywords, segmented with spaCy, and labeled additional categories using GPT-4o with few-shot prompts and category definitions. A small subset validated by clinicians yielded 61.5% correctness for the chosen prompting setup, which I then used to annotate the larger corpus for model training.

3.4 Research Questions

Does expanding from 9 to 17 ICF categories (plus None) materially increase clinically relevant coverage in Dutch rehabilitation notes? (Justification from ICF Core Set literature.) (Cieza, 2004)

Can LLM-assisted annotation (few-shot with definitions) provide useful supervision to train a Dutch clinical encoder (MedRoBERTa.nl) without harming the performance on the original nine categories?

How do a fine-tuned MedRoBERTa.nl model's predictions compare with GPT-4o direct predictions at note level across the expanded label set?

Does the expanded model reduce confusion with None and improve recall on historically hard categories (ATT/BER/INS), once additional weakly supervised data are introduced? (A-PROOF reported frequent misclassification into "no label".) (Kim et al., 2022)

3.5 Objectives

1. Extend the label space to the 17 categories listed above and implement them for sentence-to-note inference, following A-PROOF's pipeline for comparability. (Kim et al., 2022)
2. Design an LLM-assisted annotation protocol (prompting, exemples, definitions), validate a small gold subset with clinicians, and quantify LLM label quality (already piloted at 61.5% correctness).
3. Train and evaluate MedRoBERTa.nl models on manual-only data as well as manual + LLM-augmented data, reporting macro averaged precision/recall/F1 at sentence level and note level for the original 9 categories and all 17 categories (+ None). (Verkijk & Vossen, 2025)
4. Evaluate and compare to GPT-4o's direct predictions under matched label definitions and few-shot examples.
5. Analyze errors and results, with special attention to ATT/BER/INS and to None confusion previously observed in A-PROOF. (Kim et al., 2022)

3.6 Success Criteria

Primary: Both sentence-level and note-level macro F1 on the original nine is maintained or improved when training with LLM-assisted labels, relative to manual-only baselines.

Secondary: Reasonable sentence-level and note-level macro F1 across the eight new categories after clinician validation of a test subset; GPT-4o baseline is matched or exceeded by the fine-tuned MedRoBERTa.nl on the full 18-label setup.

Chapter 4

Data & Annotations

4.1 Train Sets

Train Sets Composition

The training dataset was compiled from multiple sources and evolved over the course of the project. Initially, we began with a manually annotated corpus covering the original set of 10 labels (9 informative categories plus a "None" category for irrelevant content). This included two legacy datasets (referred to as m123 and jenia_train). In total, this base corpus comprised 233,227 sentences drawn from clinical notes, all labeled by domain experts for the 9 initial ICF categories. The ICF provides a standardized taxonomy for health and functioning concepts, and the chosen categories (e.g. energy level, respiration, walking) correspond to specific ICF codes (e.g. B1300, B440, D450).

To expand the training data and cover new categories, we incorporated additional data sources and GPT-4o assisted annotations. First, a new set of 1,500 notes (40,414 sentences) from the Amsterdam UMC (AMC 2023) was selected (using keywords provided by clinicians to enrich relevant content) and added to training. These notes were initially unlabeled, so we employed GPT-4o to annotate them for all 18 target categories (the 9 original plus 8 new + None). The combined train data so far is used for stage 1 training with 273,641 sentences (9,615 notes) in total.

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	1630	0.6	1131	11.76
B140 Attention functions	577	0.21	446	4.64
B152 Emotional functions	4407	1.61	2465	25.64
B440 Respiration functions	6441	2.35	3033	31.54
B455 Exercise tolerance functions	1618	0.59	1197	12.45
B530 Weight maintenance functions	1461	0.53	1002	10.42
D450 Walking	3757	1.37	2257	23.47
D550 Eating	3067	1.12	1927	20.04
D840-D859 Work and employment	917	0.34	699	7.27
B280 Sensations of pain	9374	3.43	3357	34.91
B134 Sleep functions	3790	1.39	2198	22.86
D760 Family relationships	6643	2.43	2780	28.91
B164 Higher-level cognitive functions	5030	1.84	1945	20.23
D465 Moving around using equipment	3353	1.23	1937	20.15
D410 Changing basic body position	5542	2.03	2586	26.9
B230 Hearing functions	1709	0.62	877	9.12
D240 Handling stress and other psychological demands	5726	2.09	2569	26.72
None	218646	79.9	6248	64.98
TOTAL_SENTENCES	273641	100.0		
TOTAL_NOTES			9615	100.0

Figure 4.1: Stage 1 Initial Combined Train Data Statistics

For stage 2 experiment, we integrated extra 17-category data (with "None" filtered), including 16,781 sentences (5,767 notes) from expert-labeled hospital records (train_eb_ap_jenia_all-labels) and 42,956 sentences (7,109 notes) selected from VUMC 2023 notes (the same selection and annotation method used for AMC 2023's data). Overall, the final augmented training set contains is a heterogeneous mixture of manual and AI augmented annotations with 340,592 sentences: the original human labels for the 9 core categories were retained, and GPT-4o was used to label the new categories on those same sentences as well as to label 18 categories on unlabeled new notes drawn from AMC and VUMC sources. This resulted in a substantial increase in the coverage of previously missing categories (e.g. many sentences that were formerly unlabeled None were found to contain mentions of pain, sleep, family, etc., once those labels were introduced).

4.1.1 Annotation methods and validation

All manual annotations were provided by clinicians or expert annotators following formal definitions of each ICF category (see Appendix for category definitions). For automated GPT-based annotation, we carefully optimized the prompting procedure to ensure quality. In particular, we conducted a pilot test with 80 example sentences: GPT-4o was prompted (in zero-shot and few-shot modes) to assign labels, and two clinicians reviewed these outputs for correctness. Based on this pilot, we refined the prompt by including few-shot examples (two verified examples per category) and category definitions, which improved GPT's annotation accuracy. We also adjusted the model's temperature (set to 0.1 for more deterministic output). Using the validated prompt setting, GPT-4o then annotated the full training set for the new labels. In line

with recent findings in NLP, augmenting a small human-labeled dataset with GPT-generated labels can increase the performance of downstream classifiers (Guo et al., 2024), so this strategy was adopted to increase our training sample size. Notably, we found that a cautious approach (keeping the original human labels and only adding GPT labels for new categories or new data) was necessary, unguided GPT annotations without any human validation or thresholding tended to introduce noise (Guo et al., 2024).

4.1.2 Label distribution

The final training dataset contained 16,420 clinical notes with a total of around 340k sentences. Of these, roughly 66% of the sentences are labeled as "None", reflecting that the majority of narrative text does not belong to our specific ICF categories. This class imbalance is an important characteristic of the data. Many sentences describe general context or clinical details outside the scope of functioning categories, so the model must learn to frequently output "None". In contrast, each informative category is relatively sparse. Even the most common non-none label in training (after augmentation), Sensations of pain (B280), present in only about 5.6% of all training sentences. Other new categories also achieved non-trivial coverage: for example, Family relationships (D760) appears in around 3.2% of training sentences and Sleep functions (B134) in around 1.8%. These frequencies represent a dramatic increase from the stage 1 data. Initially, these categories were absent or very rare. By augmenting with GPT annotations and new data, we ensured that all 17 ICF categories had a reasonable number of training examples, relieving the zero-shot problem for the newly added classes. The inclusion of additional VUMC and AMC notes particularly increased the occurrence of categories like Pain and Family.

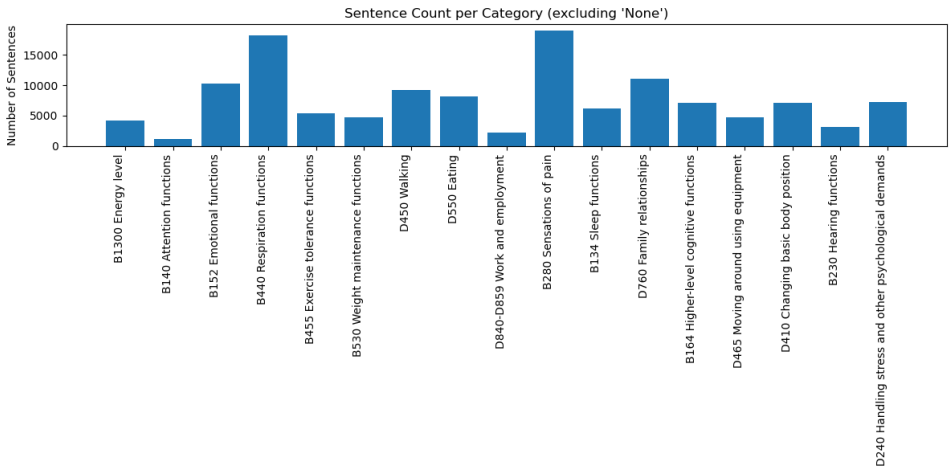


Figure 4.2: Stage 2 Final Augmented Train Data Label Distribution

Throughout this process, the original 9 categories maintained comparable representation; for instance, Respiration functions (B440) and Walking (D450) remained among the frequent categories, and their relative proportions did not drastically change with the data expansions. We refrained from aggressive down-sampling of the "None" class in training, as preliminary experiments found that maintaining the natural prevalence of "None" produced better model performance (down-sampling led to higher recall but lower precision and F1 for the informative classes, likely because the model then over-predicted labels on irrelevant sentences). The final training corpus thus reflects a balance: it is heavily skewed toward no relevant content (to mirror real note distributions), yet it contains a significantly broadened and enriched set of positive examples for each functional category.

4.2 Development Set

A portion of the data was set aside as a development (validation) set to guide model tuning and prompt configuration. This development set consisted of 10% of the annotated sentences (singled out from train data) covering the range of categories. We used another development set during the GPT prompting experiments, for example, the 80 pilot sentences validated by the clinician served as a dev sample to evaluate prompt variations and to compute macro F1 for threshold selection. In addition, the development set derived from the train data was used for iterative model hyperparameter tuning (such as adjusting learning rate and epoch schedules) and early stopping. None of the development sentences were included in training, so as to provide an unbiased estimate of model generalization during the experimentation phase.

4.3 Test Set

4.3.1 Test Set Introduction

For final evaluation, we used a test set of clinical notes that underwent a label expansion during the project. The test set is called combined_test_new_INS_fixed_FP, containing 37,355 sentences in total. The test set initially had a gold standard for 10 labels: the same 9 ICF categories as the original training set (all manually annotated by clinical experts) plus the "None" label.

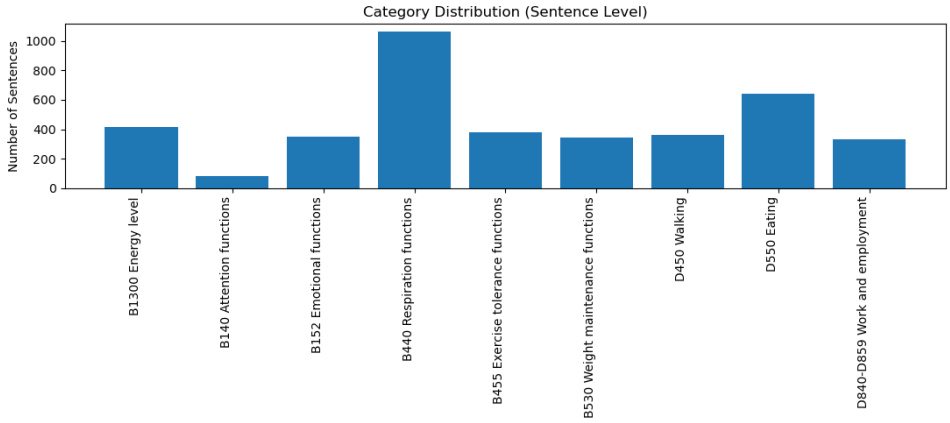


Figure 4.3: Original Test set 10-category Sentence-Level Statistics (Excluding None)

We recognized that this label set was not covering several important aspects of patient functioning documented in the notes. In particular, information about pain, sleep, cognitive function, family or social context, and certain mobility tasks was being lost under the "None" category. To address this, we expanded the taxonomy to 18 labels, adding 8 new ICF-derived categories. The new categories (with their ICF codes) were: Sensations of pain (B280), Sleep functions (B134), Family relationships (D760), Higher-level cognitive functions (B164), Moving around using equipment (D465), Changing basic body position (D410), Hearing functions (B230), and Handling stress and psychological demands (D240). These cover additional Body Function and Activity/Participation domains of the WHO ICF, aligning our annotation with a more comprehensive view of patient health. All previously defined categories remained in the scheme, so the final label set encompassed 17 ICF categories in total, plus "None".

4.3.2 Standard Update

We performed a meticulous clinician validation process to create the updated 18-label gold standard for the test set. The physician annotators reviewed every test instance predicted by the model with new categories to identify occurrences of the 8 domains, while also verifying part of the original category labels. Essentially, the test set (which contains 2,969 notes) was re-annotated under the expanded guidelines. If a sentence predicted as relevant category(ies) in a test note described, say, a pain complaint or a sleep issue, it was now labeled with the corresponding new category instead of being left as "None". The experts worked systematically through each sentence predicted by the model as new categories, and uncertained cases were confirmed via discussion to ensure consistency (for example, deciding edge cases like distinguishing general low energy from mild depression under the appropriate categories). This process produced an 18-label reference annotation that is currently considered the ground

truth for evaluation. The same members from the clinical team responsible for the initial 10-label annotations conducted the expansion, so as to maintain continuity in labeling quality and criteria.

4.3.3 Category Coverage and Shifts

The transition from 10 to 18 labels resulted in a significant increase in the amount of information labeled as relevant in the test notes. Under the original 10-label scheme, only about 10.6% of test sentences had any category label (the rest were implicitly 'None'). After introducing the new categories, approximately 15.2% of sentences are labeled with at least one category, meaning the pool of recognized informational content grew by roughly 40%. Much of this gain comes from the new categories. For example, Pain (B280) was entirely unannotated in the original version, but in the updated gold it is now the single most prevalent category: it occurs in 2.3% of all test sentences, and about 15.4% of test notes contain at least one mention of pain. This highlights how frequently pain-related statements appear in clinical narratives (e.g. symptoms, discomfort levels) that were previously overlooked by the label set. Similarly, Family relationships (D760), which covers references to family support or issues at home, is present in 1.3% of sentences (9.2% of notes) in the new gold standard, whereas such content was formerly lumped under "None". Sleep functions (B134), covering statements about sleep quality or patterns, now accounts for 0.5% of sentences (found in 4.5% of notes). Other added categories show a more modest presence, for instance, Handling stress (D240) appears in 0.6% of sentences and Hearing functions (B230) in just 0.1% – but even these provide valuable nuance by identifying patient problems that the initial categories did not address.

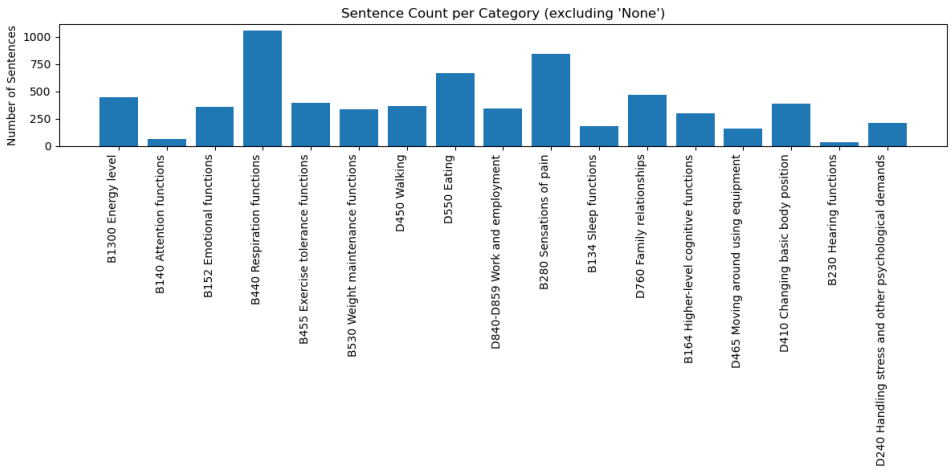


Figure 4.4: Updated Test set 18-category Sentence-Level Statistics (Excluding None)

Importantly, the incorporation of new labels did not invalidate the original annotations; rather, it augmented them. The clinicians' re-validation led to only minor adjustments in the counts of the 9 original categories. For example, a handful of sentences were reclassified upon closer review (some entries originally marked as Attention functions (B140) were corrected, reducing its count from 82 to 64 sentences, while Energy level (B1300) saw a small increase from 413 to 448 sentences after resolving ambiguities). Besides these small changes, the distribution of the legacy categories in the test set remained largely consistent, indicating that the expansion primarily added new information on top of the existing labels. After re-annotation, the most frequent original categories in the test notes are still Respiration functions (B440) and Eating (D550), each covering roughly 13-14% of notes (e.g. breathing issues and mobility are commonly documented in this patient record).

Despite the richer labeling, it should be noted that the majority of the test set's text is still labeled "None". About 84.8% of all test sentences carry no functional category label, a reflection of how much clinical documentation consists of context, history, or other details outside the specific ICF functional domains. Nearly every note (99% of notes) contains at least one such irrelevant sentence (e.g. administrative details, unrelated observations). This underscores the challenge for the model, it must identify relatively sparse signals of relevant content within a large backdrop of irrelevant text. The expansion to 18 labels helps surface more of those signals (especially for pain, cognitive and social aspects), but the task remains heavily skewed towards correctly outputting "None" when appropriate. In summary, the test set revision ensured that previously hidden but important patient information (pain, sleep, etc.) is now explicitly labeled, providing a more comprehensive evaluation benchmark. The categories chosen are grounded in the WHO ICF framework for describing health and disability, which lends clinical validity to the annotations. The refined 18-label gold standard allows us to evaluate the model's performance on a broader spectrum of functional health information, while still being comparable to the original scope for the core 9 categories.

Overall, our data preparation and annotation strategy contributed to a training set that progressively grew from a small expert-annotated base to a large GPT-augmented corpus covering all target classes, a curated development set for validation, and a re-annotated test set reflecting an expanded conceptual model of patient functioning. This approach balanced innovation (using GPT-4o to assist annotation) with expert oversight, and was guided by established frameworks like the ICF to ensure that the categories and labels have meaningful clinical interpretations. The result is a robust foundation for training and evaluating our multi-label classification model on clinical text.

