

Evaluating the Performance of LLMs in ICF Classification: Insights from Medical and General Models*

Feten Skhiri Gabbouj^{1,2}, Matei Mancas¹, Mathias Blandeau² and Laura Wallard²

Abstract—In the medical field, text data comprising personal anecdotes and detailed patient insights are often underutilized due to their unstructured nature and variability among clinicians. However, recent advances in Large Language Models (LLMs) present an opportunity to harness this data effectively. This paper explores the use of the International Classification of Functioning, Disability, and Health (ICF) framework recommended by the World Health Organization (WHO), which offers a holistic approach considering personal and environmental factors along with impairments, to structure textual descriptions systematically. The study investigates the application of medically fine-tuned LLMs, such as MedAlpaca and Meditron, for automated ICF creation, comparing their efficiency in processing real medical cases from two distinct contexts: rehabilitation and intensive care units. Additionally, we benchmark medical LLMs against general-purpose LLMs, including ChatGPT and Claude, to assess whether specialized models truly offer an advantage in medical classification tasks. Preliminary findings indicate that while medical LLMs show potential for ICF classification tasks, they may not necessarily outperform general-purpose models, as the complexity of ICF requires a deeper level of contextual understanding.

International classification of functioning, disability and health (ICF), Intensive care, large language models, AI, medical, LLM, Rehabilitation

I. INTRODUCTION

In the medical field, various data types are employed for diagnosing and categorizing patients. Among these, text data, often overlooked, serves as a valuable resource, capturing diverse patient insights, including personal anecdotes and nuanced clinical details. Despite its richness, text data presents challenges due to its inherent unstructured nature and significant inter- and intra-clinical variability. Recent advances in attention-based large language models (LLMs) [30] have opened new possibilities for leveraging textual data more effectively. However, achieving reliable medical text processing requires both well-trained models and a structured framework for data standardization.

To address this, the International Classification of Functioning, Disability, and Health (ICF) provides a comprehensive framework that transcends a purely symptom-based

approach. Instead, it integrates personal and environmental factors alongside impairments, evaluating their impact on an individual’s activities and participation in daily life [29]. Aligning unstructured patient descriptions with the ICF checklist and its hierarchical classification system ensures systematic organization of medical information, enhancing accessibility, interoperability, and clinical decision-making. However, manually filling out the ICF remains time consuming for specialists, limiting its widespread use in practice.

This study explores the potential of pre-trained medical LLMs to (semi)automatically classify patient descriptions into ICF categories, reducing the burden on clinicians. Additionally, we extend our investigation by benchmarking medical LLMs against general-purpose LLMs, assessing whether domain-specific training provides a tangible advantage in ICF related tasks. Given the complexity of ICF, which requires a deep understanding of medical context, reasoning, and structured classification, it is not evident that medical LLMs will always outperform general models. This comparison aims to determine whether specialized models justify their use in medical applications or whether general purpose LLMs, with broader training data, can perform equally well. The paper is organized as follows: introducing the scientific state of the art, describing the methodology, presenting results on various LLMs, and conclusion.

II. LITERATURE REVIEW

A. International Classification of Functioning, Disability, and Health (ICF)

Disability has traditionally been seen only through the lens of a person’s medical traits, known as the “medical model” [22], which focuses on medical treatment to improve their condition. In contrast, the “social model” [17] views disability as arising from external barriers and environmental constraints. This model suggests that improving the conditions of people with disabilities requires political and economic actions. The current approach, called the “biopsychosocial model,” combines both previous models, considering disability as a result of biological, personal, and social factors. This integrated perspective has spurred international discussions on accessibility issues. The World Health Organization (WHO) promotes the ICF framework [24], a standardized system for assessing functioning and disability across disciplines. The ICF categorizes human functioning into Body Structures and Functions, Activities, Participation, and Environmental Factors [14], [27]. It provides a structured way to evaluate patients’ conditions,

*This work was supported by the Hauts-de-France region and the Walloon region, Belgium

¹Feten Skhiri Gabbouj and Matei Mancas are with ISIA lab - Université de Mons (UMONS), Bd Dolez 31, 7000, Mons, Belgium feten.skhirigabbouj@umonts.ac.be; matei.mancas@umonts.ac.be

²Feten Skhiri Gabbouj, Mathias Blandeau and Laura Wallard are with LAMIH UMR 8201, Université Polytechnique Hauts-de-France, CNRS, F-59313, Valenciennes, France mathias.blandeau@uphf.fr; laura.wallard@uphf.fr

guiding rehabilitation, intervention planning, and long-term monitoring.

