A Cheap Lunch: Synthetic Annotation With Reduced Human Effort for Medical Text Mining

Shutao Chen, Piek Vossen
Vrije Universiteit Amsterdam
De Boelelaan 1103, 1085HV, Amsterdam, The Netherlands
realsusana.c@gmail.com, piek.vossen@vu.nl

Abstract

Electronic Health Records are rich resources of patient knowledge, among which about the functioning of patients as defined in the International Classification of Functioning (ICF) by the WHO. However, patient notes have yet to be explored as the knowledge is packaged in sometimes cryptic language exchanged between caretakers. Recent research started to use NLP techniques to extract this knowledge but this often requires laborious annotation. In this paper, we report on how the annotation can (partly) be done by a generative LLM, both for ICF categories that were previously manually annotated and for new ICF categories for which there was no annotation. We show that a domain specific encoder finetuned with both manual and synthetic annotations outperforms finetuning with just the manual or synthetic annotations on a dedicated test set that was adapted for the new categories with minimal manual effort. We also assessed the quality of the synthetic annotations of the training data. Our process shows how competitive text classifiers for medical text mining can be developed and extended to new categories with minimal manual effort by experts.

Keywords: Medical NLP, synthetic annotations, LLMs, Dutch

1. Introduction

Electronic Health Records (EHRs) are rich resources of patient knowledge that can be used for health care services. Such EHRs contain structured data on patients such as diagnosis, medication, lab results, but also notes written by caretakers to communicate the state of a patient to other caretakers. These notes contain a wealth of unexplored knowledge, often not captured by structured data. A specific type of information that falls outside the scope of structured data registration is the patient's functioning before, during, and after treatment. The World Health Organization (WHO) tries to standardize the interpretation of functioning through the International Classification of Functioning, Disability, and Health (ICF).[1] ICF is a classification schema that spans various categories of functioning (physical, mental, social) and also defines their levels in a uniform way.

Typically, ICF classifications are not included in structured patient data, but can be inferred from patient notes. As shown in (Kim et al., 2022; Meskers et al., 2022), this can be done using state-of-the-art encoder models that are finetuned with manually annotated training data. However, selecting and annotating the training data is laborious and costly. It requires developing elaborate guidelines and training annotators that, when properly trained, need to process lengthy medical notes in which most of the content is not relevant for the annotation. As reported in (Kim et al., 2022), only 5% of the sentences in medical notes give information on the functioning as defined in nine categories of ICF. As a side effect, annotators overlook relevant sentences as their attention drops reading notes, due to the mostly irrelevant text. Finally, extending a classification with other ICF categories would require reconsidering all the notes again to re-annotate each sentence for relevance for the new categories.

In this work, we therefore investigate to what extent the annotation can be done by a generative LLM and how easily the LLM can be prompted to also consider new ICF categories. The generic LLM is not adapted to the typical language we find in the notes and is also not pretrained for the task. In addition to assessing the quality of the synthetic annotations, we further investigate how useful these synthetic annotations are for finetuning an encoder model, combined with and without manual annotations. Finally, we investigate whether our method can also create annotations for new ICF categories. We describe a procedure for efficiently validating the test results for the new categories without reannotating the complete test set.

Our contributions are as follows.

- We created new training data and extended a dedicated test set with new ICF categories using both generative LLM and medical experts' validation.
- We developed a method to create ICF classifiers with reduced effort involving experts.
- We finetuned an encoder model with combinations of manual and synthetic labels, achieving

[1] https://www.who.int/standards/classifications/international-classification-of\-functioning-disability-and-health

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10353–10364
11–16 May 2026.

performance comparable to the model trained on manual annotations only, while outperforming the classifier trained solely on synthetically annotated data.

- Our finetuned encoder outperforms the classification by the prompted LLM directly, both for the old and new categories, showing that noisy annotations by the generative LLM still have added value to the manual data when given to the (faster) encoder.

Our ICF Classifier and code are freely available.[2]

2. Related work

The tracking of patient functional status is increasingly recognized as essential along with morbidity and mortality, yet much of this information appears only in free text clinical notes (Meskers et al., 2022). ICF provides a common language for documenting functional status such as body functions, activities (participation), and contextual factors, but routine ICF coding of narratives is labor-intensive and motivates automation. Foundational reviews and position papers argue for systematic ICF-based documentation and underline the gap between what is clinically recorded and what is structured, setting the stage for automated extraction from EHR text (Organization, 2007; Płaszewski and Płaszewski, 2025).

Early rule-based pipelines and classical ML established feasibility (Newman-Griffis and Fosler-Lussier, 2021; Thieu et al., 2021). Newman-Griffis and Fosler-Lussier (2021) reported strong multi-label performance on rehabilitation narratives; Thieu et al. (2021) showed high sequence-labeling scores for mobility entities. Modern neural approaches, especially transformers, improved the accuracy of sentence and note-level ICF tagging. In Dutch rehabilitation notes, the A-PROOF project (Meskers et al., 2022; Kim et al., 2022) achieved encouraging F1 on categories like *Walking* and *Emotional* functions, demonstrating that functioning signals can be retrieved from routine documentation. A recent systematic review of 37 systems found median macro F1 0.77 for extracting activities of daily living despite the heterogeneity of the data set, confirming the clinical utility (Wieland-Jorna et al., 2024).

For Dutch language deployments under privacy constraints, MedRoBERTa.nl, pretrained from scratch on hospital notes, consistently outperformed general Dutch encoders (BERTje, RobBERT) on inbuilt medical understanding and downstream clinical tasks, and was later prepared for compliant reuse (Verkijk and Vossen, 2025). Further in-domain pretraining (MedRoBERTa.nl-HAGA) improved lifestyle/status extraction (Muizelaar et al., 2024). These works support using locally runnable, in-domain encoders in care settings.