Chapter 5

Methodology

5.1 Task Definition and Label Inventory

This thesis addresses a multi-label sentence-level classification task on Dutch clinical notes, where each sentence is classified into one or more relevant categories from the ICF. Unlike single-label classification, multi-label means a sentence can be assigned multiple categories (or none) simultaneously, reflecting the fact that a single clinical statement may pertain to several aspects of patient functioning. The label inventory consists of 18 categories, derived from ICF domains of functioning and disability. Specifically, 17 categories correspond to second-level ICF codes covering mental functions, mobility, self-care, etc., and an additional "None" label indicates that no ICF-related information is present in the sentence. The labels contain original set of 9 ICF codes used in previous work as well as 8 newly introduced categories. The "None" category is used when a sentence does not describe any information relevant to these ICF domains. In total this results in 17 ICF categories plus the None class (18 labels). Thus, the classification model must learn to output a set of 0, 1, or multiple ICF codes for each sentence. We frame this as 18 parallel binary classifications (one per label), allowing any combination of labels to be assigned.

The decoded order of 18 parallel binary classification is: ["B1300 Energy level", "B140 Attention functions", "B152 Emotional functions", "B440 Respiration functions", "B455 Exercise tolerance functions", "B530 Weight maintenance functions", "D450 Walking", "D550 Eating", "D840-D859 Work and employment", "B280 Sensations of pain", "B134 Sleep functions", "D760 Family relationships", "B164 Higher-level cognitive functions", "D465 Moving around using equipment", "D410 Changing basic body position", "B230 Hearing functions", "D240 Handling stress and other psychological demands", "None"]

This task builds upon recent efforts to apply NLP to ICF-based coding of clinical text. Previous studies have demonstrated that it is feasible to classify narrative sentences or phrases into

ICF categories with encouraging accuracy. For instance, Meskers et al. (2022) annotated sentences for four ICF categories (Emotional functions, Exercise tolerance, Walking & Moving, Work & Employment) and trained a neural model, achieving F1-scores around 0.70 for certain categories. Similarly, Newman-Griffis et al. (2021) developed NLP systems for tagging patient functioning information with ICF codes, reporting over 80% macro-averaged F1 on multi-label ICF classification in rehabilitation notes. These works underscore the viability of automated ICF classification and motivate our extension to a broader set of 17 categories. Our task is novel in expanding the label space to 17 ICF categories (covering both body functions and activities/participation domains) and focusing on sentence-level granularity, which enables detailed identification of functioning information in clinical documentation.

5.2 Data Construction and Annotation Pipeline

Our data preparation involved integrating manually labeled datasets from a previous project with newly collected unlabeled notes, and then using a GPT-4o-assisted annotation strategy to assign the expanded set of categories to all sentences.

5.2.1 Manual Annotations (Original Dataset)

We began with an existing corpus of Dutch hospital notes that had been annotated by clinical experts for a set of 10 categories (the 9 ICF codes listed above, plus "None"). This dataset was compiled from prior annotation batches (referred to as Murat's batches and Jenia_new_INS_10 in our project logs) and contains thousands of sentences with gold-standard labels for the original 9 ICF categories. These manual annotations, produced with high inter-annotator agreement in earlier work, serve as a reliable foundation for the original label set. However, since this thesis aims to expand the classification to 17 ICF categories, additional data and annotations were required for the 8 new categories not present in the original scheme.

5.2.2 New Category Data

In consultation with the medical team, we identified keywords and clinical concepts related to the 8 new ICF categories to retrieve relevant sentences from unstructured clinical notes. We selected a large collection of Dutch hospital notes from 2023 (from Amsterdam UMC, labeled

here as AMC 2023) as a source of new data. Using the provided keywords (including terms related to, for example, pain, sleep, family relationships, cognitive functions, use of assistive equipment, body position changes, hearing, and psychological stress handling, as well as a minority old category B455 "Exercise tolerance"), we searched the notes for occurrences of those terms. This produced a subset of documents likely to contain information on the new categories. We then segmented these notes into sentences (detailed under Preprocessing below) and extracted the sentences containing the keywords or surrounding context. This resulted in a candidate set of sentences that potentially cover the new ICF categories. At this stage, these sentences were unlabeled (since no manual annotations existed for the new categories yet).

Next, we merged the new unlabeled sentences with the originally annotated sentences to form a comprehensive training pool. To clarify, Dataset A consisted of the original manually annotated sentences (with labels in the 10-category scheme). Dataset B consisted of the newly selected sentences from AMC 2023 notes that likely contain the new category information. We combined A and B (Dataset C) to obtain a training set covering all 18 ICF categories, though initially only the sentences in A had labels (for old categories) and sentences in B had no labels yet.

5.2.3 GPT-4o Assisted Annotation

To label Dataset C with the full 18-label inventory, we employed a weak supervision approach using OpenAI's GPT-4o model as an annotator. Recent literature suggests that LLMs like GPT-3/4 can serve as powerful labeling functions in the absence of sufficient human-labeled data (Oliveira et al., 2025). The advantage is that an LLM can encode a great deal of medical knowledge and linguistic context, potentially providing "good enough" labels at scale, which can then be used to train a dedicated model. This strategy uses the LLM's strengths without requiring it to be deployed in real time: previous work has found that using an LLM to generate training labels and then training a smaller model can outperform using the LLM directly for prediction. We adopted this paradigm, treating GPT-4o as a noisy labeler that provides initial annotations for the new categories.

We carefully designed a prompt for GPT-4o to classify a given sentence into the 17 ICF categories. The prompt provided GPT-4o with: Definitions of each ICF category, concise descriptions based on the ICF manual (e.g. B164 Higher-level cognitive functions: complex goal-directed mental functions such as decision-making, planning, and judgment), to ensure the model understood each label's meaning; Few-shot examples, we included two example sentences for each category (2*17 = 34 examples in total) illustrating a sentence and its correct

label. These examples were carefully chosen from our data: for the 9 original categories, we selected two prototype sentences from the manually labeled set (with high confidence gold labels); for each of the 8 new categories, we took two sentences from the newly collected data B that we manually verified (with the help of a clinician) as clear instances of that category. This built a robust few-shot prompt demonstrating how to label each category. We also instructed GPT-4o to assign multiple labels if appropriate or "None" if no category applied, explicitly clarifying that "None" means no relevant content.

We conducted a pilot study with 80 sentences to refine this prompt and evaluate GPT-4o's annotation quality. We first ran GPT-4o in a zero-shot mode on a subset of the unlabeled data (i.e. providing only the task instruction and definitions, with no examples) to get an initial guess of labels. From these, we selected 80 sentences that covered a variety of categories (10 sentences for each of the 8 new categories, ensuring we had some GPT-4o predicted positives for each new label). These 80 sentences (which by GPT's guess had at least one of the new categories) were then validated by clinical experts, the medical team reviewed each sentence and provided the true labels for the new categories, essentially creating a small gold standard set for evaluation. Using this as a benchmark, we experimented with different prompt settings for GPT-4: zero-shot vs. few-shot prompting; including category definitions vs. no definitions; different temperature settings.

We found that few-shot prompting with definitions at a low temperature (0.1) offered the best accuracy, with GPT-4o's labels matching the expert labels on about 61.5% of the 80 pilot sentences. In contrast, zero-shot or higher-temperature settings performed worse (we observed more inconsistent or incorrect labels when examples were not provided or when temperature was higher, leading to variability). Thus, the final prompt we adopted for full data annotation was a few-shot prompt with 34 examples (2 per category) + definitions, temperature 0.1. This framework effectively guided GPT-4o to mimic a structured classification approach, and while its accuracy (61.5%) was far from perfect, it provided a substantial starting point for training. Notably, using GPT-4o as a labeler is consistent with recent research that, despite not reaching human-level labeling quality, LLM-generated labels can be considered a "valid option" for model training when human annotation is expensive (Oliveira et al., 2025), especially if combined with some human verification.

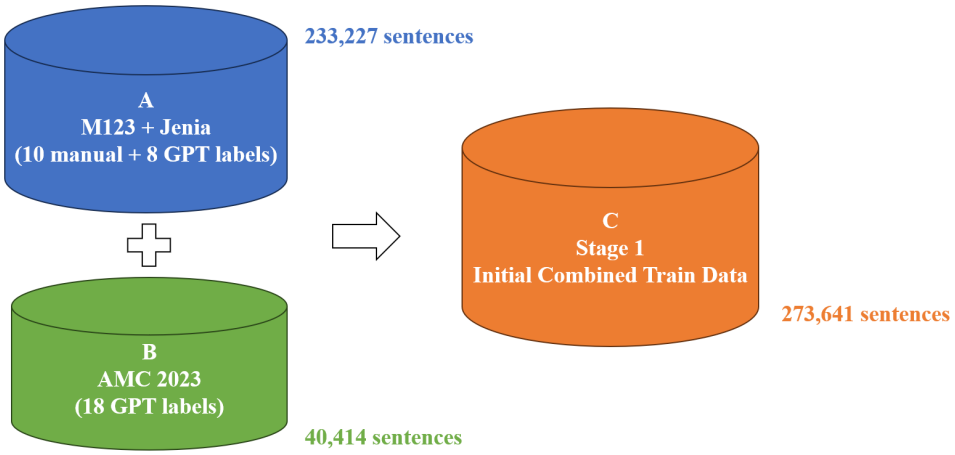


Figure 5.1: Dataset C - Initial Combined Train's Composition

With the chosen prompt, we fed all sentences in Dataset C to GPT-4o and obtained predicted labels for the full 18 classes. For sentences originating from the manual set A (which already had human labels for the old 9 categories), we retained the original human labels for those categories, and only used GPT-4o's suggestions for the 8 new categories. In other words, the human gold labels for original categories were considered authoritative (we did not override them with GPT, to avoid noise on those), and we only augmented those sentences with any new categories GPT-4o identified. For sentences from B (newly selected from AMC-2023) which had no existing labels, we took GPT-4o's multi-label output as the provisional labels for all 18 categories (which in practice often was one or more categories per sentence, or "None" if GPT thought it irrelevant). The outcome was a weakly labeled training set covering all categories: effectively, the old categories have a mix of human and GPT labels (human-annotated labels for 10-category data in set A, GPT-labeled 10-category data for set B), while the new categories have exclusively GPT-provided labels (GPT-labeled 8 new categories for set A and B).

5.2.4 Data Augmentation

After initial model experiments (described below), we took additional steps to improve the training data coverage and balance. One challenge was the dominance of the "None" class. In raw clinical text, many sentences do not describe functional issues, so "None" would be extremely frequent, which risks skewing the model to always predict no finding. In the initial combined set C, it's true that a large fraction of sentences were labeled "None" by GPT-4o. Simply downsampling or removing "None" examples indiscriminately could drop useful negative examples and hurt generalization, as we discovered: we tried a version of training data where we randomly kept only 5% of the "None" labeled sentences to balance the label distribution, but this actually degraded performance on the minority categories.

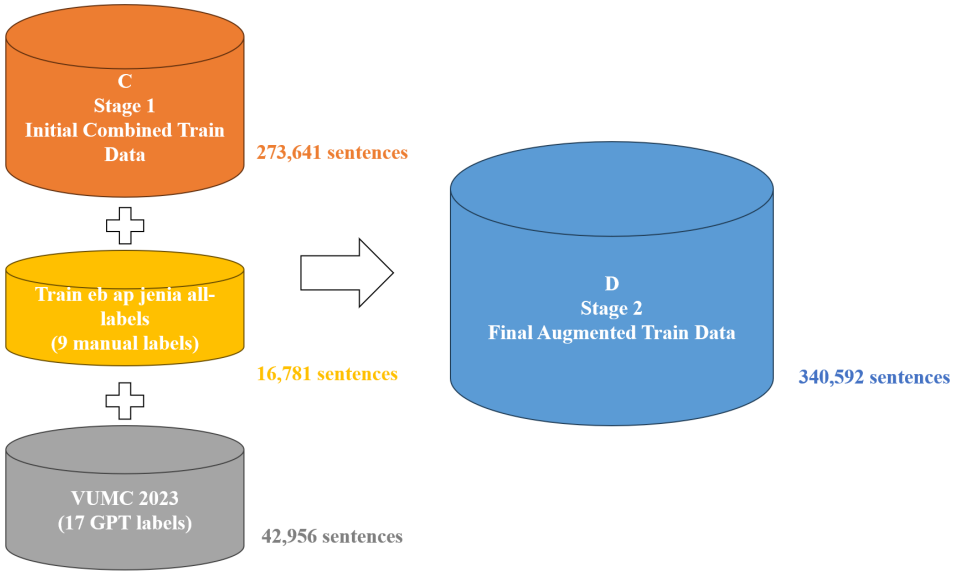


Figure 5.2: Dataset D - Final Augmented Train's Composition

Instead, we chose a more targeted augmentation strategy. We incorporated two additional datasets into training that focused on positive examples of ICF categories: an older dataset of rehabilitation notes from a previous study (train_eb_ap-labels) which contained sentences labeled with the 9 original categories excluding "None" (here, any sentence without a category had been filtered out); and a new set of 2023 notes from VUMC which we similarly annotated with GPT-4o and then filtered to remove sentences labeled "None". By adding these, we injected a larger number of category-positive instances, particularly for underrepresented categories, without further expanding the volume of "None" examples. Finally, we merged everything into one augmented training set (Dataset D): combining the GPT-annotated Dataset C with the no-None older manually annotated data and the no-None GPT annotated VUMC data. We also ensured to deduplicate this final set: some sentences, especially generic ones, could appear in multiple sources, so we removed any exact duplicate sentences that had the same note ID and sentence index to avoid overweighting them. The resulting corpus for training contains sentences from multiple origins (different hospital records and time periods), all labeled in a unified 18-category scheme. We set aside 10% of this data as a development set (validation set) for hyperparameter tuning and threshold calibration. The split was done stratified by label to the extent possible, and using distinct notes to avoid overlap between train and development set.

5.2.5 Test Set and Clinical Validation

For evaluation, we relied on a set of manually annotated data. The initial test set was derived from the earlier project's data and consisted of a collection of sentences with gold labels for the original 9 categories (and None). This test set had been curated and double-annotated by clinical experts for the old categories, providing a solid benchmark for the model's known capabilities. However, since our expanded label inventory includes 8 new categories that were never annotated in the original test, we needed to obtain ground truth for those as well to fully evaluate 18-category performance.

We thus conducted a retroactive annotation of the test set for the new categories. We fine-tuned the model with Dataset D and generated predictions on the test set, then we selected sentences predicted as new categories and have them reviewed by a clinical expert panel with knowledge of the ICF, and any applicable labels among the 8 new categories (sometimes the old 10s) were assigned. In essence, we "filled in" the new category labels for each test sentence while generally keeping the original labels for old categories (minor adjustments applied). This produced a complete 18-category gold standard test set. Note that the new category annotations for test were done after model development, so as to avoid peeking at test data during training, and were validated carefully by multiple clinicians, although due to time constraints a full multi-annotator agreement study on the new categories was not performed on the entire test (a small subset was double-annotated to ensure basic consistency). The test set ultimately allows us to evaluate how well the model identifies all 17 ICF categories in unseen sentences. Importantly, none of the GPT-generated labels were used for evaluation on test, test labels are fully derived from human experts for all categories, so that our evaluation is against a reliable ground truth. We also ensured that the test set is disjoint from the training data in terms of notes and patients, to avoid any leakage.

By the end of this data construction pipeline, we had: a large training set of 306,533 (340,592 - 34,059) sentences (several thousand, from multiple sources) labeled for 18 classes, containing weak labels with some human labels mixed in; a development set of 34,059 sentences for tuning, and a test set of 37,355 sentences with high-quality manual labels for evaluation. The combination of manual and GPT-assisted annotation allowed us to considerably expand the coverage of ICF categories without an exhausting manual labeling effort, aligning with the weak supervision approach of using LLMs to expand training data (Oliveira et al., 2025). While the GPT labels are noisy, training a dedicated model on a large quantity of such labels can still bring good performance that often surpasses using the GPT model directly on the task. Our methodology also incorporated expert validation at key points (prompt design, test labeling) to ensure clinical relevance and correct any systematic errors in the LLM annotations.

5.3 Preprocessing of Clinical Text

Before training the model, we applied a series of preprocessing steps to the notes and sentences to normalize the text and structure the inputs.

5.3.1 Sentence Segmentation

Each clinical note in the train data was split into sentences using spaCy, a robust NLP toolkit. We used a spaCy model for Dutch (expanded with medspaCy rules for clinical text) to perform sentence boundary detection. This step is important because our classification unit is the sentence. Clinical notes often contain irregular punctuation or formatting (such as newlines, bullet lists, etc.), so we refined the segmentation rules to ensure that medically meaningful segments are kept together. For instance, we prevented abbreviations or numeric lists from prematurely breaking a sentence. Each sentence was assigned a unique identifier composed of the source note ID and its sentence index, which we carry through to model inference and evaluation. This made it possible to later aggregate predictions back to the note level and also to remove duplicates as described below.

5.3.2 De-identification and Identifier Normalization

All notes were de-identified to protect patient privacy, in accordance with ethical regulations. This involved removing or masking any personal health identifiers such as patient names, dates, identification numbers, or hospital-specific codes. In the text, we replaced such identifiers with generic placeholders (e.g., "PERSON") or simply omitted them. This normalization prevents the model from treating each unique identifier as a distinct feature, instead, the model sees a consistent token for any identifier, reducing sparsity and risk of overfitting to trivial text patterns. For example, if a sentence originally was "Mevrouw Jansen liep 10 meter zonder hulp" ("Mrs. Jansen walked 10 meters without help"), it would be normalized to "PERSON liep 10 meter zonder hulp" so that the model doesn't cling to the specific surname. All notes were processed in this way before feeding into GPT-4o or the classifier, ensuring no confidential or identifying information is exposed or learned by the models.

5.3.3 Train-Dev-Test Split

The development set was sampled as 10% of the full training pool (Dataset D). We performed this split at the note level, i.e. entire notes were assigned either to train or dev, to prevent sentences from the same note appearing in both sets (which could leak contextual clues). This means some patients' data are entirely in train or entirely in dev. We also stratified by ensuring that each ICF category had at least a few positive examples in dev (given the label imbalance, a purely random split might miss some minor categories in the dev set). The final dev set was used for model selection and threshold tuning, and only after finalizing the model were the test results computed on the separate test set (which, as described, was annotated independently). The test set itself was fixed from the start (coming from the earlier dataset) and was not involved in any training or tuning decisions.