The ICF employs a hierarchical coding system to classify different aspects of health and disability. It consists of four main domains at Level 1: (b) Body Functions (e.g., psychological functions), (s) Body Structures (e.g., organs, limbs), (d) Activities and Participation (e.g., daily tasks, social involvement), and (e) Environmental Factors (e.g., social and physical surroundings). At Level 2, each domain is further divided into chapters (e.g., **b2** for Sensory Functions, **s1** for Nervous System Structures, **d4** for Mobility, **e5** for Policies and Services). The classification extends up to Level 5, offering progressively finer detail (e.g., **b210** for Seeing Functions, **b2100** for Visual Acuity, **b21000** for Binocular Distance Vision). This structure ensures precise documentation and analysis.

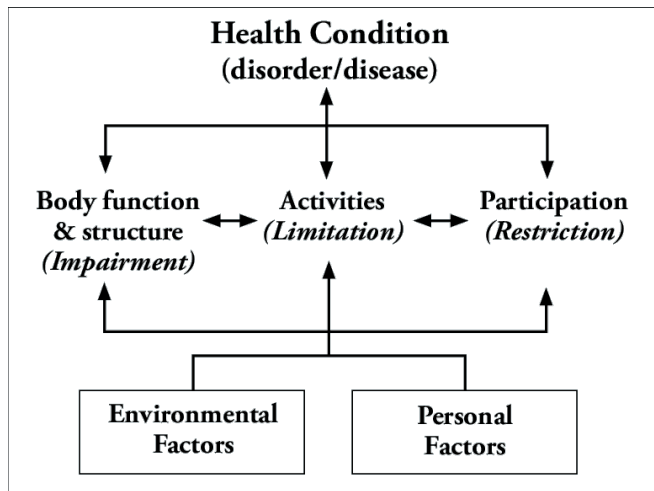


Fig. 1. The ICF model and its components: Body functions and structures, activity and participation and personal and environmental factors

The ICF is widely applied in healthcare, rehabilitation, research, and policy. Clinicians use it for patient assessment, intervention planning, and tracking progress [13], while researchers employ it to standardize data collection and compare outcomes across studies [28]. However, its complexity and extensive coding system (1,424 categories) [25] pose challenges. Proper training is required to ensure consistency, as variations in coding between clinicians can reduce data reliability and impact decision-making. Additionally, documentation can be time-consuming, making implementation difficult in busy healthcare settings.

B. Large Language Models and Knowledge Extraction in Medical Applications

The integration of electronic health records (EHRs) into clinical practice has transformed healthcare documentation but remains a challenge due to the complexity and time-consuming nature of data entry and retrieval [12]. Recent advancements in natural language processing, particularly through hybrid methods and large language models (LLMs), offer promising solutions. Biomedical text synthesis tech-

niques are being explored to enhance EHR usability, addressing issues such as fragmented interfaces and the high cognitive burden on clinicians, especially in emergency settings where they must rapidly synthesize patient histories [21]. Tools like MedKnowts unify documentation and search processes by providing concept-oriented patient record synthesis, reducing the workload associated with clinical documentation [21]. Furthermore, far-field speech recognition has been proposed for automatic International Classification of Diseases (ICD) coding, demonstrating real-time transcription capabilities using acoustic signal processing and recurrent neural networks [19], [10].

The advent of LLMs such as GPT-4, MedAlpaca, Meditron, Medicine Chat, Dr. Samantha, and BioMistral has further expanded AI applications in medicine. MedAlpaca improves medical question and answer application [2], [5], while Meditron variants support clinical decision making, particularly in low-resource settings [3], [8]. Medicine Chat improves patient-provider communication [7], and Dr. Samantha integrates support for both physical and mental health [26]. BioMistral aids clinical decision support, automates medical tasks, and supports medical research through open-source adaptability [4], [1].