The downside of the above approaches is that they need a substantial amount of training and test data that need to be annotated manually. For example, the A-PROOF dataset is limited to nine ICF categories only and it took 12 months to manually annotate the training and test data (over 6K medical notes), involving a team of eight medical students, four medical staff members, three NLP students, and one NLP expert. Therefore, practical operation in medical contexts faces two persistent obstacles. First, label sparsity: only a small fraction of sentences in notes provide functioning evidence, so sentences are dominantly labeled as "None" (irrelevant), which degrades minority class learning. Secondly, data sharing between sites is restricted, which makes it practically impossible to aggregate data from different organizations. To address the latter, Fu et al. (2024) proposed FedFSA, a hybrid rule-based plus federated transformer framework that matched single site performance without pooling patient data. Active learning studies (e.g. Weissenbacher et al. (2024) and van der Meer et al. (2024) demonstrated sizable reductions in annotation effort for rare labels, although still with modest F1, underscoring the need for complementary supervision.

To reduce expert effort and expand coverage, recent studies increasingly use LLMs to generate provisional labels that are validated by clinicians or distilled into smaller models. A growing body of literature demonstrates that LLM-derived labels can effectively augment training data. For example, Hsu and Roberts (2025) prompted Llama-2 to label data and trained a BERT model that outperformed a supervised baseline; similarly, Guo et al. (2024) found that augmenting clinical text with GPT-generated labels increased downstream accuracy (with careful noise mitigation).

Reviews of LLMs on real EHR tasks (Du et al., 2024) note a broad reliance on zero or few-shot prompting, with many settings favoring finetuned compact models for privacy. Case studies report substantial reductions in annotation time (e.g. EHRmonize with GPT-4-class models (Matos et al., 2024)) and student models that match or surpass the LLM of the trainer when trained on LLM-derived labels (Hsu and Roberts, 2025; Lopez et al., 2025). Collectively, this literature positions LLMs as practical labeling aids rather than direct production classifiers in restricted environments.

Previous ICF extraction efforts often emphasized a narrow subset of body functions, risking omission of clinically significant context (e.g. pain, sleep,

[2] huggingface.co/icf17-domains
github.com/icf17-scripts
A-PROOF

family support, device-assisted mobility, handling stress, higher-level cognition, hearing, body position). ICF Core Sets and rehabilitation selections built by expert consensus and validation repeatedly elevate these domains (e.g. low back pain, stroke) and provide concise, clinically grounded targets for documentation ((Cieza et al., 2004; Geyh et al., 2004; Prodinger et al., 2018), ICF Research Branch guidance). Newman-Griffis et al. (2022) further caution that excluding environmental factors perpetuates information inequities. We follow this guidance by adding eight categories to the A-PROOF framework of nine categories and show that enriched labeling surfaces information previously buried under "None" without undermining legacy categories.

The eight newly introduced categories, ranging from physiological functions like Sensations of pain (B280) and Hearing functions (B230) to complex activities such as Handling stress (D240), were specifically selected by medical experts from Amsterdam UMC. This selection was informed by the ICF Core Set for community-dwelling older adults, as established in the consensus study by Rink et al. (2023). These categories are essential because they represent high-prevalence functional issues in the geriatric and rehabilitative populations, which the original A-PROOF domains did not fully capture.

Aligned with these trends, we present a study that pairs LLM-assisted annotation with a domain-adapted Dutch encoder, more specifically through GPT-4 class prompting supplying weak labels (especially for the eight new categories), which are next blended with existing manually annotated data to train a multi-label MedRoBERTa.nl classifier. This strategy provides evidence that sentence-level ICF information can be reliably coded; local, in-domain encoders suit privacy constrained clinical use; and LLM-assisted supervision efficiently expand label coverage with modest expert effort, particularly for contextual and psychosocial domains.

3. Method

We approach a sentence-level multi-label classification problem: each sentence from a rehabilitation note can be assigned zero, one, or multiple ICF category labels (from a set of 18, including a "None" label for no relevant finding). Our goal is to broaden the coverage from an original 10-category scheme to 18 categories by introducing new functional, cognitive, and psychosocial codes.

To be precise, the 9 original ICF categories (with Dutch label names) were: 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.

The 8 newly introduced categories are: 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.

Figure 1 provides an overview of the annotation and training pipeline. We first prompted GPT-4 to add 8 new ICF categories to the original data set that was manually annotated with 9 categories and None (Kim et al., 2022). Next, we prompted GPT-4 to annotate a new collection of notes with the 17 ICF categories or None. To optimize the prompt, we conducted a controlled pilot on an expert-reviewed validation set. The best results (61.5% exact match with gold labels) were achieved using a configuration that combined few-shot prompting, category definitions, and a low temperature (0.1). This setting was subsequently used to generate all synthetic annotations across both old and new ICF categories. The final prompt is given in the Appendix 13.

The manual and synthetic labels were used to derive different classifiers by finetuning MedRoBERTa.nl (Verkijk and Vossen, 2025). For training, sentences were tokenized and truncated to a maximum length (512 tokens) to fit the model input size. We used the AdamW optimizer (learning rate $4\times10\div5$) and a small batch size, finetuning for a single epoch over the large augmented dataset. Although one epoch might seem brief, our training set was quite extensive.

We trained and compared three model variants to assess the impact of the expanded data and labels.

(1) Baseline-10: a MedRoBERTa.nl model fine-tuned on only the original expert-annotated data for the initial 10 categories (including None). This represents the performance of the prior 10-code system without any synthetically annotated data or new labels.

(2) Synthetic-only: a model trained exclusively on the GPT-augmented data, that is, using the weakly labeled sentences (including the newly added notes) without the original human-labeled corpus. This condition tests whether GPT-generated labels alone could suffice for training a classifier from scratch on the expanded 18-category set.