The preprocessing pipeline provides clean, sentence-level data ready for modeling. Each sentence is a standalone input, free of identifiable information and appropriately labeled, with mappings back to note-level context as needed. These steps ensure that both the GPT assisted labeling and the final model training are performed on consistent, high quality text segments.

5.4 Model Architecture and Training Procedure

For the classification model, we chose MedRoBERTa.nl, a Dutch-language Transformer model pretrained on clinical text (Verkijk & Vossen, 2021). MedRoBERTa.nl is a RoBERTa-based architecture (12-layer bidirectional Transformer) that was specifically trained from scratch on 13GB of Dutch hospital notes. Verkijk and Vossen (2021) introduced this model as the first domain-specific language model for Dutch EHRs, showing that it outperforms generic Dutch language models on understanding. We opted for MedRoBERTa.nl as the text encoder because it provides rich contextualized representations of Dutch clinical language, including medical jargon and idiosyncratic note-writing style, which we expect to be advantageous over a general-language model. Indeed, MedRoBERTa.nl has demonstrated superior performance on clinical classification tasks (e.g. identifying sentences about patient mobility) compared to base Dutch. Thus, it is well-suited for our task of ICF category classification.

We fine-tuned MedRoBERTa.nl for multi-label classification. We used the Simple Transformers library, which provides a convenient wrapper for Hugging Face Transformers in multi-label classification tasks. The training data were prepared with each text paired to a multi-hot label vector indicating all applicable categories. Input texts were tokenized and truncated or padded to a maximum length of 512 tokens to fit the model input size. We trained the model using

binary cross-entropy loss (per label) and optimized with AdamW (learning rate = 4×10ˆ5) for a single epoch, using a small batch size of 8. A held-out validation set was used to monitor performance during training (evaluating every 1,000 steps), and the best model checkpoint was saved for final evaluation. We did not employ early stopping, instead, the full training run was completed and the model with the highest validation metric was retained.

No class weighting was applied. Instead, we relied on our data augmentation (inclusion of more minority class examples) and threshold tuning to handle imbalance. The model sees all "None" vs "category" examples as per their frequency in the training data. We did consider downsampling "None" during training but removing too many "None" examples hurt the model's ability to identify truly irrelevant sentences. The final training data (Dataset D) still had a skew toward "None" (though less extreme than original), so to prevent biasing the classifier to always predict "None", we ensured the loss treated each label equally and let the data balance speak for itself. Notably, the MedRoBERTa.nl model was fine-tuned in a continued pretraining manner for some experiments: one of our trials involved first fine-tuning the model on only the manually labeled 10-category data (to give it a strong basis in those), then fine-tuning further on the 18-category data. At sentence level, this two-stage approach produced clear improvements over a single-stage fine-tune on the combined data (performance was better). Therefore, our final model was trained in a single stage on the full augmented data, starting from the base MedRoBERTa.nl pretrained weights.

The training procedure naturally integrated the GPT-labeled examples with human-labeled ones, all were in the training set together. We did not explicitly weight GPT-derived labels as less important; each training sample's contribution to the loss is the same. The rationale was that the model, if sufficiently regularized and validated on dev, would learn to generalize and could overcome some label noise by means of the large quantity of GPT-labeled samples. Essentially, the manually labeled sentences served as high quality backbones (especially for the original categories), while the GPT labels provided spaces in covering new categories and additional contexts. This approach is consistent with findings that combining a few high-quality labels with many weak labels is likely to result in the best of both worlds (Zhang et al., 2021). We did verify that performance on the original 9 categories did not degrade when training on the mixed (human + GPT) annotations compared to training on human-only data; in fact, adding the GPT-annotated data improved recall for some categories without severely hurting precision (see Results for analysis). This suggests the model managed to extract useful signal even from imperfect GPT labels, a phenomenon similar to knowledge distillation or data augmentation, where the noisy labels provided additional variance for the model to learn from. Similar observations have been made by Oliveira et al. (2025) in the legal domain, where a model trained on GPT-annotated data plus some human data achieved performance close to a fully human-trained model. In our case, the final fine-tuned model effectively extracts

GPT-4o's latent knowledge (and errors) into a smaller deployable model (MedRoBERTa.nl) that can run locally without requiring GPT-4o at inference time.

After training, the best-performing model was applied to the test set to predict labels for unseen texts. Given an input text, the model outputs a score for each of the 18 categories, which we converted into a confidence probability for that category. We then thresholded these probabilities to decide the binary label predictions. In this way, the model can assign multiple categories to a single text if their confidence exceeds the threshold. The predicted label sets (and their confidence scores) for each test instance were saved for subsequent evaluation against the ground truth labels.

5.5 Threshold Standard for Multi-Label Decision

In our multi-label setting, the model's raw confidence scores needed careful calibration to decide which labels to assign. Using a single global threshold (e.g. 0.5) was inappropriate because the model's predicted probabilities for positive labels were typically much lower (often below 0.1 in our case). Without adjustment, many true labels would be missed.

For each label, we gathered the model's confidence scores for all true positive instances and all true negative instances in the validation data. This results in two distributions per label: one for when the label is actually present, and one for when it is absent.

We identified the lowest confidence among true positives and the highest confidence among true negatives for each label. These values represent the borderline cases: a true instance barely detected, and a false instance nearly mistaken as true. In most categories, we observed significant overlap, for example, some positive examples had a confidence as low as nearly 0.08, while certain negative examples had confidence up to nearly 0.085. A few labels showed a cleaner separation (e.g. one category's lowest positive score exceeded all negative scores by a small margin).

All model output probabilities on the test set were then binarized using label-specific thresholds (rather than a generic 0.5). In practice, the calibrated thresholds turned out to be very low (around 0.08-0.09) across the 18 labels. This reflects the model's scoring tendencies, the confidences for true positive cases were relatively small due to the multi-label probability normalization. For example, the threshold for most clinical categories fell near 0.082-0.085, and even the "None" category (which had the highest baseline scores) used a threshold of about 0.09. Any label's predicted probability above its threshold was marked as present for

that sample, while lower scores were considered absent.

Using these per-label decision thresholds on the test results allowed us to improve detection of relevant categories. We avoided missing positives that had low raw scores, and simultaneously controlled false positives by not overly lowering the cut off for any label.

5.6 Evaluation Protocol

We report and evaluate the classification performance at two ranks: sentence-level (the primary unit of prediction) and note-level (aggregating sentence predictions per clinical note).

5.6.1 Sentence-Level Evaluation

This treats each sentence as an independent instance with one gold label set and one predicted label set. We compute standard classification metrics: for each ICF category, we calculate precision, recall, and F1 score comparing the predicted vs. true labels across all sentences in the test set. We then report the macro average of these metrics over the 17 ICF categories. Macro F1 is our key measure, as it gives equal weight to performance on each category, aligning with our goal to do well even on the less frequent categories. Sentence-level evaluation directly measures how well the model can identify specific functional information in individual utterances. A true positive occurs when a sentence that should be labeled with category X is indeed predicted as X; a false negative is when the model misses a category present in the sentence, and a false positive is when the model assigns a category that isn't actually applicable to that sentence. Because our test set is fully labeled for all categories, any label not assigned in the gold standard is considered truly absent for that sentence, so we can count false positives for all categories (this includes if the model predicts a new category for a test sentence that was originally labeled "None" or only had old categories, and the annotators confirmed that new category was not actually present).

5.6.2 Note-Level Evaluation (OR-aggregation)

Many practical use cases (e.g. generating a patient's functional summary or tagging an entire clinical note with relevant codes) operate at the note level. We want to know, for each note,

which ICF categories are mentioned anywhere in that note. Since our model processes one sentence at a time, we provide note-level predictions by an OR-aggregation rule: for each category, if any sentence in the note is predicted to have that category, then the note as a whole is labeled with that category. In other words, a note is positive for a category if at least one of its sentences was identified as such by the model. This approach is logical because if, say, "Walking" is discussed in any sentence, the entire note is clearly relevant to Walking (at least at some point).

We are interested in note-level macro F1 since that speaks to how well the system would perform in tagging entire documents (which is relevant for integrating with an Electronic Health Record system, e.g. automatically labeling a patient's record with their functioning issues). It aligns with clinical expectations because if the model finds at least one mention of a problem, the patient likely has that issue noted.

In short, the evaluation will report performance at both levels. The sentence-level evaluation reflects the model's fine grained accuracy and is useful for analyzing which categories are detected well and which are missed at the level of individual statements. The note-level evaluation reflects the end users' perspective, if this system were used to auto label entire notes, how often would it correctly identify that a note contains info about, say, "Pain" or "Walking", and how often would it miss it or raise a false alarm.

5.6.3 Baseline Comparison (GPT-4o Inference)

In addition to our fine-tuned MedRoBERTa.nl model, we evaluated GPT-4o's performance as a direct classifier on this task as a baseline for comparison. Large language models like GPT-4o have shown few-shot capabilities and one might wonder if we could skip model training altogether and just ask GPT-4o to label new sentences (zero-shot or few-shot). Using the same prompt design developed during annotation, we prompted GPT-4o to label each test sentence and measured its accuracy against the gold standard. Specifically, we used the best performing prompt identified earlier: a few-shot prompt with two examples per category and the category definitions provided, instructing GPT-4o to list all applicable categories for the given test sentence. This prompt to GPT-4o at inference time is essentially the same setup as we used to annotate training data (except the examples in the prompt remained the ones from training phase; we did not include any test sentence-specific hints, of course). We also experimented with a zero-shot prompt (just instructive text and definitions, no examples) as an alternate baseline, to assess how much the few-shot examples help GPT-4o.

The GPT-4o baseline is interesting for a few reasons. It represents the performance of a state-of-the-art LLM without fine-tuning, which is useful to see how far a large model can go on this classification out-of-the-box; it provides a point of reference for our fine-tuned model: if our approach is effective, the fine-tuned MedRoBERTa.nl should meet or exceed GPT-4o's accuracy, validating the weak supervision and finetune strategy (Zhang et al., 2021); and it helps quantify the benefit of task-specific training. We note that GPT-4o was run under the same constraints as our model (it only saw each sentence, not full notes at once, since our task is sentence classification; for note-level results we aggregate GPT's sentence predictions with OR as well). We did not apply threshold calibration to GPT-4o's outputs; instead, we interpret its output labels directly (GPT-4o was instructed to only output the labels it considers positive, which inherently is its own internal thresholding). In practice, GPT-4o sometimes missed subtle categories or over-predicted some categories; the few-shot prompt mitigated some of that by giving it clear examples of each category's context. The results in the next chapter will show how GPT-4's macro-F1 compares to the fine-tuned model. Previous research in analogous settings (e.g. legal text classification) has found that models trained on GPT-labeled data can approach or exceed the zero or few-shot GPT performance itself (Oliveira et al., 2025), which is what we anticipate: the fine-tuned model effectively specializes and possibly corrects some inconsistencies of GPT. Indeed, using the LLM in a loop (for data generation) rather than directly is likely to result in better results.

All evaluations (for both our model and GPT-4o) were done under identical test conditions, using the same set of test sentences and gold labels. We compute macro metrics, and we also examine per category's scores to see where each method does well or poorly. We will present confusion matrices and example errors in the Error Analysis chapter to provide further insight beyond summary metrics.

5.7 Reproducibility and Implementation Details

We are committed to making this work reproducible. All data preprocessing, model training, and evaluation steps are implemented in documented scripts. The code (in Python, using libraries such as Hugging Face Transformers, spaCy, and scikit-learn) is available in the project repository. We fixed random seeds in all experiments (for example, the train/dev split selection, weight initialization, and shuffling in training) to ensure that results can be replicated exactly. The label ordering (i.e., the index assignment of the 18 labels) was standardized and kept consistent throughout the project, this is important because the model's output vector has to be interpreted in the correct order. We provide a reference mapping of label names to indices (for instance, index 0 = B1300 Energy level ... index 16 = D760 Family relationships, index 17 =

None) in our documentation, so that anyone loading the model or using our evaluation script will apply thresholds to the correct output dimension and calculate metrics for the correct label.

For data splits, we archive the exact lists of note IDs (or sentence IDs) that went into the train, dev, and test sets. This means someone with access to the same raw data could reconstruct the splits exactly. Our test set in particular is identified by a set of unique keys (combining note and sentence identifiers); this allows comparing any future model's predictions on the same test instances.

Our use of GPT-4o in data annotation is also documented: we provide the exact prompt template used, including the instructions, category definitions, and example format, so that one could re-run GPT-4o on the same sentences and verify the label outputs. Note that GPT-4o, being a non-deterministic model (especially with temperature 0.1, though that is low, some minor randomness remains), may not produce identical labels every run. To mitigate this, we recorded the outputs from our annotation run. Those outputs (the GPT-4o labeled dataset) are stored so one can use them directly rather than calling the API again.

With similar clinical data and access to MedRoBERTa.nl and GPT-4o, our pipeline following the provided documentation can be replicated. Reproducibility is significant in clinical NLP, and we have made efforts to be transparent at every step, from data preprocessing decisions to model hyperparameters and evaluation scripts.

Chapter 6

Results

6.1 Baseline Performance (Run 0: 10-Category Model)

Our starting point is a MedRoBERTa classifier trained only on the original manually annotated dataset with 10 ICF categories (including the "None" class for no relevant content). This baseline model (Run 0) achieved a macro F1 of approximately 0.59 at the sentence level for the 10 categories, with macro precision around 0.77 and recall 0.51. In other words, the model was fairly precise in its predictions but missed nearly half of the relevant category instances. The high precision and low recall pattern suggests a conservative classifier that avoids false positives but at the cost of many false negatives.

Table 6.1: Five training runs and macro scores

Run	System	Data	Training	Macro F1 (10)	Macro F1 (18)
0	Baseline 10-cat	A (manual annotations only)	1 epoch, LR 4e-5	0.59	
1	+GPT data 18-cat	C (0 + AMC 2023 GPT labeled)	1 epoch, LR 4e-5	0.62	0.61
2	None downsample	1 with “None” class 5% sampling	1 epoch, LR 4e-5	0.60	
3	Aug 18-cat	D (1 + extra manual & VUMC 2023 GPT labeled)	1 epoch, LR 4e-5	0.67	0.68
4	Longer training	D	5 epochs, LR 4e-5	0.64	
5	GPT-4o		few-shot, definitions, temp 0.1	0.57	0.58

6.2 Incorporating GPT4o Annotations (Run 1))

To address the data scarcity in underrepresented categories, we expanded the training set with GPT-4o assisted annotations. Specifically, additional sentences (from recent AMC 2023 notes) were labeled by GPT-4o (with few-shot prompting and ICF definitions) for the full 18-category inventory, then combined with the original manual data. Run 1 refers to the model trained on this combined dataset (set C) without any downsampling. Importantly, this introduced 8 new ICF categories (e.g. pain, sleep, family support) that were previously not present. On the original 10 categories, the effect of adding GPT-labeled data was immediately apparent: macro F1 improved to 0.62, driven largely by an increase in macro-average recall to 0.54 (vs 0.51 baseline) while precision stayed high (0.77). The micro F1 also rose (to roughly 88%), indicating better overall accuracy. In other words, adding the GPT-synthesized training data enabled the model to catch more positives across categories without sacrificing precision. This suggests the synthetic labels were of reasonable quality and helped learn decision boundaries for minority classes that the baseline often missed. For example, category B140 Attention functions, which had been missed in many cases by the baseline, saw a noticeable jump in recall. We also observed improvement in detecting work-related activity (D840-859) and respiration functions (B440) sentences, as the GPT-augmented data included more varied expressions of fatigue and work participation that the model could learn from. These trends confirm that data augmentation via GPT-4o can mitigate some class imbalance issues by providing pseudo-labeled examples for rarer categories. This approach is supported by recent studies showing that LLM generated data can improve multi-label text classification in low-resource settings (Hu et al., 2025; Zhao et al., 2024).

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW
precision	0.87	0.91	0.70	0.81	0.64	0.67	0.62	0.80
recall	0.49	0.37	0.36	0.63	0.57	0.70	0.17	0.64
f1-score	0.63	0.52	0.47	0.71	0.60	0.69	0.26	0.71
support	1063.00	82.00	331.00	413.00	640.00	362.00	382.00	341.00

	STM	none	micro avg	macro avg	weighted avg	samples avg
precision	0.70	0.97	0.95	0.77	0.95	0.84
recall	0.62	0.87	0.84	0.54	0.84	0.84
f1-score	0.66	0.92	0.89	0.62	0.89	0.84
support	348.00	33769.00	37731.00	37731.00	37731.00	37731.00

TRUE \ PREDICT	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	none
ADM	522	1	0	25	10	19	15	5	5	386
ATT	1	30	1	6	0	1	0	0	2	4
BER	1	2	118	9	0	5	3	5	3	90
ENR	26	2	4	259	4	12	26	9	4	34
ETN	6	0	1	5	362	1	0	36	2	222
FAC	10	0	0	7	0	254	14	0	3	13
INS	22	3	3	41	1	58	64	3	3	83
MBW	2	0	0	6	48	1	2	218	8	83
STM	5	0	7	3	1	3	1	3	216	37
none	68	3	49	50	184	100	19	42	80	29545

Figure 6.1: Run 1 initial combined pool (10-category results)

It is worth noting that because the model was now trained on 18-category data, it occasionally predicted the new categories on the test set, which at this stage had gold labels for only the original 10. These predictions (e.g. flagging "family relationships" or "pain" in a sentence) were counted as false positives under the old evaluation. In fact, some sentences labeled as "None" in the original gold were being flagged with new ICF codes by the model. For instance, the model might tag a sentence mentioning family with D760 (Family relationships) or a sentence describing pain with B280 (Sensation of pain), predictions that were technically errors given the incomplete gold standard. This phenomenon highlighted an important point: the model was discovering previously unannotated ICF information. Later, when we obtained human validation for the new categories on the test set, many of these early "false positives" turned out to be legitimate positives (e.g. the patient's note did discuss family support or pain). This demonstrates the value of the expanded label set: the system began to identify clinically important factors that were outside the scope of the initial annotation. In summary, Run 1 showed a clear upward trend in performance (macro F1 from 0.59 to 0.62) attributable to the GPT-augmented data, and it set the stage for handling a broader set of ICF codes.