(3) Model-18 (Manual + Synthetic): trained on the combined dataset that includes both the human-labeled examples and all GPT-labeled augmentations for all 18 categories.

By comparing these systems, we can quantify the contribution of human vs. LLM-provided anno-

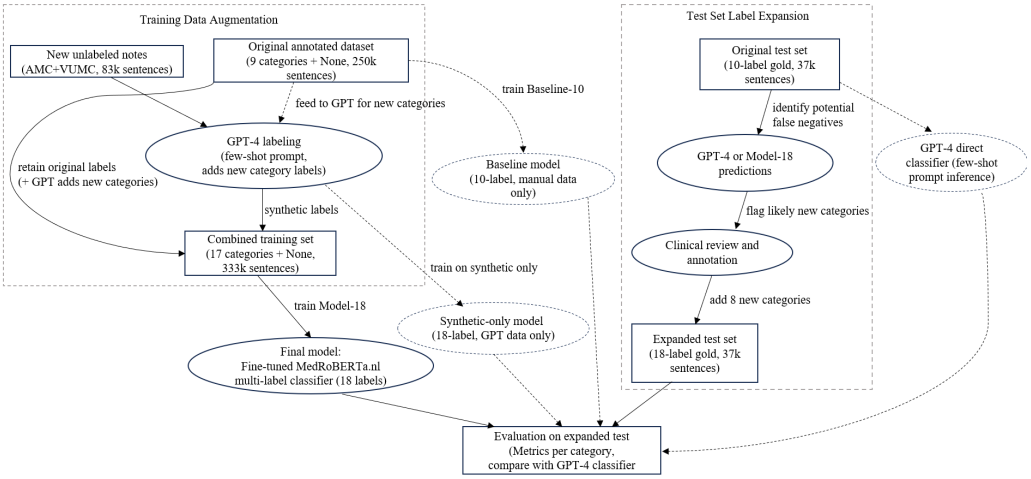

Figure 1: Overview of the experimental pipeline. The process begins with manual annotations (Baseline-10), followed by GPT-4-assisted data augmentation for new ICF categories, resulting in the combined Model-18 (manual + synthetic) classifier.

tations. In particular, we examine whether adding GPT-labeled training data improves the model over the baseline trained on limited manual data, and how the fully augmented 18-category model withstands a model that relies purely on synthetic labels. For completeness, we also evaluate GPT-4 itself as a few-shot classifier on this task to benchmark the finetuned models against a large LLM.

As a test set, we reused the 10-category manually annotated set from the A-PROOF project. Rather than re-annotating the entire corpus, we used the different models to identify candidate sentences for the eight new categories. These candidates were manually reviewed and labeled by a panel of three clinical experts.

4. The annotated data

4.1. The training data

The training datasets and their statistics are summarized in Table 1. The original dataset (used for "Baseline-10") consists of roughly 250K sentences (from 13,882 notes) that were manually labeled by clinicians for nine ICF categories and in case none applied with "None". We augmented this dataset by running GPT-4 on the same sentences: whenever a sentence was already labeled with one of the 9 legacy codes, we kept the expert labels and only added new labels that GPT-4 suggested. Thus in the original dataset, labels for the 9 old categories are expert (gold) labels.

The synthetic dataset ("Extension" in Table 1) contains about 83,370 sentences (8,609 notes) from recent clinical notes at AMC and VUMC (year 2023). These sentences were not manually labeled for ICF categories. Instead, we prompted GPT-4 on each sentence (with a few-shot prompt) to assign any of the 18 ICF codes. In other words, every sentence in this set receives a multi-label annotation purely from GPT-4. We filtered out senseless GPT outputs, and we further removed half of the sentences that GPT-4 labeled only as "None", this was done to enrich the data with positive examples and reduce the extreme None-ratio. The end result is 83K GPT-annotated sentences covering all 17 functional codes.

Table 1: Statistics on the sentences and notes used as training data with their annotation labels: M=manual, S=Synthetic, 9=old categories, 8=new categories, N=None and 18=all categories combined. Original are the notes from the original data set and Extension are sentences from new notes annotated by ChatGPT.

Dataset	Labels	Sentences	Notes
Original	M9 + S8 + None	250,008	13,882
Extension	S18	83,370	8,609
Combined	**M9+S18**	333,378	22,491

The final combined train data is the union of these two sets. We merged the 250K expert-derived sentences with the 83K synthetically-labeled sentences to obtain about 333,378 sentences (22,491 notes) in total. This combined set is annotated with all 18 ICF categories (17 functional codes + "None"). The original 250k sentences have expert labels for the 9 original categories and GPT labels for the 8 new ones, while the 83k newly selected sentences have GPT labels for both old and new categories.

As expected, a large majority (66.26%) of sentences in the combined training data carry the "None" label. The functional categories are relatively rare: e.g. the category *Sensations of Pain*

(B280) appears in only about 5-6% of sentences after augmentation, and most others are in the 1-3% range. The rarest categories (e.g. *Hearing functions B230*) are under 1%. Importantly, the augmentation preserved the relative distribution of the original 9 codes: dominant domains like *Respiration functions (B440)* and *Walking (D450)* remain among the most frequent.

4.2. The test data

For the test set, we reused the held-out set of 37,355 sentences from the A-PROOF project that had been manually annotated by clinical experts using the earlier 10-category scheme. Obviously, this test set lacks labels for the newly introduced categories, while most sentences (90.4%) have the label None. Instead of annotating the 3K notes in the test set all over again for the eight new categories, we performed a retroactive annotation. We first used our model trained on the augmented data to flag test sentences that might contain one of the new category concepts. Those candidate sentences were reviewed by a panel of clinicians, who assigned ground-truth labels for the eight new categories as needed (while preserving the original expert labels for the old categories).[3] This resulted in test data in which the new labels were added when predicted. [4]

Table 2: Sentence-Level statistics for the test data before and after validation

Version	Old 9 %	New 8 %	None %
Original	9.6	0	90.4
Labels Updated	10.84	6.97	84.84

The result is shown in Table 3. After adding the new labels, about 15.2% of sentences have at least one ICF label (up from 10.6% originally). In the updated test set, 10.8% of sentences carry one of the original 9 codes, 6.97% carry one of the 8 new codes, and 84.8% are "None". In practical terms, we captured new information, for example, pain (B280) occurred in 2.3% of sentences (15.4% of notes), family relationships (D760) in 1.3% of sentences, and sleep (B134) in 0.5%.

To evaluate the reliability of the newly-expanded manual annotations, the inter-annotator agreement (IAA) was assessed among three medical experts from Amsterdam UMC for a random selection of 155 sentences. The analysis resulted in an overall Average Percent Agreement of 79.65%. This reflects a moderate-to-substantial agreement, which aligns with the benchmark performance for the original 10-class research, despite the increased complexity of classifying 18 categories. The levels of agreement still varied strongly across the ICF domains. Highly objective functional domains, such as D410 (Changing basic body position) and D450 (Walking), achieved near-perfect agreement (up to 100%). In contrast, the agreement was markedly lower for psychosocial categories such as D240 (Handling stress) and D760 (Family relationships). These results underscore the high subjectivity and linguistic overlap in clinical notes when documenting cognitive or social functioning, a challenge previously documented in A-PROOF (Kim et al., 2022), where domains like "Exercise tolerance" also showed lower expert consistency.

Table 3: IAA Example Highlights by ICF Category

Level	ICF Categories	% Agreement
High	D410, D450	82.2% – 100%
Moderate	B280, B230	75.4% – 81.5%
Low	D240, D760, B164	< 70%

To further validate the reliability of the updated test set annotations, we performed a re-evaluation of potential false negatives. Specifically, we randomly sampled 405 sentences that were labeled as "None" in the gold-standard test set and were also predicted as "None" by both our final model and by GPT-4. This subset was again reviewed by clinical experts at VUMC to assess whether any functioning content might have been overlooked. The experts identified 3 out of 405 sentences (0.74%) as false negatives, i.e., they should have been assigned one or more ICF categories. This produces an estimated undetected false negative rate of approximately 0.0074, suggesting that the remaining error in "None" assignments is minimal.

These findings demonstrate that while some ICF codes remain challenging for manual validation, the 18-category pipeline maintains an expert-level reliability sufficient for clinical application, providing more granular data than the original 10-domain models.

5. Model results

Table 4 shows that the MedRoBERTa model trained on both manual and synthetic data (Manual + Synthetic 18) achieves the highest overall accuracy. Its micro average F1 score is 0.90, nearly matching the manual-only model's 0.91 and far exceeding the synthetic-only model's 0.80. Likewise, its

[3]The panel of clinicians consisted of three medical experts from Amsterdam UMC who were qualified for the task. All three were main contributors to the original A-PROOF project and were centrally involved in developing the formal ICF annotation guidelines used in the previous Dutch clinical NLP benchmark (Kim et al., 2022).

[4]Notably, all final test labels were provided or verified by humans – no GPT-generated labels were used in the test set – ensuring that our evaluation uses a high-quality expert reference standard.

Table 4: Overview of recall, precision and F1 per category and overall, split by four classifiers: Manual 10, Synthetic 18, Manual + Synthetic 18, and ChatGPT4o 18.

Category	Manual 10			Synthetic only 18			Manual + Synthetic 18			ChatGPT4o 18			
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Support
B1300	0.55	0.82	0.66	0.66	0.7	0.68	0.63	0.83	0.72	0.67	0.63	0.65	448
B140	0.58	0.88	0.7	0.52	0.87	0.65	0.66	0.89	0.76	0.73	0.6	0.66	64
B152	0.7	0.67	0.69	0.48	0.5	0.49	0.7	0.66	0.68	0.42	0.55	0.48	362
B440	0.64	0.85	0.73	0.73	0.36	0.48	0.65	0.85	0.73	0.6	0.54	0.57	1060
B455	0.37	0.44	0.4	0.33	0.21	0.25	0.34	0.4	0.36	0.36	0.26	0.3	398
B530	0.74	0.78	0.76	0.86	0.23	0.37	0.69	0.78	0.73	0.73	0.47	0.57	338
D450	0.75	0.65	0.7	0.78	0.42	0.55	0.8	0.63	0.71	0.74	0.6	0.66	368
D550	0.66	0.62	0.64	0.51	0.36	0.42	0.65	0.61	0.63	0.46	0.4	0.43	665
D840-D859	0.5	0.68	0.58	0.6	0.54	0.57	0.48	0.67	0.56	0.58	0.6	0.59	342
B280				0.91	0.69	0.79	0.93	0.83	0.88	0.78	0.79	0.78	847
B134				0.84	0.56	0.68	0.91	0.61	0.73	0.81	0.63	0.71	183
D760				0.88	0.49	0.63	0.92	0.54	0.68	0.6	0.61	0.6	472
B164				0.45	0.7	0.55	0.75	0.73	0.74	0.37	0.77	0.5	303
D465				0.67	0.6	0.63	0.79	0.55	0.65	0.48	0.57	0.52	162
D410				0.45	0.6	0.52	0.68	0.59	0.63	0.4	0.67	0.5	390
B230				0.81	0.28	0.42	0.92	0.33	0.49	0.76	0.4	0.52	37
D240				0.25	0.38	0.3	0.77	0.53	0.63	0.4	0.56	0.47	212
None	0.97	0.91	0.94	0.8	0.98	0.88	0.92	0.97	0.95	0.92	0.95	0.93	31692
Micro Avg	0.93	0.89	0.91	0.77	0.83	0.8	0.88	0.92	0.9	0.86	0.88	0.87	38343
Macro Avg	0.65	0.73	0.68	0.64	0.53	0.55	0.73	0.67	0.68	0.6	0.59	0.58	38343

: Highest
: Medium
: Lowest

macro average F1 is 0.68, essentially identical to the manual-only model's 0.68 and well above the synthetic-only model's 0.55 (and also above GPT-4o's 0.58). In other words, adding the synthetic annotations preserved the strong overall performance of the expert-trained model while significantly improving upon the synthetic-only baseline.