6.3 Addressing Class Imbalance with Downsampling (Run 2)

One straightforward strategy to handle the severe class imbalance (over 85% of sentences are "None") was to down-sample the majority class. In Run 2, we experimented with keeping only 5% of "None" examples in training, aiming to strengthen the relative proportion of meaningful ICF category instances. The intuition was that this would force the model to focus on minority classes rather than defaulting to predicting "None". As expected, this did increase the model's sensitivity: the macro recall jumped (the model found more of the rare category instances). However, this came at a steep cost to precision. The model began over-predicting ICF categories in sentences that truly had no relevant content, causing many false positives. Consequently, macro precision dropped from 0.77 to 0.64, and the overall macro F1 fell slightly to 0.60 (down from 0.62 in Run 1). The micro F1 similarly declined. In other words, the modest gains in detecting minority classes were offset by a flood of false alarms on the majority class. This outcome is consistent with general NLP findings that naive resampling can hurt more than help, especially when classes have overlapping features (Henning et al., 2023). In our case, many "None" sentences contain generic clinical text that shares vocabulary with the ICF categories (e.g. mention of walking or mood without actual impairment), so removing most "None" examples made the model less calibrated and more likely to mislabel such sentences as positive.

Given the drop in overall F1 score, we abandoned further downsampling. Run 2 confirmed that class imbalance is challenging to address with sampling alone, matching literature that emphasizes the difficulty of finding one-size-fits-all solutions for imbalance in multi-label settings (Henning et al., 2023). We therefore reverted to using the full training set (with all "None" instances) in subsequent runs, and instead focused on data augmentation and model tuning, aiming to improve minority class performance.

6.4 Expanded Label Inventory and Augmented Data (Run 3)

Run 3 represents our major experimental milestone, where we fully embrace the 18-category ICF system and substantially increase the training data. For this run, we created an augmented training set (D) by combining: the previous combined data (manual + GPT from Run 1); and additional manually annotated sentences and newly GPT-labeled notes from other sources (including VUMC data) to further increase the cases of underrepresented categories. The model was trained on this enriched 18-category dataset for 1 epoch with an initial learning rate of 4e-5 (earlier experiments indicated one epoch was optimal to prevent overfitting).

6.4.1 10-Category Performance

On the core 10 categories, Run 3 achieved a macro F1 of 0.67, a sizeable improvement over Run 1's 0.62. Impressively, recall climbed to 0.63 while precision held around 0.72, indicating the model was now catching many more of the relevant sentences with only a minor dip in precision compared to the baseline. This reflects the effect of both more training examples and wider label coverage: the model had seen a greater variety of ways each functional issue can be described, improving generalization. The micro F1 (10-category) kept high in 89%, showing that even the frequently occurring classes benefited from augmentation. For example, ETN (Eating) and MBW (Weight maintenance functions), two categories that often overlap in content, both saw F1 gains and fewer confusions with each other. Earlier runs often confused these more seriously, but with more data the model learned to more stably assign both labels.

Noticeably, comparing to the original 10-category manual label system, by adding ai-assisted annotations, recall of some of the weak categories improves. Attention (ATT, B140) sees recall rise from 0.23 (manual-only 10-cat model) to 0.52 with the augmented 18-cat model. F1 moves from 0.37 to 0.67 with precision unchanged. On the confusion matrix, correct hits increase (19 to 43 on the diagonal. Work & employment (BER, D840-D859) has recall increases from 0.42 to 0.48, and F1 from 0.51 to 0.55, true positive counts grow evidently (140 to 159), showing fewer misses to None. Exercise tolerance (INS, B455), the hardest of the three improves as well: recall 0.20 to 0.34, F1 0.29 to 0.36. The diagonal increases (75 to 131) and true INS to None drops (200 to 70), which caps recall compared with ATT and BER, suggesting new GPT-synthesized sentences could cover more instances involve capability to endure physical exertion that the model had not seen before; but residual confusion with Energy level (B1300) persists.

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW
precision	0.84	0.91	0.65	0.81	0.61	0.61	0.39	0.78
recall	0.63	0.52	0.48	0.67	0.67	0.78	0.34	0.68
f1-score	0.72	0.67	0.55	0.73	0.64	0.69	0.36	0.73
support	1063.00	82.00	331.00	413.00	640.00	362.00	382.00	341.00

	STM	none	micro avg	macro avg	weighted avg	samples avg
precision	0.63	0.98	0.94	0.72	0.95	0.85
recall	0.69	0.87	0.84	0.63	0.84	0.84
f1-score	0.66	0.92	0.89	0.67	0.89	0.84
support	348.00	33769.00	37731.00	37731.00	37731.00	37731.00

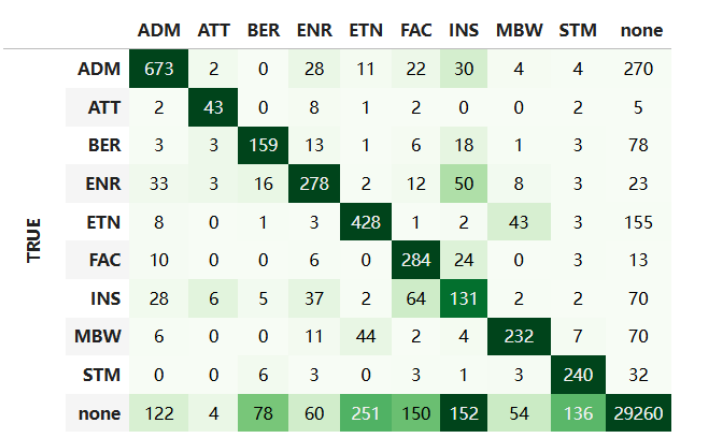


Figure 6.2: Run 3 augmented pool (10-category results)

In short, Run 3 demonstrated that strategic data augmentation has potential to improve minority class detection, an approach echoed by other multi-label text classification studies that use data synthesis or augmentation to balance long-tailed label distributions (Hu et al., 2025; Zhao et al., 2024).

6.4.2 Full 18-Category Performance

After Run 3, we obtained an updated gold standard on the test set that includes all 18 categories (the new labels were manually validated by clinical experts for the test notes). This allowed us to evaluate the model's performance on the entire label set. The Run 3 model achieved a macro F1 of 0.68 across all 18 categories (macro precision 0.67, recall 0.73). The fact that this is on a level with the 10-category F1 (0.67) indicates that the model handled the new categories reasonably well without sacrificing overall performance. In other words, expanding the label inventory from 10 to 18 did not cause a drop in aggregate accuracy, which is a reassuring result for our thesis objective of covering more ICF aspects. Some of the new categories had particularly strong results: B280 (Sensation of pain) stood out with high precision and recall, suggesting that pain-related sentences (e.g. reports of chronic pain or pain interfering with activities) were often detected correctly. This is likely because pain descriptions are usually explicit (mentions of "pain" or specific pain behaviors) and we had ample training examples synthesized for B280. The model's ability to pick up pain is a clinically significant gain, given

that pain is a major cause of disability and often under-documented (WHO, 2001). Another new category, B134 (Sleep functions), also showed good performance, suggesting the model might have learned to identify references to sleep quality or insomnia which were common in notes (and these were previously just ignored by the 10-category model).

Table 6.2: Run 1 to Run 3 (18-category): Changes of New Categories' Results (Part)

Frequency	Category	F1 Change	Precision Change	Recall Change
High	B280 Pain	+0.03	+0.05	+0.01
High	D760 Family	+0.02	+0.02 (still low)	0
Mid	B164 Cognition	+0.07	+0.10	+0.04
Mid	D410 Position Change	+0.13	+0.20	0
Mid	D240 Stress	+0.20	+0.23	-0.01

Other new categories were more challenging. D760 (Family relationships) had one of the lowest precisions among all labels. While the recall for D760 was moderate (suggesting the model might have found many sentences where family or support network was mentioned), it frequently over-predicted this category. As a result, D760's precision lagged. This implicates a limitation that the model might lack deeper context to distinguish mention of a factor from actual functional impact. Improving this may require more nuanced training data or incorporating context beyond single sentences. However, the inclusion of D760 is still valuable; even if some false alarms occur, the system can flag notes for potential social support issues that clinicians might review. Importantly, many of those were previously "invisible" to the automated system. The ICF framework emphasizes environmental factors like family support as crucial to functioning, so having any capacity to detect related content is an improvement in the model's clinical utility (WHO, 2001).

	B1300	B140	B152	B440	B455	B530	D450	D550	\
precision	0.83	0.89	0.66	0.85	0.40	0.78	0.63	0.61	
recall	0.63	0.66	0.70	0.65	0.34	0.69	0.80	0.65	
f1-score	0.72	0.76	0.68	0.73	0.36	0.73	0.71	0.63	
support	448.00	64.00	362.00	1060.00	398.00	338.00	368.00	665.00	

	D840-D859	B280	B134	D760	B164	D465	D410	B230	\
precision	0.67	0.83	0.61	0.54	0.73	0.55	0.59	0.33	
recall	0.48	0.93	0.91	0.92	0.75	0.79	0.68	0.92	
f1-score	0.56	0.88	0.73	0.68	0.74	0.65	0.63	0.49	
support	342.00	847.00	183.00	472.00	303.00	162.00	390.00	37.00	

	D240	None	micro avg	macro avg	weighted avg	samples avg
precision	0.53	0.97	0.92	0.67	0.92	0.89
recall	0.77	0.92	0.88	0.73	0.88	0.89
f1-score	0.63	0.95	0.90	0.68	0.90	0.89
support	212.00	31692.00	38343.00	38343.00	38343.00	38343.00

Figure 6.3: Run 3 augmented pool (18-category results)

Similarly, D240 (Handling stress and other psychological demands) and B164 (Higher-level cognitive functions) showed some confusion with each other and with B152 (emotional functions). These three categories all relate to cognitive or psychological aspects, and a given sentence can invoke multiple. If the gold standard only tagged one or two of these, the model might predict a different one, appearing as an error. This suggests some blurry boundary between mental function categories, which is understandable given their conceptual similarity. Despite these confusions, the model dramatically expanded coverage of the cognitive domain, i.e., B164 (new in the extended set) had decent performance. Overall, while category-level performance varied, the trend was clear, the model in Run 3 successfully learned the new categories with no collapse in performance on the original ones.

Table 6.3: Run 0 to Run 3: Changes of Old Categories' Results

Category	F1 Change	Precision Change	Recall Change
B1300 Energy level	+0.08	+0.01	+0.12
B140 Attention functions	+0.31	-0.01	+0.36
B152 Emotional functions	+0.01	-0.13	+0.13
B440 Respiration functions	+0.12	-0.05	+0.19
B455 Exercise tolerance functions	+0.08	-0.18	+0.15
B530 Weight maintenance functions	+0.03	-0.03	+0.08
B530 Weight maintenance functions	+0.03	-0.03	+0.08
D450 Walking	+0.09	-0.01	+0.19
D550 Eating	+0.09	-0.03	+0.08
D840-D859 Work and employment	+0.04	-0.01	+0.06

After updates of the gold labels, relative to the initial manual-10 model, the model finetuned with 18-category augmented data contributes to significant improvements for the original domains, but the way this happens is consistent: recall rises broadly while precision softens slightly. In other words, augmentation mainly converts "near misses" that previously fell into None into true positives at a modest cost in false alarms. This suggests that the augmented data turns a precision-dependent, risk-oriented classifier into a more balanced one, fewer misses to None, slightly broader decision boundaries, and consistent F1 improvements. This indicates that multi-label training with a more complete label set could be mutually beneficial. This is possibly due to shared context, for example, mentioning pain (B280) often co-occurs with noting limitations in exercise tolerance (B455) or sleep (B134), so the model's representation of those contexts became richer by learning all labels jointly.

6.5 Fine-Tuning Variations and Final Model (Run 4)

After identifying Run 3's configuration (MedRoBERTa, 1 epoch, LR 4e-5 on augmented data) as optimal, we conducted a few fine-tuning variations to see if performance could be further imporved. These are summarized as Run 4 in the table, though internally we tried multiple tweaks: increasing training epochs, lowering the learning rate, and using a two stage fine-tuning (first on 10-cat data, then on 18-cat data). The rationale was that additional epochs might allow the model to learn harder patterns or that a model already familiar with the 10 original categories could adapt better to the expanded set.

The outcome, however, was that more training was not necessarily better. Running 5 epochs on the same data possibly led to overfitting: macro F1 dropped to 0.64 on the 10 categories (vs 0.67 at 1 epoch). The precision actually ticked up slightly (to 0.74) but recall fell sharply (0.58), indicating the model might have become too conservative again, memorizing frequent patterns and missing less common ones. We suspect that by epoch 5, the model had over-learned the dominant "None" and high-frequency signals despite early stopping attempts. This aligns with our earlier observation that a single epoch on a relatively large dataset was enough to converge; beyond that, the model begins to specialize on the training distribution at the expense of generalization. We also tried a smaller learning rate (2e-5) for 1 epoch to see if a gentler update would help, but the F1 (0.66) was essentially similar but not surpassing the original run 3. Another variant was initializing the model from the Run 0 (10-cat) checkpoint and then training on the 18-cat data (to give it a head start on those original categories). This too resulted in no notable gain (macro F1 0.64). In sum, none of these fine-tuning variations surpassed the simpler one-epoch training from scratch on the full augmented set.

These findings suggest a practical assumption: more data and coverage could help (Run 3), but more epochs might not. It appears the augmented dataset was large and diverse enough that one epoch was sufficient to extract the needed signal with a high learning rate, whereas longer training started to over-prioritize common labels. Similar behavior has been reported in multi-label literature, Chen et al. (2024) note that fine-tuning even a fraction of available data can outperform extensive prompting of an LLM. Our scenario mirrors that: the best model remained the one from Run 3, which we will treat as the final fine-tuned model for subsequent analysis. The consistency of results across these variations also adds confidence that the performance might plateau around macro F1 0.67-0.68 for this architecture and data; reaching significantly higher would likely require even more (and higher quality) training data.

6.6 Comparison with GPT-4o (Zero/Few-Shot Classification)

A key question for this project was how a fine-tuned specialist model compares to a large generalist model (GPT-4o) on the task of sentence-level ICF coding. To explore this, we evaluated GPT-4o in a few-shot classification setting: we provided GPT-4o with the definitions of the 18 ICF categories and 2 example sentences for each category (drawn from our data), then prompted it to label test sentences. This setup uses GPT-4o as an out-of-the-box classifier without task-specific fine-tuning. The results showed that our fine-tuned MedRoBERTa model markedly outperformed GPT-4o on this task. GPT-4o's overall macro F1 was 0.58 for the 18 categories (precision 0.59, recall 0.60), which is substantially lower than the 0.68 achieved by the fine-tuned model. In fact, GPT-4o's performance was closer to the level of our early Run 0/1 models. The fine-tuned model had a 10 point advantage in F1, suggesting that domain-specific training still confers a big benefit. This finding aligns with reports that fine-tuned smaller models can rival or exceed large LLMs for specialized classification tasks (Chen et al., 2024), especially when the LLM is not explicitly tuned to the task.

Examining by category, GPT-4o did reasonably well on very frequent, overt categories (it was good at identifying B280 Sensations of pain and B134 Sleep functions, likely due to clear keywords). However, GPT-4o struggled with subtle categories and often missed multi-label situations. For example, in a sentence describing pain leading to sleep problems, GPT-4 might only output "pain" but miss the sleep dysfunction, whereas our fine-tuned model would be more likely to correctly assign both B280 and B134 labels. GPT-4o also had difficulty with the more abstract categories like B164 (higher-level cognition), the low racall (0.37) suggests that the model might ignored them if not explicitly mentioned. These errors limit recalls for those abstract labels.

	B1300	B140	B152	B440	B455	B530	D450	D550	\
precision	0.63	0.60	0.55	0.54	0.26	0.47	0.60	0.40	
recall	0.67	0.73	0.42	0.60	0.36	0.73	0.74	0.46	
f1-score	0.65	0.66	0.48	0.57	0.30	0.57	0.66	0.43	
support	448.00	64.00	362.00	1060.00	398.00	338.00	368.00	665.00	

	D840-D859	B280	B134	D760	B164	D465	D410	B230	\
precision	0.60	0.79	0.63	0.61	0.77	0.57	0.67	0.40	
recall	0.58	0.78	0.81	0.60	0.37	0.48	0.40	0.76	
f1-score	0.59	0.78	0.71	0.60	0.50	0.52	0.50	0.52	
support	342.00	847.00	183.00	472.00	303.00	162.00	390.00	37.00	

	D240	None	micro avg	macro avg	weighted avg	samples avg
precision	0.56	0.95	0.88	0.59	0.88	0.88
recall	0.40	0.92	0.86	0.60	0.86	0.87
f1-score	0.47	0.93	0.87	0.58	0.87	0.87
support	212.00	31692.00	38343.00	38343.00	38343.00	38343.00

Figure 6.4: Few-shot GPT4o (18-category results)

It's important to note that GPT-4o's few-shot capability is impressive given it had no

explicit training on our dataset, achieving 0.58 macro F1 out-of-the-box on a nuanced multi-label task shows the power of LLMs' latent knowledge. With more prompt engineering or chain-of-thought prompting, GPT-4o might improve. However, our results indicate that for fine-grained clinical text classification, a fine-tuned model is still superior. The fine-tuned model not only learned the general patterns but also the specific quirks and context of our data (e.g. how Dutch clinical notes phrase certain impairments), which GPT-4o could not fully replicate with a generic prompt. This is echoed by recent studies in biomedicine where smaller pre-trained models fine-tuned on target data outperformed GPT-3.5/4 on tasks like ICD coding and reasoning with clinical text (Chen et al., 2024). In our case, the fine-tuned model provides more consistent and reliable outputs, which is important for practical deployment in a clinical setting (where one would prefer a model that behaves predictably, even if GPT-4o occasionally shows flashes of brilliance).

The comparative analysis favors our fine-tuned approach, which achieved higher F1 across the board and especially excelled in multi-label scenarios, whereas GPT-4o, without task-specific tuning, left many labels on the table. This demonstrates the contribution of our work in creating a customised model for ICF classification. That said, GPT remains a strong baseline and could potentially be used to generate additional training data or as an ensemble component in the future.

6.7 Note-Level Check (OR-Aggregation over Sentences)

Because clinical use often concerns the presence of a function anywhere in a note, we report an aggregated note-level view (a note is positive for class x if any sentence in that note is positive). This section is a sanity check; detailed errors analysis remain at sentence level.

6.7.1 Aggregation Rule

For each category, a note is marked positive if any sentence in that note is positive for that category (logical OR across sentences). With probabilistic outputs, this is equivalent to taking the maximum sentence probability per note. OR-aggregation is not a majority vote; one positive sentence is sufficient. In practice this tends to raise recall and lower precision. We report note-level metrics as a sanity check, detailed errors remain at sentence level.

6.7.2 Initial Combined Model (Note Level)

On 2,969 notes, the macro precision/recall/F1 = 0.71 / 0.76 / 0.69. Compared with the sentence-level figures for the same model, recall rises (as expected when multiple sentences give multiple chances), while precision dips slightly, producing a comparable F1. Per-class patterns mirror the sentence level but with notable improvements: Respiration and Eating reach F1 0.84 and 0.81, Emotional 0.75, and Weight maintenance 0.87. Some new categories still lag in precision despite higher recall. For Family (D760), precision is about 0.67 with recall around 0.96 (F1 near 0.79), indicating frequent over-assignment on administrative mentions even when aggregation is used. Handling stress (D240) remains precision limited (precision 0.41, recall 0.88, F1 0.56); aggregation improves recall but cannot fix boundary errors where sentences lack explicit coping actions.

```
=== NOTE-LEVEL CLASSIFICATION REPORT (18 cats) ===
              B1300   B140    B152    B440    B455    B530    D450    D550  \
precision      0.87   0.94    0.82    0.89    0.71    0.95    0.65    0.87
recall         0.66   0.61    0.69    0.80    0.23    0.81    0.82    0.75
f1-score       0.75   0.74    0.75    0.84    0.35    0.87    0.72    0.81
support      339.00  51.00  232.00  397.00  276.00  212.00  184.00  309.00

             D840-D859    B280    B134    D760    B164    D465    D410    B230  \
precision         0.84    0.86    0.50    0.67    0.67    0.33    0.49    0.31
recall            0.44    0.95    0.96    0.96    0.81    0.91    0.76    0.82
f1-score          0.57    0.90    0.66    0.79    0.73    0.49    0.59    0.45
support         268.00  457.00  132.00  274.00  143.00  111.00  201.00  22.00

               D240     None  micro avg  macro avg  weighted avg  samples avg
precision      0.41     0.92       0.76       0.71          0.81         0.80
recall         0.88     0.90       0.79       0.76          0.79         0.81
f1-score       0.56     0.91       0.77       0.69          0.78         0.79
support      132.00  1602.00    5342.00    5342.00       5342.00      5342.00
```

Figure 6.5: Note-level initial combined pool

6.7.3 Final Augmented Model (Note Level)

On the same set of notes, macro precision/recall/F1 increases to 0.74 / 0.81 / 0.76. Gains concentrate in categories that augmentation explicitly enriched. Among the new codes, Sleep (B134) improves from F1 0.66 to 0.80 driven by precision 0.50 to 0.70; Changing body position (D410) from 0.59 to 0.73 with precision 0.49 to 0.69; Using equipment (D465) from 0.49 to 0.71, precision rise 0.33 to 0.64 with a small recall reduction; and Handling stress (D240) from 0.56 to 0.74 as precision lifts 0.41 to 0.64 while recall stays high (0.88 to 0.89). Legacy hard categories also benefit: Work & employment (D840-D859) moves 0.57 to 0.65 and Exercise tolerance (B455) 0.35 to 0.48; Respiration (B440) consolidates at 0.87 F1, and Eating (D550) rises 0.81 to 0.83. A minor exception is Walking (D450), essentially unchanged (0.72 to 0.71). The None label remains stable at F1 0.91; aggregation continues to favour recall for frequent functions while leaving boundary-precision issues visible.

```
=== NOTE-LEVEL CLASSIFICATION REPORT (18 cats) ===
              B1300   B140    B152    B440    B455    B530    D450    D550  \
precision      0.86   0.93    0.75    0.85    0.48    0.95    0.59    0.84
recall         0.70   0.76    0.76    0.88    0.47    0.82    0.88    0.82
f1-score       0.77   0.84    0.76    0.87    0.48    0.88    0.71    0.83
support      339.00  51.00  232.00  397.00  276.00  212.00  184.00  309.00

              D840-D859    B280    B134    D760    B164    D465    D410   B230  \
precision          0.77    0.89    0.70    0.69    0.78    0.64    0.69   0.32
recall             0.56    0.96    0.94    0.97    0.80    0.79    0.78   0.91
f1-score           0.65    0.92    0.80    0.81    0.79    0.71    0.73   0.48
support          268.00  457.00  132.00  274.00  143.00  111.00  201.00  22.00

               D240     None  micro avg  macro avg  weighted avg  samples avg
precision      0.64     0.94       0.80       0.74          0.82         0.82
recall         0.89     0.89       0.83       0.81          0.83         0.83
f1-score       0.74     0.91       0.81       0.76          0.82         0.82
support      132.00  1602.00    5342.00    5342.00       5342.00      5342.00
```

Figure 6.6: Note-level final augmented pool

6.7.4 GPT-4o (Note Level) and Comparison

On the same note set, GPT-4o attains macro precision/recall/F1 0.72 / 0.70 / 0.69. The F1 is comparable to the initial combined MedRoBERTa run (0.71 / 0.79 / 0.69) but precision and recall are lower, showing performance clearly below the final augmented model (0.74 / 0.81 / 0.76). The largest gaps align with the categories that benefited most from augmentation: Handling stress (D240) and Using equipment (D465) show markedly lower F1 for GPT-4o than for the augmented model; Changing body position (D410) is also weaker. By contrast, highly lexical categories such as Pain (B280) and Sleep (B134) are closer across systems at note level. The GPT-4o confusion matrix continues to show broader activations into common functions within notes that lack definite evidence, resulting in lower precision in the same areas where the sentence-level analysis found over prediction.

```
=== NOTE-LEVEL CLASSIFICATION REPORT (18 cats) ===
              B1300   B140    B152    B440    B455    B530    D450    D550  \
precision      0.75   0.76    0.74    0.74    0.40    0.70    0.58    0.68
recall         0.78   0.82    0.45    0.80    0.48    0.89    0.83    0.69
f1-score       0.76   0.79    0.56    0.77    0.43    0.78    0.68    0.69
support      339.00  51.00  232.00  397.00  276.00  212.00  184.00  309.00

              D840-D859    B280    B134    D760    B164    D465    D410   B230  \
precision          0.74    0.91    0.76    0.76    0.87    0.78    0.73   0.43
recall             0.69    0.84    0.86    0.73    0.48    0.59    0.45   0.82
f1-score           0.71    0.87    0.81    0.75    0.62    0.67    0.56   0.56
support          268.00  457.00  132.00  274.00  143.00  111.00  201.00  22.00

               D240     None  micro avg  macro avg  weighted avg  samples avg
precision      0.66     0.92       0.77       0.72          0.78         0.78
recall         0.48     0.84       0.74       0.70          0.74         0.76
f1-score       0.55     0.88       0.75       0.69          0.75         0.76
support      132.00  1602.00    5342.00    5342.00       5342.00      5342.00
```

Figure 6.7: Note-level GPT-4o's predictions

Sentence level remains our primary evaluation (what the model actually predicts and where it fails), while this note-level check shows that the augmented model's advantages translate to the clinical unit of analysis. Functions repeatedly mentioned within a note (Respiration, Eating, Emotional) present notably higher note-level F1; categories with fuzzy boundaries

(Family, Handling stress) improve but remain precision constrained, emphasizing the need for boundary-aware calibration rather than wholesale data growth. The relative ranking of systems is unchanged by aggregation: augmented MedRoBERTa > initial combined baseline > GPT-4o.

Chapter 7

Error Analysis

7.1 Flipped Errors: Old None to New Categories

A major source of apparent error in the 10-category evaluation stemmed from the absence of labels for clinically meaningful content. Sentences that should have belonged to the newly added categories were annotated as None, so model predictions for them were treated as false positives. After clinical validation with the 18-category gold, however, these predictions were recognized as correct. These "flipped errors" show that part of the model's apparent over-prediction was actually sensitivity to functional content excluded from the old scheme. This is the classic incomplete-label problem in multi-label classification, unlabeled positives are counted as errors, which depresses apparent precision until the label space is completed, or missing labels are modeled explicitly (Chen et al., 2024; Bhowmick et al., 2008)

Table 7.1: Flipped Error Examples and Patterns

Category	**Text**	**Pattern (old gold | pred | new gold)**
B280 Pain	veel hoofdpijn afgelopen dagen.	old=['None'] | pred=['B280'] | new=['B280']
B280 Pain	pijn in de vingers en kniee.	old=['None'] | pred=['B280'] | new=['B280']
D410 Position Change	op luchtkussen met rotatie in romp (3 x 10 hh)	old=['None'] | pred=['D410'] | new=['D410']
D760 Family	herstel op activiteit oppassen op de kleinzoon 1x per week 8 uur van PSK 10 naar PSK <	old=['None'] | pred=['D760'] | new=['D760']
B164 Cognition	heeft moeite om dingen te begrijpen en te luisteren	old=['None'] | pred=['B164'] | new=['B164']
B164 Cognition	belemmert mij bij het uitvoeren van bepaalde taken	old=['None'] | pred=['B164'] | new=['B164']

B280 Sensations of pain

B280 Sensations of pain is defined as "the sensation of unpleasant feeling indicating potential or actual damage to some body structure, including generalized or localized pain" (World Health Organization [WHO], 2001). Several sentences explicitly describing pain were first annotated as None but re-labeled as Pain in the expanded gold. For example, "veel hoofdpijn afgelopen dagen." ("a lot of headache in the past days.") and "pijn in de vingers en kniee." ("pain in the fingers and knees.") were counted as errors in the 10-class setting. The model predicted Pain in both cases, which became true positives under the revised gold. These examples show the model was correctly identifying straightforward symptom mentions that the restricted label set could not capture.

D410 Changing basic body position

D410 Changing basic body position is defined as "getting into and out of a body position and moving from one location to another, such as rolling, sitting, standing, or bending" (WHO, 2001). Sentences describing posture or trunk movement were misclassified as None before validation. For example, "op luchtkussen met rotatie in romp (3 x 10 hh)" ("on air cushion with trunk rotation (3 x 10 reps)") was originally marked None. The model predicted D410 Changing basic body position, which matched the updated gold labels. Such cases illustrate how motor function exercises were systematically overlooked in the old scheme, while the model nevertheless detected their functional relevance.

D760 Family relationships

D760 Family relationships is defined as "creating and maintaining kinship relationships, such as with members of the nuclear family, extended family, or legal guardians" (WHO, 2001). One representative example is "herstel op activiteit oppassen op de kleinzoon 1x per week 8 uur van PSK 10 naar PSK <" ("recovery activity of babysitting the grandson once a week for 8 hours from PSK 10 to PSK <"). The key cues are the role ("oppassen op de kleinzoon"), frequency and duration ("1x per week, 8 uur"), which together denote ongoing kinship care responsibilities. This sentence was originally labeled None, but the model predicted Family relationships, which was later confirmed by clinical validation. Such cases point out how relational responsibilities were systematically overlooked in the 10-category scheme but identified correctly by the model.

B164 Higher-level cognitive functions

B164 Higher-level cognitive functions are "specific mental functions especially dependent on the frontal lobes of the brain, including decision-making, planning, and mental flexibility" (WHO, 2001). Cognitive difficulties appeared as flipped errors in several sentences. For instance, "heeft moeite om dingen te begrijpen en te luisteren" ("has trouble understanding and listening") was originally tagged None. The model predicted Cognition, which was confirmed in the updated gold. Another example, "belemmert mij bij het uitvoeren van bepaalde taken"

("hinders me in carrying out certain tasks"), likewise illustrates subtle cognitive dysfunction that the model identified but the old annotation missed.

Taken together, flipped errors reveal that many supposed false positives were not random noise but genuine detections of pain, body position, family, and cognitive impairments. Pain (B280) flips are mostly lexically sufficient terms; Family (D760) flips require pragmatic inference that the kinship mention is functionally relevant, therefore performs weaker in precision once the category exists. Expanding the label set corrected these systematic misclassifications and demonstrates how limited taxonomies can underestimate model capability in clinically relevant domains.

7.2 False Negatives: Missed Categories

In this study, false negatives are sentences whose gold annotation contains one or more ICF categories while the model predicts None. These errors depress recall and are especially informative about the kinds of functional information that fail to trigger a category decision in brief, telegraphic clinical prose. Below I analyse three categories with low recall: B1300 Energy level, D840-D859 Work and employment, and B455 Exercise tolerance functions, using representative sentences from the test set. In all cited cases, the prediction vector was exactly None while the gold vector contained the respective category.

Table 7.2: False Negative Examples and Cues

Category	Recall	Text	Cue
B1300 Energy level	0.63	voor verwijzing covid vermoeidheidsklachten	vermoeidheidsklachten
B1300 Energy level	0.63	aan alles dat ze fitter word P : HUR + conditie rondje	fitter
B1300 Energy level	0.63	rijst eiwit en erwten eiwit combineren met elkaar , verder aandacht voor verdelen van energie , c over 2-3 weken via whatsapp	verdelen van energie
D840-D859 Work	0.48	van aanpak / uitgevoerde behandeling S : gaat heel goed . werken	werken
D840-D859 Work	0.48	dag gewerkt : goed	gewerkt
D840-D859 Work	0.48	uurtje gewerkt . geen	gewerkt
B455 Exercise	0.34	activiteiten (Uzelf aankleden , wassen) Helemaal	
B455 Exercise	0.34	merkt zelf ook dat de oefeningen beter gaan .	oefeningen
B455 Exercise	0.34	is door covid verzwakt en haar smaak en reuk ontbreken .	verzwakt

B1300 Energy level

In the ICF, Energy level covers stamina, vigour and fatigability. Several genuine energy-related statements were missed by the model despite containing clear lexical cues. The fragment "voor verwijzing covid vermoeidheidsklachten" explicitly names vermoeidheidsklachten (fatigue complaints) in a referral context, yet the model returned None. Likewise, "aan alles dat ze fitter word P : HUR + conditie rondje" documents improved fitness and conditioning, signalling the same functional axis from the opposite direction (recovery rather than impairment); the heavy use of abbreviations (e.g. HUR) and a list-like structure likely blunted the classifier's evidence. A third sentence, "rijst eiwit en erwten eiwit combineren met elkaar , verder aandacht voor verdelen van energie , c over 2-3 weken via whatsapp", centres on pacing (verdelen van energie), which is a solid energy management cue, yet the nutrition frame and telegraphic style again appear to have pushed the decision boundary toward None. Across these examples, the common pattern is that energy-related content is present but packaged in minimal, note-style constructions with abbreviations and topic shifts; the model seems to rely on fuller syntactic or contextual composition before committing to B1300.

D840-D859 Work and employment

Work participation in these notes is typically recorded in highly compressed status updates rather than full sentences, and those entries were systematically missed. In "van aanpak / uitgevoerde behandeling S : gaat heel goed . werken" the final token werken sits as a bare fragment within a progress line; without explicit arguments (employer, hours, role) the model treats it as meta-text rather than an ICF relevant activity. The entry "dag gewerkt : goed" is an even starker diary-style report that directly evidences engagement in work but lacks grammatical context; it, too, was predicted as None. Moreover, "uurtje gewerkt . geen" conveys an hour of work, then trails into an elliptical "geen" (likely beginning a negation or separate item), a pattern that may increase uncertainty and push the model to abstain. The error profile suggests that the style of sentence segmentation in the test data could affect the results, classifier sometimes underweights schematic mentions of work when they are not embedded in a narrative sentence, even though by ICF definition they constitute direct evidence of participation in employment.

B455 Exercise tolerance functions

Exercise tolerance pertains to the capacity to sustain physical exertion; here, the model often ignored exertion cues unless they were quantified or elaborated. For example, "activiteiten (Uzelf aankleden , wassen) Helemaal", this sentence contains only ADL or self-care cues "aankleden, wassen" (dressing, washing), and there is no endurance or exertion marker (e.g. lopen, traplopen). The telegraphic style ("activiteiten... Helemaal") also lacks difficulty terms, making a sentence-level model unlikely to infer exercise tolerance without broader note context. In "merkt zelf ook dat de oefeningen beter gaan .", the patient reported improvement in exercises implies an endurance trajectory, but the absence of intensity or duration markers appears to have weakened the category signal. Similarly, "is door covid verzwakt en haar smaak en reuk ontbreken ." mentions verzwakt (weakened) in a post-COVID context; the co-occurrence with taste and smell may have diffused the model's attention away from exertional capacity which was not expressed through a complete sentence. Together these cases indicate that, unless exertion is quantified (load, duration) or explicitly tied to endurance terms, the model defaults to None even when an endurance related change is evident.

In sum, false negatives for Energy, Work, and Exercise tolerance usually due to short, incomplete and context lacking expressions that use abbreviations, list syntax, or multi-item symptom framing. Even when key lexical cues are present (vermoeidheid, gewerkt, niet inspannen), the absence of sentence-level structure or explicit quantification leads the model to errors. This points to two practical mitigations for future work: further adjust sentence segmentation design for part of the test data, expanding training exposure to note style and abbreviated structures of functional content, and enriching the modelling context (e.g. with

section headers or neighbouring sentences) so that sparse cues are interpreted as functionally meaningful rather than dismissed as None.

7.3 Systematic Confusions between Related Categories

This section analyses three most significant confusion pairs shown by the confusion matrix: B1300 Energy level - B455 Exercise tolerance functions, B152 Emotional functions - D240 Handling stress and other psychological demands, D450 Walking - D465 Moving around using equipment. Because in confusion cases, many example sentences involve more than one category in gold labels and predictions, I organise the discussion as the table shows:

Table 7.3: A & B Confusion Patterns

Pattern	**Meaning**
clean_single	a true swap (gold has only A, prediction has only B)
tp_plus_extra_single	gold has only A but the model predicts A plus extra labels (often B)
gold_multi_simple	gold has A with other labels, but the model predicts only B
gold_multi_no_tp	gold has A with others, and the model predicts B with others (but not A)
gold_multi_tp_plus_extra	both gold and prediction include A and B (sometimes with additional labels)

B1300 Energy level - B455 Exercise tolerance functions

Energy level refers to stamina and fatigability; Exercise tolerance concerns capacity under exertion. A genuine swap appears in clean_single: "zich fit O covid revalidatie P HUR , fietsen 80–120 watt , balansoefeningen" was gold labeled as Energy level, but the model predicted Exercise tolerance. The structured exercise context (cycling with wattage) likely pulled the model toward exertion than focusing on vigour and stamina conveyed by the information. In tp_plus_extra_single, the model tends to over assign: "ben over 2 weken's avonds niet meer moe na een dag wandelen en vissen B:9" is gold Energy level, and the model predicts Energy level + Exercise tolerance (and Walking). Here stamina and exertion are mentioned together, so the model sees them as parallel information than a package that simply demonstrates stamina. In gold_multi_simple, the model reframes a mixed gold as exertion: "aan de trap ben ik wel moe th HUR Fietsen in de heuvel vorm (niveau 5) 7" has gold Energy level + Walking, but the model outputs Exercise tolerance only, showing that the model clings on the exercise content, overshadowing the stair-walking and fatigue cues that drive the gold labels. There's also lexical ambiguity, trap can mean "stairs" (gold wants D450) or "to pedal", which co-occurs with fietsen and could result in bias toward Exercise tolerance. Overall, this pair shows that

when fatigue and exercise context co-occur in one sentence, the model often assigns both labels or shifts toward exertion. This is less a failure to detect the gold concept and more a tendency to read stamina and endurance as a package when exercise details are present.