More specifically, the combined model's macro average recall (0.73) is substantially higher than the manual model's 0.65 (and higher than synthetic's 0.64), indicating that it identifies more relevant cases on average. Its macro precision (0.67) is lower than the manual model's 0.73 (but higher than synthetic's 0.53). Basically, the combined training provided a more balanced trade-off, allowing the model to catch more true positives with modest increases in false positives, leading to the best macro-F1 scores in Table 4.

In the original nine ICF categories, the combined model closely matches the performance of the manual-only (Baseline 10) model, showing that synthetically-annotated data did not degrade the expert-trained system. For example, Attention function (B140) saw its F1 rise from 0.70 (Manual 10) to 0.76 (Manual + Synthetic 18), driven by an increase in recall from 0.58 to 0.66 while precision remained high (0.88 to 0.89). Other original categories show similar patterns: *Energy level (B1300)* improved from F1 0.66 to 0.72, *Respiration functions (B440)* stayed high (0.73 vs 0.73), and *Walking (D450)* rose slightly (F1 0.70 to 0.71). In all of these cases, precision in the combined model remains comparable to the manual-only model while recall is equal or higher.

There are a few minor exceptions. For example, *Exercise tolerance (B455)* had low F1 under both models (0.40 for manual-only vs 0.36 for combined) because both precision and recall are low. Similarly, *Weight maintenance (B530)* saw a modest F1 drop (0.76 to 0.73) because recall fell (0.74 to 0.69), although precision stayed at 0.78. Overall, however, the combined model's F1 scores on all original categories are within a few points of the manual model's scores. Its macro F1 on the full 18-category test (0.68) is essentially the same as its macro F1 on just the original categories. In other words, incorporating the eight new labels did not decrease the accuracy of the original codes.

By contrast, the synthetic-only model generally performs worse on the original categories. It often has higher recall on some categories, for example, B530 recall is 0.86 vs 0.74 for manual, but this comes at the cost of very low precision. In nearly every original category, synthetic-only F1 is far below manual-only and combined. For instance, *B152 Emotional functions* drops from F1 0.69 (manual) to 0.49 (synthetic) and the combined model holds it at 0.68. In short, without any expert labels the synthetic model identifies more positive cases but also produces many false ones, so its overall accuracy lags behind.

On the eight categories introduced by the synthetic annotations, the combined model vastly outperforms both the synthetic-only model and GPT-4o. GPT-4o's prediction results are produced using a few-shot prompting approach (providing definitions and examples for each ICF code), for the purpose of understanding how a large language model without finetuning compares to our specialized model. As expected, the manual-only classifier cannot detect these new labels at all (its recall is essentially 0), so we focus on comparing the three 18-category models. Across these categories, the combined model achieves the highest F1 in almost every case. For example, *Sensations of pain (B280)* jumps to F1 0.88 under the com-

bined model, compared to 0.79 for synthetic-only and 0.78 for GPT-4o. *Higher-level cognition (B164)* shows a striking gain as well: F1 0.74 combined vs 0.55 synthetic vs only 0.50 for GPT-4o (with combined recall 0.75 vs synthetic 0.45). *Handling stress (D240)* shows the largest relative jump, combined model's F1 stands at 0.63, synthetic only 0.30, GPT-4o 0.47. One new label where GPT-4o slightly edges the combined model is *Hearing functions (B230)*. B230 is very rare (support of 37 instances) and difficult to interpret. Nevertheless, even here the combined model has a higher recall.

Analyzing precision and recall reveals the models' strengths and weaknesses. The synthetic-only model has moderate recall (0.64) but low precision (0.53), meaning it flags many false positives. The combined model strikes a balance, its recall (0.73) is much higher than manual-only, while its precision (0.67) is still substantially higher than the synthetic-only model. In effect, the combined model finds more true positives (fewer missed) while maintaining most of the precision of the manual model. GPT-4o's few-shot classifier produces roughly balanced precision and recall (both around 0.60) but both are lower than the finetuned combined model, so GPT-4o's F1 (0.58 macro) is about ten-point lower.

Overall, the results show that including synthetic annotations improved the detection of both common and rare categories. The combined model achieves much higher recall across the board than the manual model, and substantially higher precision than the synthetic-only model.

6. Error Analysis: MedRoBERTa-18 vs. GPT-4o

The quantitative performance gap identified in the results, most notably the ten-point macro-F1 margin between the combined MedRoBERTa model and GPT-4o, is driven by three primary qualitative error patterns. While GPT-4o demonstrates sophisticated general linguistic understanding, it lacks the domain-specific boundary awareness and clinical rigor required for precise ICF coding. This deficiency explains the lower precision and recall observed in the few-shot approach, as the general-purpose model often struggles to distinguish between everyday activities and formal functional categories. Our analysis divides these differences to three distinct failure modes.

Over-Generalization on Definition Boundaries

A notable error type involves over-prediction, where GPT-4 assigns a label to general lifestyle descriptions that do not meet the strict clinical criteria. For instance, GPT-4 labels "het bereiden van gezonde avondmaaltijden" (preparing healthy evening meals) as D550 Eating. In ICF context, "Eating" is strictly reserved for the physical act of intake (swallowing, chewing), whereas meal preparation is an instrumental activity. The finetuned MedRoBERTa model, having been exposed to these specific boundary cases during training, correctly predicted "None", showing a better understanding of category constraints.