B152 Emotional functions - D240 Handling stress and other psychological demands
Emotional functions are about mood and affect; Handling stress is about coping with demands. In clean_single, "het sporten weer heel prettig en vind dat de begeleiding erg professioneel is hier P : kracht" is gold Emotional functions (positive affect) but the model predicts Handling stress, likely because the sentence sits in a structured activity context and the model reads it as coping rather than mood. In tp_plus_extra_single, the model often keeps the gold and adds coping: "Anamnese Vond het spannend om naar huis te gaan ." is gold Emotional functions, while the model predicts Emotional functions + Handling stress. The wording ("spannend" before going home) naturally evokes both emotion and coping of psychological demands. In gold_multi_simple, affect appears with activity and the model collapses to stress coping: "van aanpak / uitgevoerde behandeling S : gaat lekker . is druk aan het sporten en vind het leuk P" has gold Emotional functions + Exercise tolerance, but the model predicts Handling stress only, possibly because the mention of "druk" triggers implication of possible stress management in busy situation. The pattern is consistent, once psychological language is present, the model tends to double tag affect with stress coping or to prefer the stress coping label when activity or responsibility is also mentioned.

D450 Walking - D465 Moving around using equipment
Walking is an unaided movement; Moving around using equipment is a mobility with a device. A true swap is rare but clear in clean_single: "Volgens dagprogramma en looptraining met de fysio." is gold Walking, yet the model predicts Using equipment, probably because therapy contexts often imply device use. In tp_plus_extra_single, the model frequently assigns both when a device is hinted: "Morgen FAC 4 evt zonder rollator met steun op infuuspaal ." is gold Walking, while the model predicts Walking + Using equipment. Multi-label gold strengthens this tendency. In gold_multi_simple, the model generalises assisted mobility to walking alone: "Dhr is gemobiliseerd morgens na terugkomst van de pacu en in de middag heeft dhr gefietst ." has gold Using equipment + Changing basic body position, but the model predicts Walking only. The phrase "heeft... gefietst" is a strong lexical cue for Using equipment (stationary bike), but the model fails to recognize it, suggesting the model's weak sensitivity of bicycle-related terms or weak coverage for such terms in the D465 class in the train data. In gold_multi_no_tp, it drops the device label when both themes are present: "Zo wordt zij in rolstoel gereden voor grote stukken lopen (wordt na een stuk lopen wel moe) ." has gold Energy level + Exercise tolerance + Using equipment, while the model predicts Exercise tolerance + Walking (no device). The most common outcome, however, is gold_multi_tp_plus_extra, where both gold and prediction include Walking and Using equipment, as in "Mw loopt zelfstandig rond de

afdeling met rollator ." In short, the model is good at detecting mobility but often assigns both aided and unaided movements when devices and walking appear together, and sometimes drops the device when it reads a general "walking" cue.

Across the three pairs, two points are clear. First, clean_single swaps show that single sentences can be genuinely ambiguous. Exercise details can be read as endurance rather than stamina; emotions around a life change can be read as stress coping; therapy contexts can imply device use. Second, the more common outcome is over-assignment in multi-label sentences: the model keeps the gold label and adds the related one (tp_plus_extra_single), or it picks the related label when gold already mixes categories (gold_multi_simple). This suggests practical fixes, such as adding simple competition rules for known pairs, and give the model a little more context (e.g. nearby lines or headers) so it can favour the intended reading, stamina vs exercise, mood vs coping, unaided vs aided movements, instead of predicting both.

7.4 Difficult New Categories (Low Precision)

This subsection examines three newly introduced categories with low precision: D760 Family relationships, B230 Hearing functions, and D240 Handling stress and other psychological demands. In each case I analyse false positives, namely, sentences the model labeled with the category even though the gold label did not include it. The aim is to identify systematic triggers that lead to over-assignment and to relate these triggers to the ICF's intended scope.

D760 Family relationships

The ICF defines Family relationships as creating and maintaining kinship relationships (e.g. parent-child, siblings, extended family) and their functional aspects, not merely recording civil status or family composition. Many false positives arise from demographic statements that mention family but do not describe any functional role or limitation. For example, "Uw familiegegevens U komt uit een gezin van twee kinderen (1 zus) ." and "Gehuwd" were predicted as D760, yet both are administrative facts with no action or participation to evaluate. A second error type is co-mention with affect, where the model adds D760 because "family" appears in an emotional context without a functional claim; e.g., "Familie is heel bang dat als mw ." (gold: B152 Emotional functions). These patterns explain the low precision: the model treats any "family" cue (demographic or affective) as evidence of Family relationships, whereas the ICF category requires relational participation (e.g. caregiving, conflict, support) or its impact on functioning. Tightening decision rules to require a relational action or role (e.g. "zorgt voor...", "past op...", "steunt...") would likely reduce these false positives.

B230 Hearing functions

Hearing functions cover sensing sound and discriminating its qualities (e.g. loudness, pitch, speech discrimination), not voice production or articulation. Here, false positives are driven by lexical confusions. Phrases about voice or articulation (e.g. "Articulatie en spraak : ongestoord ."; "Direct na operatie hees .") were labeled B230 even though they concern speech or voice quality, which the ICF treats outside hearing. Another confusion case is idiomatic "geen gehoor" used for phone contact failure, for example, "...vandaag patiënt niet kunnen bereiken , geen gehoor .", which the model misreads as "no hearing" rather than "no answer". Precision suffers because the classifier maps surface cues ("spraak", "hees", "gehoor") to hearing without disambiguating speech vs. auditory perception or idiom vs. sensory loss. Simple lexical thresholds (e.g. discount hees, stem, articulatie unless hearing terms occur; treat geen gehoor in contact or telephone frames as non-hearing) or more context features (phone call sections, contact attempts) could reduce these errors.

D240 Handling stress and other psychological demands

The ICF intends this category for carrying out coordinated actions to manage demands, responsibilities, time pressure, or crises (e.g. taking exams, meeting deadlines), not for general psychiatric status or service planning. The low precision here comes from two regular triggers. First, service or plan mentions without a coping action, such as "Het verdere beleid ten aanzien van zijn depressie kan besproken worden met de ambulant psychiater (Hoenderdos) ." and "Op order van de psychiater maatschappelijk werk... gevraagd ." The model assigns D240 whenever psychiatric care or social work appears, but the sentences describe referrals or consults, not the person's handling of demands. Second, emotion-coping mixture: when emotional language is present, the model often adds D240 to B152 Emotional functions, as in "Overdracht... patiënt is erg bang... wil graag dat de sertraline herstart wordt ." The presence of fear or treatment preference is not, by itself, evidence of dealing with externally imposed demands. To improve precision, the classifier needs features that prioritize task-oriented coping cues (e.g. meeting deadlines, multitasking, caregiving under pressure) over clinical administration (referrals, diagnoses) and pure emotions.

Across these three categories, the common cause of low precision is over-reliance on surface cues in note-style text. Mentions of family, hearing, voice, psychiatry, or social work act as strong lexical triggers even when the ICF-relevant action is absent. To address the issue, we can seek to encode anchors aligned with ICF definitions, such as verbs and expressions that connect with participation, sensory discrimination, or stress coping actions, and suppress known baiting contexts (demographics, telephony, service). Light context can also be added in the train data to help the model distinguish functional participation from administrative or affective mentions. These adjustments should reduce over assignment and raise the precisions for D760, B230, and D240 without sacrificing recall.

Chapter 8

Discussion & Future Work

This thesis extends sentence-level ICF classification from 10 to 18 categories in Dutch rehabilitation notes by combining manual labels with GPT-4o assisted weak supervision and in-domain fine-tuning (MedRoBERTa.nl). On the updated 18-label gold, the best model achieves macro F1 0.68 at sentence level, competitive for multi-label clinical NLP and sufficient to broaden functional coverage beyond the original system. Crucially, performance on the original 10 categories improves (macro F1 0.67 vs 0.59 for a manual-only baseline), indicating that the added weak labels enriched the decision boundary without harming previous capabilities. In a direct comparison, the fine-tuned MedRoBERTa.nl outperforms GPT-4o used as a few-shot classifier (macro F1 0.58), suggesting that domain-adapted, locally deployable models can extract useful signal from LLM generated labels and generalize better to clinical language use. This aligns with broader evidence that weak supervision can supply effective training signal when gold labels are scarce and that LLM-assisted annotation can reduce human effort for text labeling tasks (Ratner et al., 2020 & Gilardi et al., 2023).

Expanding the label set surfaced substantial previously unlabeled content (e.g. pain, family, equipment assisted mobility). In practice, this means fewer sentences default to None, and more clinically relevant signals are captured for downstream note-level aggregation. Conceptually, the expansion follows WHO's ICF view that functioning is multi-dimensional and context dependent; adding domains such as pain (b280), sleep (b134), cognition (b164), and family relationships (d760) moves the extractor closer to the ICF's intended breadth (WHO, 2001).

Per category's behavior aligns with previous literature: domains with explicit lexical cues (e.g. pain) are easier, while context dependent or demographic mentions (e.g. family relationships) have lower precision, partly a limitation of sentence-level scope and partly ambiguity of definitions. The result echoes calls to include social factors to reduce inequities in disability documentation, while also stressing the modeling challenge those factors pose (Newman-Griffis et al., 2022).

8.1 Limitations

Despite its positive outcomes, this work has several limitations that must be acknowledged. First, the quality of the GPT-4o generated annotations, while reasonable, is imperfect. The LLM was correct roughly 60% of the time on a validation sample, which means a substantial portion of the synthetic training labels were noisy or completely wrong. Our model had to learn from these noisy data. In practice, the model likely averaged the noise given the large volume of examples, but some errors in GPT labeling could have introduced biases. For example, if GPT-4o tended to over-assign certain phrases as a category, the model might internalize that bias (which we saw with the Family relationships false positives). We attempted to counter this by prompt engineering and selecting a high precision mode (temperature 0.1 with definitions), but a more systematic filtering of GPT outputs (or iterative refinement where the model's predictions are checked) could further improve training data quality (Ratner et al., 2020).

Second, there are limitations in the gold standard evaluation for the new categories. Due to time constraints, we did not have full manual annotation of all test sentences for the eight new categories. Instead, we relied on an approach where new-category instances predicted by the model were validated by clinicians. This means it's possible that some sentences in the test set truly contain a new category issue but were never identified by the model (and thus never validated as gold). Those would still be marked as "None" in our gold labels, potentially counting as false negatives that we did not recognize. In other words, our evaluation of recall for new categories could be overly optimistic, since we only evaluated on the cases the model found. A more rigorous evaluation would involve having experts annotate a random sample of test notes for all categories comprehensively. The lack of a fully blinded, comprehensive gold standard for new categories is a limitation that leaves some uncertainty in the reported metrics.

8.2 Future Work

Improving the quality of synthetic labels is a clear next step. One idea is to employ iterative weak supervision, where the model trained on initial GPT labels is used to re-annotate the data or to identify likely mistakes, which are then corrected by an expert or by another round of GPT with a different prompt. Alternatively, multiple LLMs or multiple prompts could label the data, and one could take the consensus as a more reliable label (a form of ensemble labeling).

Incorporating an active learning loop (where the model identifies sentences it is most uncertain about, and those are sent for manual annotation) could efficiently concentrate expert effort on the trickiest cases and improve the model further (Ratner et al., 2020 & Fries et al., 2021).

As discussed, one limitation is the lack of broader context for each sentence. Future work could explore hybrid models that combine sentence-level and note-level approaches. For example, a hierarchical model might first read the entire note to get an idea of the overall context (e.g. the patient's primary issues), and then classify each sentence with that context in mind. Another approach is to use post-processing rules: if a note has multiple sentences all related to a certain topic (say family situation), ensure consistent labeling across them or aggregate them to decide if the theme is truly significant. Sequence models or joint inference across sentences (using techniques like conditional random fields or attention across sentence embeddings) could also help resolve ambiguities that a single-sentence model cannot (Yang et al., 2016). LLMs can also generate clinical-style text to alleviate data scarcity. The MedSyn framework combines a knowledge graph with GPT-4 or LlamA to produce synthetic discharge summaries; adding these notes improved ICD-10 accuracy by up to 17.8% on tail codes, subject to quality controls and bias checks (Kumichev et al., 2024). I did not generate new notes in this study, but this line of work motivates future augmentation once the eight new ICF domains are fully validated.

Finally, an important future direction is to test the system in real clinical workflows. This includes conducting user studies where clinicians use the model's outputs. Their feedback could identify if the extracted ICF labels align with clinical intuition and needs. For deployment, considerations like model interpretability (providing highlights of text for each predicted category), reliability, and integration into electronic health record systems will be key. Additionally, evaluating the model's impact, for example, does providing automated functional labels help in care planning or in monitoring patient progress? Further evaluations would help demonstrate the true utility of this work. Even if the model is not perfect, if it can flag, say, that a patient has pain or social support issues, it could alert providers to address those aspects, so as to improve comprehensive care (Gilardi et al., 2023).

Chapter 9

Conclusion

Effective management and research of health conditions require understanding not only what diseases patients have, but also how those conditions affect their daily functioning. This thesis tackled that challenge by extending a clinical text classification model to recognize a wider spectrum of functioning related information, as defined by the ICF framework. We focused on Dutch rehabilitation notes and implemented a novel strategy that combined limited expert-labeled data with GPT-4o assisted annotations for new categories. The extended model can automatically identify 17 ICF categories (plus "None") in narrative sentences, ranging from physical functions like pain, mobility, and sleep to psychological and social aspects like family relationships and stress handling.

Our approach proved that using a LMM for weak supervision is both practical and beneficial. We showed that an in-domain MedRoBERTa model trained on a mixture of gold and synthetic labels achieved robust performance (macro-F1 0.68) in multi-label classification, outperforming the direct GPT-4o classification baseline and maintaining high accuracy on previously established categories. In doing so, we substantially increased the coverage of relevant information: the system now sees details that were previously missed (for example, identifying pain or family support issues mentioned in text that earlier would have gone unlabeled). These results underscore the potential of combining human domain knowledge with AI assistance to rapidly adapt NLP tools to evolving information needs in healthcare.

The findings of this work are significant for both clinical practice and future research. From a clinical standpoint, the augmented classifier could be used to enrich electronic health records with structured data about patient functioning. This could enable better tracking of rehabilitation progress, facilitate case mix comparisons (by quantifying functioning profiles), and ensure that "hidden" factors like psychosocial support are brought to light in care planning. From a research perspective, our methodology contributes to the growing evidence that small, carefully constructed datasets augmented with LLM-generated labels can contribute to high-performing models, even in specialized domains. It also demonstrates a viable workflow for

updating NLP systems, rather than requiring a full new annotation effort, one can make use of an existing model (GPT-4o in this case) to extend the label space with relatively low cost.

In closing, this thesis advances the goal of scalable, comprehensive ICF classification from clinical text. We have shown that it is possible to broaden an NLP system's understanding of patient records, capturing not just a narrow set of categories but a richer coverage of a patient's functioning and environment without a heavy annotation burden. The approach and insights here lay the foundation for future improvements, such as incorporating more context, refining category definitions, and expanding to additional domains. By continuing to improve on these fronts, we move towards NLP tools that better support clinicians and patients, ensuring that key aspects of health and disability are not lost in the narrative but are systematically recognized and acted upon. The contribution of this work is a step toward the vision that an AI-assisted pipeline brings us closer to holistic patient information extraction, ultimately aiding in more informed and equitable healthcare.

Appendix A

Additional Tables and Figures

18 ICF Categories & Definitions

1. Old Categories (9)

B1300 Energy level
Mental functions that produce vigour and stamina.