Failure to Identify Clinical Nuance

GPT-4 often fails to map specific physiological recovery terms to ICF categories, frequently defaulting to "None". For example, "herstel op activiteit tenniswedstrijd 1x" (recovery after a tennis match) was missed by GPT-4 but correctly identified by MedRoBERTa as B455 Exercise tolerance functions. This indicates that GPT-4 fails to link the recovery mention to the underlying physiological function defined in the ICF framework. Similar patterns were observed with Dutch medical terms such as "vermoeidheid" (fatigue), which GPT-4 possibly often treated as general sentiment rather than the functional category B1300 Energy level.

Excessive Reliance on Keywords and Conflict on Multiple Themes

GPT-4 can fall into a keyword trap when a sentence contains multiple functional themes. In the case where a patient mentioned a distance (15km) followed by moving back in with their mother ("Woont sinds eerdere psychische klachten weer bij haar moeder in huis"), GPT-4 prioritized the distance metrics and predicted D450 Walking. The finetuned MedRoBERTa however correctly predicted the primary functional context as D760 Family relationships. This suggests that general-purpose LLMs struggle to weigh the clinical hierarchy of information when sentences are flexibly structured or are complex due to multiple themes.

7. Discussion

The performance across the original eight categories allows us to analyze the impact of synthetic data. We see that adding synthetically-annotated data to the manual data did not result in a significant improvement to using only the manual data for training, as the results are very similar. Training on only synthetically-labeled data results in a model that performs significantly lower. We can conclude that the synthetic labels did not add much, but also did not hurt the model.

The performance of GPT4o is very close to that of the MedRoBERTa.nl model finetuned with synthetically-labeled data only but performs much lower when manual annotations are added. This highlights that a domain-specific, finetuned model provides more consistent and comprehensive classification than a few-shot GPT-4 in this setting. The 10 point macro-F1 gap in favor of the augmented model suggests that GPT-4, even with its vast general knowledge, falls short on identifying the fine-

grained clinical details and multiple coexisting labels that our task-specific model handles. It appears that training a pretrained domain model (the encoder MedRoBERTa.nl) with manual annotations in addition to the synthetic annotations was crucial for learning the idiosyncrasy of how ICF concepts are reported in Dutch rehabilitation notes (e.g. subtle phrasing or context that indicates a functional problem). Such nuances could not fully be evoked from GPT-4 by a generic prompt.

Our work underscores the value of using LLM-generated annotations as a form of weak supervision to rapidly expand the coverage of clinical NLP systems. With no manually labeled data initially available for the eight new ICF categories, GPT-4 (with definitions and few-shot examples) was used to annotate thousands of sentences for those labels. Although the GPT-4 labels are noisy, their inclusion provided a critical signal to train the model on otherwise unseen classes. However, the results also show that the manual annotations of the original nine categories helps the model to detect the eight new categories. The performance of the model trained with manual and synthetically-annotated data outperforms the synthetic only model on the nine categories by a large margin (+13 percent points macro averaged F1). This implies that the MedRoBERTa.nl model learns to generalize better from the manual annotations in general and not just for the manually annotated categories.

From a clinical informatics perspective, the ability to broaden the ICF classification scope (from 9 up to 17 functional categories) is highly significant. We effectively surfaced substantial clinically meaningful content that was previously invisible to automated extraction. For example, including categories like *Sensations of Pain (B280)* and *Sleep functions (B134)* enabled the model to capture important symptoms (pain, insomnia) that the old scheme simply ignored; indeed, the model showed excellent performance on these domains, with pain-related sentences being detected with high precision and recall. Given that pain is a major driver of disability and was under-documented in structured data (Organization, 2007), this is a clinically meaningful improvement in what our NLP system can do. Likewise, adding *Higher-level cognitive functions (B164)* allowed identification of subtle cognitive impairment mentions, and adding *Family relationships (D760)* let the model detect references to the patient's support network. Notably, our expansion touches on social and contextual factors that are seldom identified but crucial for holistic rehabilitation care. This echoes calls in recent literature to include social factors and context in health documentation (Newman-Griffis et al., 2022). Our results show that an NLP model can be feasibly extended to cover these components provided that part of the data is manually annotated by experts to leverage the synthetic annotations, an important step toward more equitable and comprehensive health information systems.

8. Conclusion

We report on the ability of a generative LLM, GPT-4, to annotate Dutch clinical notes with WHO-ICF categories on patient functioning. We compared few-shot prompting of GPT-4 with the encoder model MedRoBERTa.nl that is finetuned with manual, synthetic (from GPT-4), and combined annotations. We observed that finetuning with manual annotations outperforms GPT-4 and adding synthetic annotations did not hurt the performance. Performance on new categories benefits too from the original manual annotations, even though there are no manual annotations for the new. Hence, MedRoBERTa.nl generalized knowledge from the manual annotations that is also useful for extending the classifier to new categories.

GPT-4's performance is still decent given it had no direct training on our dataset: achieving $\tilde{0}.58$ macro F1 out-of-the-box on a complex multi-label task demonstrates the power of large language models' latent understanding. With more sophisticated prompting or as part of an ensemble, GPT-4 could potentially be used to further improve results (for example, by additionally annotating training examples or catching cases which the finetuned model misses). However, at present, the finetuned augmented model clearly produced the best performance across the board, combining competitive precision and recall for both common and rare ICF categories. These results validate our approach of augmenting limited expert annotations with GPT-generated labels: it enabled a substantial expansion of the classifier's scope (from 10 to 18 codes) without any loss in accuracy. Notably, the finetuned model is also magnitudes faster than ChatGPT and has a much lower footprint.