B140 Attention functions
Specific mental functions of focusing on an external stimulus or internal experience for the required period of time.
Inclusions: functions of sustaining attention, shifting attention, dividing attention, sharing attention; concentration; distractibility
Exclusions: consciousness functions; energy and drive functions; sleep functions; memory functions; psychomotor functions; perceptual functions

B152 Emotional functions
Specific mental functions related to the feeling and affective components of the processes of the mind.
Inclusions: functions of appropriateness of emotion, regulation and range of emotion; affect; sadness, happiness, love, fear, anger, hate, tension, anxiety, joy, sorrow; lability of emotion; flattening of affect
Exclusions: temperament and personality functions; energy and drive functions

B440 Respiration functions
Functions of inhaling air into the lungs, the exchange of gases between air and blood, and exhaling air.
Inclusions: functions of respiration rate, rhythm and depth; impairments such as apnoea, hyperventilation, irregular respiration, paradoxical respiration, and brochial spasm, and as in

pulmonary emphysema; upper pulmonary obstruction, reduction in airflow through upper and lower airways
Exclusions: respiratory muscle functions; additional respiratory functions; exercise tolerance functions

B455 Exercise tolerance functions
Functions related to respiratory and cardiovascular capacity as required for enduring physical exertion.
Inclusions: functions of physical endurance, aerobic capacity, stamina and fatiguability
Exclusions: functions of the cardiovascular system; haematological system functions; respiration functions; respiratory muscle functions; additional respiratory functions

B530 Weight maintenance functions
Functions of maintaining appropriate body weight, including weight gain during the developmental period.
Inclusions: functions of maintenance of acceptable Body Mass Index (BMI); and impairments such as underweight, cachexia, wasting, overweight, emaciation and such as in primary and secondary obesity
Exclusions: assimilation functions; general metabolic functions; endocrine gland functions

D450 Walking
Moving along a surface on foot, step by step, so that one foot is always on the ground, such as when strolling, sauntering, walking forwards, backwards, or sideways.
Inclusions: walking short or long distances; walking on different surfaces; walking around obstacles
Exclusions: transferring oneself; moving around

D550 Eating
Indicating need for, and carrying out the coordinated tasks and actions of eating food that has been served, bringing it to the mouth and consuming it in culturally acceptable ways, cutting or breaking food into pieces, opening bottles and cans, using eating implements, having meals, feasting or dining.
Exclusion: drinking

D840-D859 Work and employment
Apprenticeship (work preparation)
Engaging in programmes related to preparation for employment, such as performing the tasks required of an apprenticeship, internship, articling and in service training.
Exclusion: vocational training

Acquiring, keeping and terminating a job
Seeking, finding and choosing employment, being hired and accepting employment, maintaining and advancing through a job, trade, occupation or profession, and leaving a job in an appropriate manner.
Inclusions: seeking employment; preparing a resume or curriculum vitae; contacting employers and preparing interviews; maintaining a job; monitoring one's own work performance; giving notice; and terminating a job

Remunerative employment
Engaging in all aspects of work, as an occupation, trade, profession or other form of employment, for payment, as an employee, full or part time, or self-employed, such as seeking employment and getting a job, doing the required tasks of the job, attending work on time as required, supervising other workers or being supervised, and performing required tasks alone or in groups.
Inclusions: self-employment, part-time and full-time employment

Non-remunerative employment
Engaging in all aspects of work in which pay is not provided, full-time or part-time, including organized work activities, doing the required tasks of the job, attending work on time as required, supervising other workers or being supervised, and performing required tasks alone or in groups, such as volunteer work, charity work, working for a community or religious group without remuneration, working around the home without remuneration.
Exclusion: Domestic Life

Work and employment, other specified and unspecified

2. New Categories (8)

B280 Sensations of pain
Sensation of unpleasant feeling indicating potential or actual damage to some body structure.
Inclusions: sensations of generalized or localized pain, in one or more body part, pain in a dermatome, stabbing pain, burning pain, dull pain, aching pain; impairments such as myalgia, analgesia and hyperalgesia

B134 Sleep functions
General mental functions of periodic, reversible and selective physical and mental disengagement from one's immediate environment accompanied by characteristic physiological changes.

Inclusions: functions of amount of sleeping, and onset, maintenance and quality of sleep; functions involving the sleep cycle, such as in insomnia, hypersomnia and narcolepsy
Exclusions: consciousness functions; energy and drive functions; attention functions; psychomotor functions

D760 Family relationships
Creating and maintaining kinship relationships, such as with members of the nuclear family, extended family, foster and adopted family and step relationships, more distant relationships such as second cousins, or legal guardians.
Inclusions: parent-child and child-parent relationships, sibling and extended family relationships

B164 Higher-level cognitive functions
Specific mental functions especially dependent on the frontal lobes of the brain, including complex goal-directed behaviours such as decision-making, abstract thinking, planning and carrying out plans, mental flexibility, and deciding which behaviours are appropriate under what circumstances; often called executive functions.
Inclusions: functions of abstraction and organization of ideas; time management, insight and judgement; concept formation, categorization and cognitive flexibility Exclusions: memory functions; thought functions; mental functions of language; calculation functions

D465 Moving around using equipment
Moving the whole body from place to place, on any surface or space, by using specific devices designed to facilitate moving or create other ways of moving around, such as with skates, skis, scuba equipment, swim fins, or moving down the street in a wheelchair or a walker.
Exclusions: transferring oneself; walking; moving around; using transportation; driving

D410 Changing basic body position
Getting into and out of a body position and moving from one location to another, such as rolling from one side to the other, sitting, standing, getting up out of a chair to lie down on a bed, and getting into and out of positions of kneeling or squatting.
Inclusion: changing body position from lying down, from squatting or kneeling, from sitting or standing, bending and shifting the body's centre of gravity Exclusion: transferring oneself

B230 Hearing functions
Sensory functions relating to sensing the presence of sounds and discriminating the location, pitch, loudness and quality of sounds.
Inclusions: functions of hearing, auditory discrimination, localization of sound source, lateralization of sound, speech discrimination; impairments such as deafness, hearing impairment

and hearing loss
Exclusions: perceptual functions and mental functions of language

D240 Handling stress and other psychological demands
Carrying out simple or complex and coordinated actions to manage and control the psychological demands required to carry out tasks demanding significant responsibilities and involving stress, distraction, or crises, such as taking exams, driving a vehicle during heavy traffic, putting on clothes when hurried by parents, finishing a task within a time-limit or taking care of a large group of children.
Inclusions: handling responsibilities; handling stress and crisis

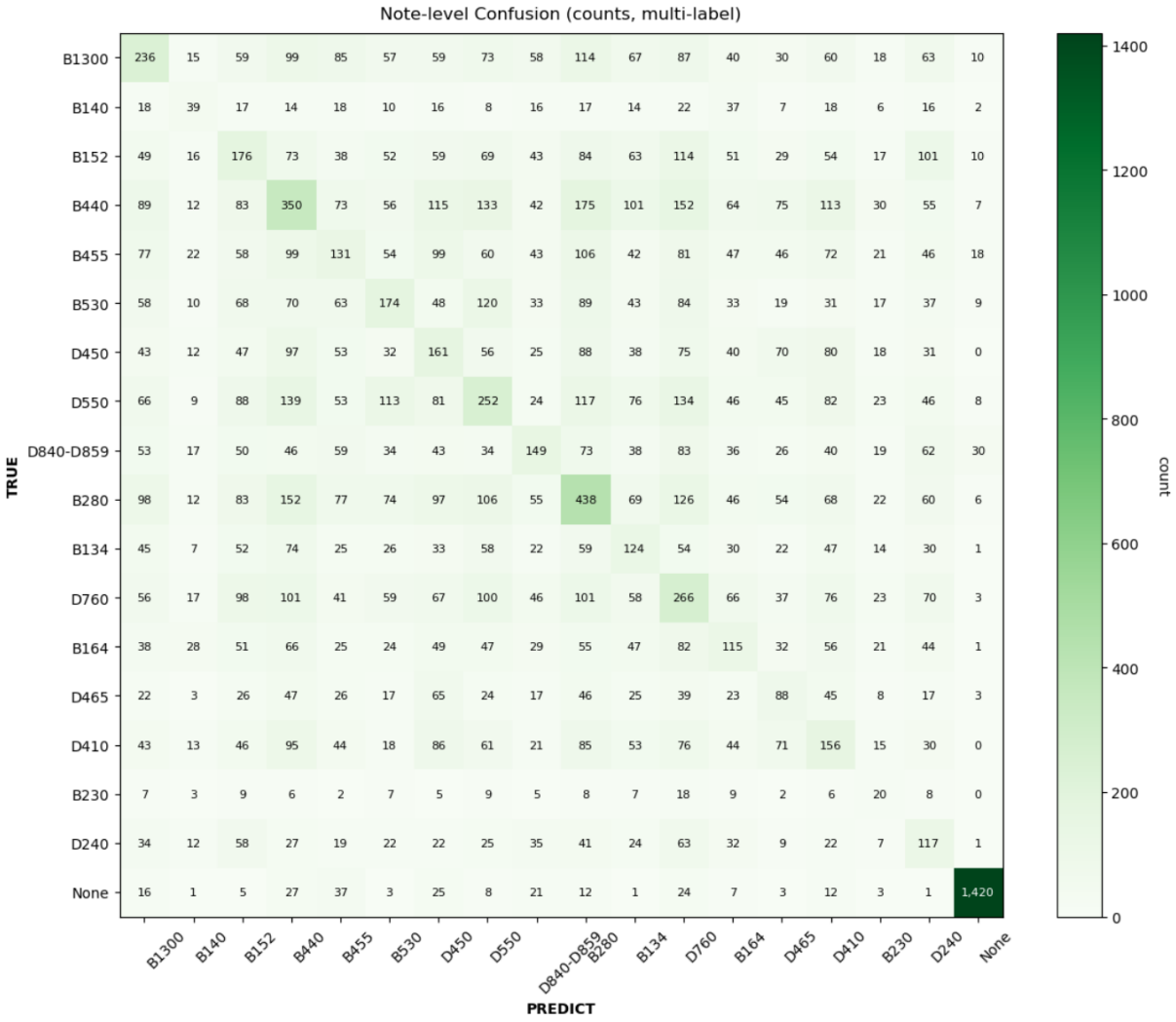


Figure A.1: Final Augmented Data (D) Note-Level Confusion Matrix

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	1121	0.48	838	10.33
B140 Attention functions	407	0.17	335	4.13
B152 Emotional functions	3490	1.5	2097	25.84
B440 Respiration functions	5047	2.16	2484	30.61
B455 Exercise tolerance functions	1132	0.49	914	11.26
B530 Weight maintenance functions	904	0.39	703	8.66
D450 Walking	2645	1.13	1786	22.01
D550 Eating	2592	1.11	1624	20.01
D840-D859 Work and employment	610	0.26	508	6.26
B280 Sensations of pain	6982	2.99	2560	31.55
B134 Sleep functions	3118	1.34	1820	22.43
D760 Family relationships	5289	2.27	2231	27.49
B164 Higher-level cognitive functions	4367	1.87	1670	20.58
D465 Moving around using equipment	2748	1.18	1592	19.62
D410 Changing basic body position	5079	2.18	2316	28.54
B230 Hearing functions	1082	0.46	693	8.54
D240 Handling stress and other psychological demands	5263	2.26	2334	28.76
None	190359	81.62	4771	58.79
TOTAL_SENTENCES	233227	100.0		
TOTAL_NOTES			8115	100.0

Figure A.2: Dataset A annotated: train_jenia_murat statistics

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	509	1.26	293	19.53
B140 Attention functions	170	0.42	111	7.4
B152 Emotional functions	917	2.27	368	24.53
B440 Respiration functions	1394	3.45	549	36.6
B455 Exercise tolerance functions	486	1.2	283	18.87
B530 Weight maintenance functions	557	1.38	299	19.93
D450 Walking	1112	2.75	471	31.4
D550 Eating	475	1.18	303	20.2
D840-D859 Work and employment	307	0.76	191	12.73
B280 Sensations of pain	2392	5.92	797	53.13
B134 Sleep functions	672	1.66	378	25.2
D760 Family relationships	1354	3.35	549	36.6
B164 Higher-level cognitive functions	663	1.64	275	18.33
D465 Moving around using equipment	605	1.5	345	23.0
D410 Changing basic body position	463	1.15	270	18.0
B230 Hearing functions	627	1.55	184	12.27
D240 Handling stress and other psychological demands	463	1.15	235	15.67
None	28287	69.99	1477	98.47
TOTAL_SENTENCES	40414	100.0		
TOTAL_NOTES			1500	100.0

Figure A.3: Dataset B annotated: AMC 2023 statistics

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	989	5.89	707	12.26
B140 Attention functions	247	1.47	175	3.03
B152 Emotional functions	3390	20.2	1989	34.49
B440 Respiration functions	4988	29.72	2345	40.66
B455 Exercise tolerance functions	1967	11.72	1260	21.85
B530 Weight maintenance functions	755	4.5	546	9.47
D450 Walking	2489	14.83	1631	28.28
D550 Eating	2420	14.42	1416	24.55
D840-D859 Work and employment	486	2.9	381	6.61
TOTAL_SENTENCES	16781	100.0		
TOTAL_NOTES			5767	100.0

Figure A.4: Dataset D's 1st composition: train_eb_ap_jenia_all-labels statistics

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	1545	3.6	1148	16.15
B140 Attention functions	380	0.88	266	3.74
B152 Emotional functions	2414	5.62	1256	17.67
B440 Respiration functions	6657	15.5	2877	40.47
B455 Exercise tolerance functions	1750	4.07	1145	16.11
B530 Weight maintenance functions	2512	5.85	1392	19.58
D450 Walking	2990	6.96	1703	23.96
D550 Eating	2596	6.04	1561	21.96
D840-D859 Work and employment	813	1.89	545	7.67
B280 Sensations of pain	9645	22.45	3833	53.92
B134 Sleep functions	2368	5.51	1701	23.93
D760 Family relationships	4397	10.24	2306	32.44
B164 Higher-level cognitive functions	2113	4.92	923	12.98
D465 Moving around using equipment	1318	3.07	848	11.93
D410 Changing basic body position	1516	3.53	934	13.14
B230 Hearing functions	1449	3.37	585	8.23
D240 Handling stress and other psychological demands	1456	3.39	781	10.99
None	0	0.0	0	0.0
TOTAL_SENTENCES	42956	100.0		
TOTAL_NOTES			7109	100.0

Figure A.5: Dataset D's 2st composition: VUMC 2023 statistics

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	4163	1.22	2267	13.81
B140 Attention functions	1204	0.35	680	4.14
B152 Emotional functions	10224	3.0	3727	22.7
B440 Respiration functions	18192	5.34	5930	36.11
B455 Exercise tolerance functions	5335	1.57	2852	17.37
B530 Weight maintenance functions	4748	1.39	2346	14.29
D450 Walking	9238	2.71	3960	24.12
D550 Eating	8129	2.39	3502	21.33
D840-D859 Work and employment	2222	0.65	1191	7.25
B280 Sensations of pain	19019	5.58	7185	43.76
B134 Sleep functions	6158	1.81	3894	23.71
D760 Family relationships	11040	3.24	5084	30.96
B164 Higher-level cognitive functions	7143	2.1	2838	17.28
D465 Moving around using equipment	4669	1.37	2782	16.94
D410 Changing basic body position	7054	2.07	3514	21.4
B230 Hearing functions	3158	0.93	1460	8.89
D240 Handling stress and other psychological demands	7182	2.11	3346	20.38
None	225662	66.26	6257	38.11
TOTAL_SENTENCES	340592	100.0		
TOTAL_NOTES			16420	100.0

Figure A.6: Dataset D annotated statistics

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	413	1.11	310	10.44
B140 Attention functions	82	0.22	61	2.05
B152 Emotional functions	348	0.93	215	7.24
B440 Respiration functions	1063	2.85	383	12.9
B455 Exercise tolerance functions	382	1.02	252	8.49
B530 Weight maintenance functions	341	0.91	214	7.21
D450 Walking	362	0.97	176	5.93
D550 Eating	640	1.71	299	10.07
D840-D859 Work and employment	331	0.89	267	8.99
TOTAL_SENTENCES	37355	100.0		
TOTAL_NOTES			2969	100.0

Figure A.7: Test data statistics before label updates

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	413	1.11	310	10.44
B140 Attention functions	82	0.22	61	2.05
B152 Emotional functions	348	0.93	215	7.24
B440 Respiration functions	1063	2.85	383	12.9
B455 Exercise tolerance functions	382	1.02	252	8.49
B530 Weight maintenance functions	341	0.91	214	7.21
D450 Walking	362	0.97	176	5.93
D550 Eating	640	1.71	299	10.07
D840-D859 Work and employment	331	0.89	267	8.99
TOTAL_SENTENCES	37355	100.0		
TOTAL_NOTES			2969	100.0

Figure A.8: Test data statistics before label updates

Category	Sentence Count	Sentence %	Note Count	Note %
B1300 Energy level	448	1.2	339	11.42
B140 Attention functions	64	0.17	51	1.72
B152 Emotional functions	362	0.97	232	7.81
B440 Respiration functions	1060	2.84	397	13.37
B455 Exercise tolerance functions	398	1.07	276	9.3
B530 Weight maintenance functions	338	0.9	212	7.14
D450 Walking	368	0.99	184	6.2
D550 Eating	665	1.78	309	10.41
D840-D859 Work and employment	342	0.92	268	9.03
B280 Sensations of pain	847	2.27	457	15.39
B134 Sleep functions	183	0.49	132	4.45
D760 Family relationships	472	1.26	274	9.23
B164 Higher-level cognitive functions	303	0.81	143	4.82
D465 Moving around using equipment	162	0.43	111	3.74
D410 Changing basic body position	390	1.04	201	6.77
B230 Hearing functions	37	0.1	22	0.74
D240 Handling stress and other psychological demands	212	0.57	132	4.45
None	31692	84.84	2947	99.26
TOTAL_SENTENCES	37355	100.0		
TOTAL_NOTES			2969	100.0

Figure A.9: Test data statistics after label updates

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	\
precision	0.64	0.90	0.57	0.77	0.40	0.55	0.42	0.55	
recall	0.76	0.34	0.57	0.64	0.86	0.66	0.28	0.81	
f1-score	0.69	0.50	0.57	0.70	0.55	0.60	0.34	0.65	
support	1063.00	82.00	331.00	413.00	640.00	362.00	382.00	341.00	

	STM	none	micro avg	macro avg	weighted avg	samples avg
precision	0.56	0.99	0.91	0.64	0.95	0.69
recall	0.63	0.68	0.68	0.62	0.68	0.69
f1-score	0.59	0.81	0.78	0.60	0.79	0.69
support	348.00	33769.00	37731.00	37731.00	37731.00	37731.00

		PREDICT									
		ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	none
TRUE	ADM	812	1	0	20	12	17	33	7	3	73
	ATT	1	28	0	8	1	1	0	0	2	1
	BER	2	3	188	6	2	6	5	1	4	6
	ENR	26	2	8	263	5	12	41	12	6	2
	ETN	9	0	0	5	551	1	0	39	1	27
	FAC	10	0	0	6	0	240	21	0	4	6
	INS	30	3	4	39	2	62	108	3	2	20
	MBW	1	0	0	8	63	1	4	277	6	16
	STM	2	0	8	2	2	3	2	4	220	5
	none	456	3	139	66	786	163	107	216	162	23103

Figure A.10: Medroberta 10-Category Results, Dataset C with Down-sampled 'None'

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	\
precision	0.87	0.86	0.67	0.79	0.63	0.67	0.40	0.82	
recall	0.58	0.44	0.40	0.62	0.54	0.69	0.38	0.61	
f1-score	0.70	0.58	0.50	0.69	0.58	0.68	0.39	0.70	
support	1063.00	82.00	331.00	413.00	640.00	362.00	382.00	341.00	

	STM	none	micro avg	macro avg	weighted avg	samples avg
precision	0.68	0.97	0.94	0.74	0.94	0.84
recall	0.64	0.87	0.84	0.58	0.84	0.84
f1-score	0.66	0.92	0.89	0.64	0.89	0.84
support	348.00	33769.00	37731.00	37731.00	37731.00	37731.00

		PREDICT									
		ADM	ATT	BER	ENR	ETN	FAC	INS	MBW	STM	none
TRUE	ADM	615	2	0	25	8	18	30	5	4	333
	ATT	3	36	1	7	1	2	8	0	3	4
	BER	5	2	131	14	1	6	23	0	4	99
	ENR	31	4	12	254	5	12	69	9	3	45
	ETN	7	0	2	3	346	1	2	37	3	242
	FAC	14	0	0	5	0	250	31	0	5	20
	INS	24	3	6	25	1	56	144	2	3	89
	MBW	8	0	1	12	48	2	3	208	7	96
	STM	1	0	8	5	1	3	4	3	224	40
	none	87	6	58	58	176	94	148	41	97	29456

Figure A.11: Medroberta 10-Category Results, Dataset D, 5 epochs, LR 4e-5

	ADM	ATT	BER	ENR	ETN	FAC	INS	MBW
precision	0.86	0.85	0.62	0.80	0.64	0.69	0.42	0.78
recall	0.56	0.49	0.41	0.60	0.56	0.72	0.40	0.67
f1-score	0.68	0.62	0.49	0.68	0.59	0.71	0.41	0.72
support	1063.00	82.00	331.00	413.00	640.00	362.00	382.00	341.00

	STM	none	micro avg	macro avg	weighted avg	samples avg
precision	0.62	0.97	0.94	0.72	0.94	0.84
recall	0.61	0.87	0.84	0.59	0.84	0.84
f1-score	0.62	0.92	0.89	0.64	0.89	0.84
support	348.00	33769.00	37731.00	37731.00	37731.00	37731.00

		PREDICT								
		ADM	**ATT**	**BER**	**ENR**	**ETN**	**FAC**	**INS**	**MBW**	**STM**	**none**
TRUE	**ADM**	592	3	0	25	11	21	35	8	7	348
	ATT	3	40	2	7	1	3	2	0	3	7
	BER	5	2	135	9	2	6	23	4	5	102
	ENR	27	5	16	246	4	13	56	13	6	38
	ETN	8	0	1	3	356	1	2	40	2	245
	FAC	9	0	0	1	0	262	39	0	5	20
	INS	20	6	6	25	2	59	152	3	6	90
	MBW	6	0	0	11	51	2	4	230	8	93
	STM	1	0	8	1	1	3	10	5	214	40
	none	90	7	74	55	178	85	152	53	117	29368

Figure A.12: Dataset A finetuned Medroberta, 10-Category Results, Dataset D, LR gradually decrease from 4e-5

Appendix B

References

Adejumo, P., Thangaraj, P. M., Dhingra, L. S., Aminorroaya, A., Zhou, X., Brandt, C., Xu, H., Krumholz, H. M., & Khera, R. (2024). Natural Language Processing of Clinical Documentation to Assess Functional Status in Patients With Heart Failure. JAMA network open, 7(11), e2443925.

Akbasli, I. T., Birbilen, A. Z., & Teksam, O. (2025). Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-english languages. BMC medical informatics and decision making, 25(1), 154.

Altalla, B., Abdalla, S., Altamimi, A. et al. (2025). Evaluating GPT models for clinical note de-identification. Sci Rep 15, 3852.

Bhowmick, P. K., Basu, A., & Mitra, P. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In COLING 2008: Proceedings of the Workshop on Human Judgements in Computational Linguistics (pp. 58–65). Coling 2008 Organizing Committee.

Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., & Manning, C.D. (2024). BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text.

Bolton, E., Xiong, B., Muralidharan, V., Schamroth, J., Muralidharan, V., Manning, C., & Daneshjou, R. (2024). Assessing the potential of mid-sized language models for clinical QA.

Bosma, J.S., Dercksen, K., Builtjes, L. et al. (2025). The DRAGON benchmark for clinical NLP. npj Digit. Med. 8, 289.

Chen, Q., Hu, Y., Peng, X. et al. (2025) Benchmarking large language models for biomed-

ical natural language processing applications and recommendations. Nat Commun 16, 3280.

Chen, S., Li, Y., Lu, S., Van, H., Aerts, H. J. W. L., Savova, G. K., & Bitterman, D. S. (2024). Evaluating the ChatGPT family of models for biomedical reasoning and classification. Journal of the American Medical Informatics Association : JAMIA, 31(4), 940–948.

Chen, Y., Li, C., Dai, X., Li, J., Sun, W., Wang, Y., Zhang, R., Zhang, T., & Wang, B. (2024). Boosting single positive multi-label classification with generalized robust loss. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24) (pp. 3825–3833). IJCAI.

Cho, H. N., Jun, T. J., Kim, Y. H., Kang, H., Ahn, I., Gwon, H., Kim, Y., Seo, J., Choi, H., Kim, M., Han, J., Kee, G., Park, S., & Ko, S. (2024). Task-Specific Transformer-Based Language Models in Health Care: Scoping Review. JMIR medical informatics, 12, e49724.

Chung, P., Yun, S. J., & Khan, F. (2014). A comparison of participation outcome measures and the International Classification of Functioning, Disability and Health Core Sets for traumatic brain injury. Journal of rehabilitation medicine, 46(2), 108–116.

Cieza, A., Brockow, T., Ewert, T., Amman, E., Kollerits, B., Chatterji, S., Ustün, T. B., & Stucki, G. (2002). Linking health-status measurements to the international classification of functioning, disability and health. Journal of rehabilitation medicine, 34(5), 205–210.

Cieza, A., Stucki, G., Weigl, M., Disler, P., Jäckel, W., van der Linden, S., Kostanjsek, N., & de Bie, R. (2004). ICF Core Sets for low back pain. Journal of rehabilitation medicine, (44 Suppl), 69–74.

Cieza, A., Stucki, G., Weigl, M., Kullmann, L., Stoll, T., Kamen, L., Kostanjsek, N., & Walsh, N. (2004). ICF Core Sets for chronic widespread pain. Journal of rehabilitation medicine, (44 Suppl), 63–68.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model.

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based language model. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 3255–3265). Association for Computational Linguistics.

Du, X., Wang, Y., Zhou, Z., Chuang, Y.-W., Yang, R., Zhang, W., Wang, X., Zhang, R., Hong, P.,

Bates, D., & Zhou, L. (2024). Generative large language models in electronic health records for patient care since 2023: A systematic review. medRxiv.

Fries, J.A., Steinberg, E., Khattar, S. et al. (2021). Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun 12, 2017.

Fu, S., Jia, H., Vassilaki, M., Keloth, V. K., Dang, Y., Zhou, Y., Garg, M., Petersen, R. C., St Sauver, J., Moon, S., Wang, L., Wen, A., Li, F., Xu, H., Tao, C., Fan, J., Liu, H., & Sohn, S. (2024). FedFSA: Hybrid and federated framework for functional status ascertainment across institutions. Journal of biomedical informatics, 152, 104623.

Geyh, S., Cieza, A., Schouten, J., Dickson, H., Frommelt, P., Omar, Z., Kostanjsek, N., Ring, H., & Stucki, G. (2004). ICF Core Sets for stroke. Journal of rehabilitation medicine, (44 Suppl), 135–141.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120.

Grill, E., Joisten, S., Swoboda, W., & Stucki, G. (2007). Early-stage impairments and limitations of functioning from the geriatric ICF core set as determinants of independent living in older patients after discharge from post-acute rehabilitation. Journal of rehabilitation medicine, 39(8), 591–597.

Gül, H., Çınar, M. A., & Bayramlar, K. (2025). ChatGPT as a collaborative research assistant in the ICF linking process of the brief version of the Burn Specific Health Scale. Burns : journal of the International Society for Burn Injuries, 51(7), 107609.

Guo, Y., Ovadje, A., Al-Garadi, M. A., & Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. Journal of the American Medical Informatics Association : JAMIA, 31(10), 2181–2189.

Hakbijl, A., Rohn, E., Tate, D., van Leeuwen, C., Forchheimer, M., Stolwijk-Swüste, J., Charlifue, S., Greve, J., New, P., & Post, M. (2023). The social dimension of quality of life following spinal cord injury or disease: An international ICF-linking study. Spinal Cord, 62.

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A survey of methods for addressing class imbalance in deep-learning-based natural language processing. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 523–540). Association for Computational Linguistics.

Hernández-Lázaro, H., Jiménez-Del Barrio, S., Ceballos-Laita, L., Lahuerta-Martin, S., Medrano-de-la-Fuente, R., Hernando-Garijo, I., & Mingo-Gómez, M. T. (2023). Multicentre cross-sectional study assessing content validity of the International Classification of Functioning, disability and health core set for post-acute musculoskeletal conditions in primary care physiotherapy services. Journal of rehabilitation medicine, 55, jrm11950.

Hsu, E., Roberts, K. (2025). Leveraging large language models for knowledge-free weak supervision in clinical natural language processing.

Hu, W., Fan, Q., Yan, H., Xu, X., Huang, S., & Zhang, K. (2025). A Survey of Multi-Label Text Classification Under Few-Shot Scenarios. Applied Sciences, 15(16), 8872.

Kim, J., Verkijk, S., Geleijn, E., Leeden, M., Meskers, C., Meskers, C., van der Veen, S., Vossen, P., & Widdershoven, G. (2022). Modeling Dutch medical texts for detecting functional categories and levels of COVID-19 patients.

Kumichev, G., Blinov, P., Kuzkina, Y., Goncharov, V., Zubkova, G., Zenovkin, N., Goncharov, A., & Savchenko, A. (2024). MedSyn: LLM-based synthetic medical text generation framework. In Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X (pp. 215–230). Springer.

Lopez, I., Swaminathan, A., Vedula, K. et al. (2025). Clinical entity augmented retrieval for clinical information extraction. npj Digit. Med. 8, 45.

Lu, Q., Li, R., Wen, A., Wang, J., Wang, L., & Liu, H. (2025). Large language models struggle in token-level clinical named entity recognition. AMIA Annual Symposium Proceedings, 2024.

Matos, J., Gallifant, J., Pei, J., & Wong, A. (2024). EHRmonize: A framework for medical concept abstraction from electronic health records using large language models.

Menezes, M., Hoffmann, A., Tan, A., Nalbandyan, M., Omenn, G., Mazzotti, D., Hernández-Arango, A., Visweswaran, S., Venkatesh, S., Mandl, K., Bourgeois, F., Lee, J., Makmur, A., Hanauer, D., Semanik, M., Kerivan, L., Hill, T., Forero, J., & Restrepo, C. (2024). The potential of generative pre-trained transformer 4 (GPT-4) to analyse medical notes in three different languages: A retrospective model-evaluation study. The Lancet Digital Health, 7, e35–e43.

Meskers, C. G. M., van der Veen, S., Kim, J., Meskers, C. J. W., Smit, Q. T. S., Verkijk, S.,

Geleijn, E., Widdershoven, G. A. M., Vossen, P. T. J. M., & van der Leeden, M. (2022). 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.

Muizelaar, H., Haas, M., van Dortmont, K. et al. (2024). Extracting patient lifestyle characteristics from Dutch clinical text with BERT models. BMC Med Inform Decis Mak 24, 151.

Murphy, R.M., Dongelmans, D.A., de Keizer, N.F. et al. (2025). Creation of a gold standard Dutch corpus of clinical notes for adverse drug event detection: the Dutch ADE corpus. Lang Resources & Evaluation 59, 2763–2779.

Naguib, M., Tannier, X., & Névéol, A. (2024). Few-shot clinical entity recognition in English, French and Spanish: Masked language models outperform generative model prompting. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 6829–6852). Association for Computational Linguistics.

Newman-Griffis, D. R., Hurwitz, M. B., McKernan, G. P., Houtrow, A. J., & Dicianno, B. E. (2022). A roadmap to reduce information inequities in disability with digital health and natural language processing. PLOS digital health, 1(11), e0000135.

Newman-Griffis, D., & Fosler-Lussier, E. (2021). Automated Coding of Under-Studied Medical Concept Domains: Linking Physical Activity Reports to the International Classification of Functioning, Disability, and Health. Frontiers in digital health, 3, 620828.

Nieminen, L., Ketamo, H., & Kankaanpää, M. (2025). ICF coding automated: A validation study for self-supervised architecture in electronic health records. Research Square.

Oliveira, V., Nogueira, G., Faleiros, T. et al. (2025). Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. Artif Intell Law 33, 361–381.

Persoon, A., Banningh, L. J., van de Vrie, W., Rikkert, M. G., & van Achterberg, T. (2011). Development of the Nurses' Observation Scale for Cognitive Abilities (NOSCA). ISRN nursing, 2011, 895082.

Płaszewski, M., & Płaszewski, K. (2025). ICF-based assessment of functioning–State-of-the-art and challenges: A user's perspective. Preprints.

Prodinger, B., Tennant, A., & Stucki, G. (2018). Standardized reporting of functioning information on ICF-based common metrics. European journal of physical and rehabilitation medicine, 54(1), 110–117.

Ratner, A., Bach, S.H., Ehrenberg, H. et al. (2020). Snorkel: rapid training data creation with weak supervision. The VLDB Journal 29, 709–730.

Rink, L., Tomandl, J., Womser, S., Kühlein, T., & Sebastião, M. (2023). Development of a subset of the international classification of functioning, disability and health as a basis for a questionnaire for community-dwelling older adults aged 75 and above in primary care: a consensus study. BMJ open, 13(8), e072184.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., Prakash, S., … Natarajan, V. (2025). Toward expert-level medical question answering with large language models. Nature medicine, 31(3), 943–950.

Stallinga, H. A., Bakker, J., Haan, S. J., van Os-Medendorp, H., Kars, M. C., Overgoor, L., Stewart, R. E., & Roodbol, P. F. (2021). The Usability of the Preliminary ICF Core Set for Hospitalized Patients After a Hematopoietic Stem Cell Transplantation From the Perspective of Nurses: A Feasibility Study. Frontiers in rehabilitation sciences, 2, 710127.

Stucki, A & Cieza, A & Michel, F & Stucki, Prof. Dr. med. Gerold & Bentley, Alison & Culebras, A & Tufik, S & Kotchabhakdi, Naiphinich & Tachibana, Naoko & Ustun, Tevfik & Partinen, Markku. (2008). Developing ICF Core Sets for persons with sleep disorders based on the International Classification of Functioning, Disability and Health. Sleep medicine. 9. 191-8. 10.1016/j.sleep.2007.01.019.

Tan, B., Liu, L., & Yu, L. (2025). ICF: The endpoint of rehabilitation assessment? Regenesis Repair Rehabilitation, 1.

Thieu, T., Maldonado, J. C., Ho, P. S., Ding, M., Marr, A., Brandt, D., Newman-Griffis, D., Zirikly, A., Chan, L., & Rasch, E. (2021). A comprehensive study of mobility functioning information in clinical notes: Entity hierarchy, corpus annotation, and sequence labeling. International journal of medical informatics, 147, 104351.

Timilsina, M., Buosi, S., Razzaq, M. A., Haque, R., Judge, C., & Curry, E. (2025). Harmonizing foundation models in healthcare: A comprehensive survey of their roles, relationships,

and impact in artificial intelligence's advancing terrain. Computers in biology and medicine, 189, 109925.

Valadkevičienė, D., Jatužis, D., Žukauskaitė, I., Danylaitė Karrenbauer, V., & Bileviciute-Ljungar, I. (2024). Revision of the brief international classification of functioning, disability and health core set for multiple sclerosis: a study of the comprehensive icf core set for multiple sclerosis with participants referred for work ability assessment. Journal of rehabilitation medicine, 56, jrm19671.

van der Meer, M., Falk, N., Murukannaiah, P. K., & Liscio, E. (2024). Annotator-Centric Active Learning for Subjective NLP Tasks. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (pp. 18537-18555). (EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference). Association for Computational Linguistics (ACL).

van Es, B., Reteig, L.C., Tan, S.C. et al. (2023). Negation detection in Dutch clinical texts: an evaluation of rule-based and machine learning methods. BMC Bioinformatics 24, 10.

Verkijk, S., & Vossen, P. (2025). Creating, anonymizing and evaluating the first medical language model pre-trained on Dutch Electronic Health Records: MedRoBERTa.nl. Artificial Intelligence in Medicine, 167, 1-13. Article 103148. Advance online publication.

Verkijk, S., & Vossen, P. . (2021). MedRoBERTa.nl: A Language Model for Dutch Electronic Health Records. Computational Linguistics in the Netherlands Journal, 11, 141–159.

Weissenbacher, D., Courtright, K., Rawal, S., Crane-Droesch, A., O'Connor, K., Kuhl, N., Merlino, C., Foxwell, A., Haines, L., Puhl, J., & Gonzalez-Hernandez, G. (2024). Detecting goals of care conversations in clinical notes with active learning. Journal of biomedical informatics, 151, 104618.

Wieland-Jorna, Y., van Kooten, D., Verheij, R. A., de Man, Y., Francke, A. L., & Oosterveld-Vlug, M. G. (2024). Natural language processing systems for extracting information from electronic health records about activities of daily living. A systematic review. JAMIA open, 7(2), ooae044.

Wiśniowska-Szurlej, A., Sozańska, A., Barrio, S. J., Sozański, B., Ceballos-Laita, L., & Hernández-Lázaro, H. (2024). ICF based comparison of musculoskeletal health in regions of Poland and Spain. Scientific reports, 14(1), 27671.

Workman, T. E., Ahmed, A., Sheriff, H. M., Raman, V. K., Zhang, S., Shao, Y., Faselis, C., Fonarow, G. C., & Zeng-Treitler, Q. (2024). ChatGPT-4 extraction of heart failure symptoms and signs from electronic health records. Progress in cardiovascular diseases, 87, 44–49.

World Health Organization. (2001). International Classification of Functioning, Disability and Health (ICF). Geneva: WHO.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E.H. (2016). Hierarchical Attention Networks for Document Classification. North American Chapter of the Association for Computational Linguistics.

Zhang, J., Yu, Y., Li, Y., Wang, Y., Yang, Y., Yang, M., & Ratner, A. (2021). WRENCH: A comprehensive benchmark for weak supervision.

Zhang, T., Cai, L., Li, J., Roberts, N., Guha, N., Lee, J., & Sala, F. (2025). Stronger than you think: Benchmarking weak supervision on realistic tasks.

Zhao, H., Chen, H., Ruggles, T., Feng, Y., Singh, D., & Yoon, H.-J. (2024). Improving text classification with large language model-based data augmentation. Electronics, 13, 2535.