9. Limitations

We have covered 17 ICF categories in Dutch EHR notes. Obviously, ICF has many more categories that could have been covered and that could be important for other medical services. It is not given that our approach will work for other ICF categories. Much depends on how these are expressed in the notes. Our notes come from two hospitals that were also used to pretrain MedRoBERTa.nl. Other hospitals may have different styles of describing functioning for which our models may not work well. Furthermore, our classifiers are limited to Dutch notes and cannot handle notes in other languages. In future research, multilingual models my be used

to transfer the annotations to other languages. Finally, other caretakers in the Dutch health system also make notes on their patients. Their practice is not included, which creates an omission of patient data that extends beyond the hospital context.

Another limitation of this study is the persistent challenge of class imbalance, characterized by the overdominance of the "None" class and a sparse distribution of specific clinical ICF codes. Despite the implementation of mitigation strategies to balance the dataset, a notable gap remains between the micro F1 and macro F1 scores. This imbalance indicates that while the model performs robustly on high-frequency categories, it continues to struggle with learning the nuanced patterns of minority classes. This phenomenon stems largely from real clinical documentation, where certain functional domains (such as mobility) are recorded with much greater frequency than others, such as specific psychological demands. Future work might focus on addressing this long-tail distribution problem through more advanced data augmentation techniques.

10. Ethics

Medical data are very sensitive when it comes to privacy. Special care was taken to anonymize MedROBERTa.nl and also this research was carried out in a secure environment that was approved by the privacy officer. The hospital has a specific contract with Microsoft to use ChatGPT under strict privacy conditions.

11. Bibliographical References

A. Cieza, G. Stucki, M. Weigl, P. Disler, W. Jäckel, S. van der Linden, N. Kostanjsek, and R. de Bie. 2004. Icf core sets for low back pain. *Journal of Rehabilitation Medicine*, 44 Suppl:69–74.

X. Du, Y. Wang, Z. Zhou, Y.-W. Chuang, R. Yang, W. Zhang, X. Wang, R. Zhang, P. Hong, D. Bates, and L. Zhou. 2024. Generative large language models in electronic health records for patient care since 2023: A systematic review. *medRxiv*. Preprint.

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

S. Geyh, A. Cieza, J. Schouten, H. Dickson, P. Frommelt, Z. Omar, N. Kostanjsek, H. Ring, and G. Stucki. 2004. Icf core sets for stroke. *Journal of Rehabilitation Medicine*, 44 Suppl:135–141.

F. Gilardi, M. Alizadeh, and M. Kubli. 2023. Chatgpt outperforms crowd workers for text annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Y. Guo, A. Ovadje, M. A. Al-Garadi, and A. Sarker. 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.

Stephen J Halpin, Claire McIvor, Gemma Whyatt, Anastasia Adams, Olivia Harvey, Lyndsay McLean, Christopher Walshaw, Steven Kemp, Joanna Corrado, Rajinder Singh, et al. 2021. Postdischarge symptoms and rehabilitation needs in survivors of covid-19 infection: A cross-sectional evaluation. *Journal of medical virology*, 93(2):1013–1022.

Enshuo Hsu and Kirk Roberts. 2025. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific Reports*, 15(1):8241.

Jenia Kim, Stella Verkijk, Edwin Geleijn, Marieke van der Leeden, Carel Meskers, Caroline Meskers, Sabina van der Veen, Piek Vossen, and Guy Widdershoven. 2022. Modeling dutch medical texts for detecting functional categories and levels of covid-19 patients. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4577–4585.

I. Lopez, A. Swaminathan, K. Vedula, S. Narayanan, F. Nateghi Haredasht, S. P. Ma, A. S. Liang, S. Tate, M. Maddali, R. J. Gallo, N. H. Shah, and J. H. Chen. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Rehab Mahmoud, Nashwa El-Bendary, Hoda MO Mokhtar, and Aboul Ella Hassanien. 2014. Icf based automation system for spinal cord injuries rehabilitation. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)*, pages 192–197. IEEE.

J. Matos, J. Gallifant, J. Pei, and A. Wong. 2024. Ehrmonize: A framework for medical concept abstraction from electronic health records using large language models. *arXiv preprint arXiv:2407.00242*. Preprint.

Carel GM Meskers, Sabina van der Veen, Jenia Kim, Caroline JW Meskers, Quirine TS Smit, Stella Verkijk, Edwin Geleijn, Guy AM Widdershoven, Piek TJM Vossen, and Marike van der Leeden. 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.

H. Muizelaar, M. Haas, K. van Dortmont, P. van der Putten, and M. Spruit. 2024. Extracting patient lifestyle characteristics from dutch clinical text with bert models. *BMC Medical Informatics and Decision Making*, 24(1):151.

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

Denis Newman-Griffis and Eric Fosler-Lussier. 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.

World Health Organization. 2007. *International Classification of Functioning, Disability, and Health: Children & Youth Version: ICF-CY.* World Health Organization.

Maciej Płaszewski and Karol Płaszewski. 2025. Icf-based assessment of functioning–state-of-the-art and challenges: A user's perspective.

B. Prodinger, A. Tennant, and G. Stucki. 2018. Standardized reporting of functioning information on icf-based common metrics. *European Journal of Physical and Rehabilitation Medicine*, 54(1):110–117.

A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. 2020. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29:709–730.

Laura Rink, Johanna Tomandl, Sonja Womser, Thomas Kühlein, and Maria Sebastião. 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.

Thanh Thieu, Jonathan Camacho Maldonado, Pei-Shu Ho, Min Ding, Alex Marr, Diane Brandt, Denis Newman-Griffis, Ayah Zirikly, Leighton Chan, and Elizabeth Rasch. 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.

M. van der Meer, N. Falk, P. K. Murukannaiah, and E. Liscio. 2024. Annotator-centric active learning for subjective nlp tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

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

Y. Wieland-Jorna, D. van Kooten, R. A. Verheij, Y. de Man, A. L. Francke, and M. G. Oosterveld-Vlug. 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.

12. Language Resource References

Verkijk, Stella and Vossen, Piek. 2025. *Creating, anonymizing and evaluating the first medical language model pre-trained on Dutch Electronic Health Records: MedRoBERTa. nl*. Elsevier.

13. Appendix

Synthetic Labels Generation - Prompt for GPT4 Annotation

The following prompt template was used to generate annotations using the GPT-4 model. The prompt consists of a system role definition, detailed category descriptions, and few-shot examples to ensure high-quality, structured output.

System Instructions:

You are an annotation assistant. You will receive sentences from a Dutch clinical note. The sentences are already split in a list. For each sentence in the list you should choose zero, one, or more ICF categories. For each sentence, if the category 'None' is chosen, no other category should be added. Return a JSON array of objects.

Category Definitions:

"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",

"B152 Emotional functions": "Specific mental functions related to the feeling and affective components of the processes of the mind, "B440 Respiration functions": "Functions of inhaling air into the lungs, the exchange of gases between air and blood, and exhaling air",

"B455 Exercise tolerance functions": "Functions related to respiratory and cardiovascular capacity as required for enduring pjysical exertion",

"B530 Weight maintenance functions": "Functions of maintaining appropriate body weight, including weight gain during the development period",

"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. Include: walking short or long distances; walking on different surfaces; walking on different surfaces; walking around obstacles",

"D550 Eating": "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. Exclude: ingestion functions (chewing, swallowing, etc.), appetite",

"D840-D859 Work and employment": "apprenticeship (work preparation); acquiring, keeping and terminating a job; remunerative employment; non-remunerative employment",

"B280 Sensations of pain": "Sensation of unpleasant feeling indicating potential or actual damage to some body structure",

"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",

"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",

"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",

"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",

"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",

"B230 Hearing functions": "Sensory functions relating to sensing the presence of sounds and discriminating the location, pitch, loudness and quality of sounds",

"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",

"None": "Does not belong to any of the ICF categories in the list".

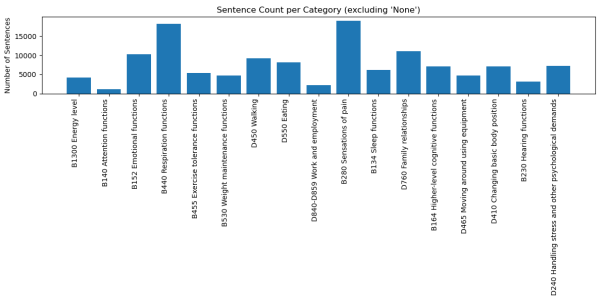

Figure 2: Final Augmented Train Data Label Distribution (Excluding None)

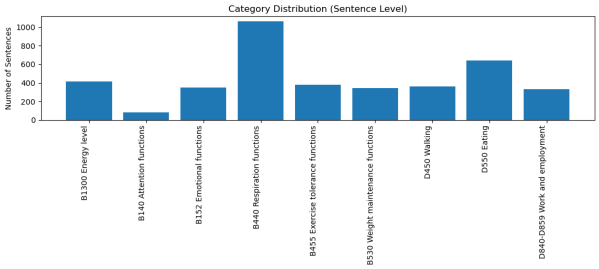

Figure 3: Original Test Set 10-Category Sentence-Level Statistics (Excluding None)

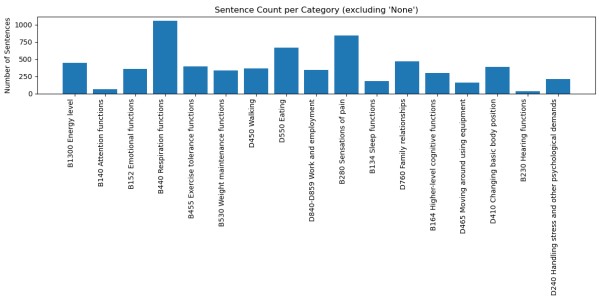

Figure 4: Updated Test Set 18-Category Sentence-Level Statistics (Excluding None)

	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 5: MedRoBERTa 18-Category Classification Report

	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: Few-Shot GPT-4o 18-Category Classification Report

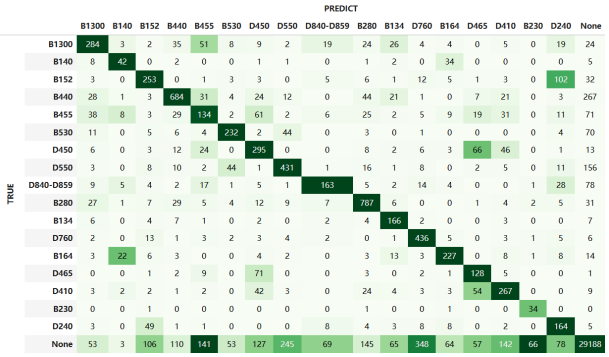

Figure 7: MedRoBERTa 18-Category Confusion Matrix

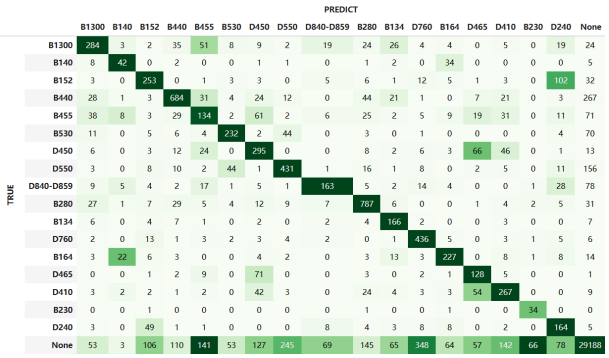

Figure 8: Few-Shot GPT-4o 18-Category Confusion Matrix