Despite these advancements, the application of LLMs to the International Classification of Functioning, Disability, and Health (ICF) remains underexplored. Some studies indicate potential, such as the use of ChatGPT-4 in generating rehabilitation prescriptions and ICF codes, which were evaluated by licensed PMR clinicians in a stroke case study [33]. Although minor discrepancies were observed, the majority of treatment plans and ICF codes were accurately generated, demonstrating the potential of LLMs in clinical rehabilitation. This highlights an opportunity for further research into how LLMs can improve the documentation and application of ICF, paving the way for more efficient and accurate clinical workflows.

III. METHODOLOGY

Our methodology involves prompting LLMs to generate ICF classifications from text descriptions and comparing their outputs to ground-truth labels to assess performance. We evaluated medical LLMs both quantitatively and qualitatively, identifying the best performing models and subsequently comparing them to general-purpose LLMs.

A. Datasets and preprocessing

The medical field from which the data originates greatly influences the text data reported by the clinicians. To capture this variability, we decided to work with medical information from two distinct fields.

- **Rehabilitation dataset:** Rehabilitation professionals in centre hospitalier de valencienne provided real patient data, offering standardized descriptions for 6 of their patients. These descriptions included key details such as reasons for hospitalization, medical and surgical history,

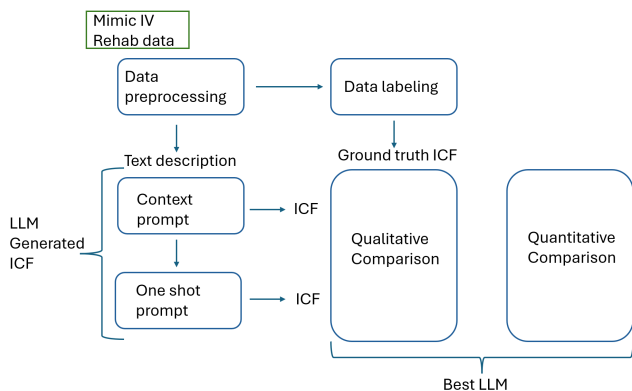


Fig. 2. The methodology used in this study

allergies, current treatments, lifestyle factors (family situation, professional status, hobbies), and initial clinical examinations. These descriptions were in French.

To ensure ethical compliance and protect patient privacy, all data was anonymized by removing identifiable details such as names, birth dates, physician names, and treatment locations. In addition, medical and family history were excluded as they may not accurately reflect the current condition of the patient and are typically not part of the ICF framework. Finally, a medical expert manually assigned ground-truth ICF classifications to each patient, serving as a reference to evaluate LLM-generated classifications.

- **Mimic IV dataset** : We used 10 patient descriptions from the Mimic-IV data set [20] to supplement our analysis. Mimic-IV is a publicly available database derived from electronic health records that contains detailed patient data, such as measurements, diagnoses, treatments, and procedures. Unlike rehabilitation data, which include rich personal and environmental factors, MIMIC-IV focuses on intensive care, providing a contrasting medical domain to evaluate LLM performance and adaptability in diverse clinical contexts.

The preprocessing involved generating structured text summaries for each patient, including medical status, prescribed treatments, hospitalization reasons, and demographic details (age, gender, and marital status). This allowed for uniformity in comparison with the rehabilitation dataset, ensuring consistency in the extracted features.

For labeling and providing ground truth, each patient description was manually annotated by a medical expert. Due to the complexity of ICF classification, this process was time-consuming and labor-intensive. Given the challenges in obtaining labeled data and the difficulty of manual ICF annotation, the sample size remained small. Labeling the entire Mimic-IV dataset was not feasible, as the task requires substantial expertise and effort from medical professionals.

All procedures involving human data were conducted in accordance with the Declaration of Helsinki (1975, revised

2000). For the MIMIC-IV dataset, data usage was approved by the Institutional Review Boards of the Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology.

B. Prompt engineering

Five medical LLMs were selected for our study based on their relatively small sizes, each with 7 billion parameters, except for MedAlpaca, which has 13 billion parameters: Meditron [32], MedAlpaca [18], Dr. Samantha [26], Medicine Chat [16], and BioMistral [31]. Smaller LLMs are more practical for real-world applications, as they can be deployed on standard hardware, unlike larger models such as Med42 [15], which, despite its high accuracy, requires substantial resources for deployment. For testing, we used LM Studio [6], a versatile platform that enabled efficient evaluation of model performance through both quantitative and qualitative assessments.

To ensure a comprehensive comparison, we applied three quantization versions (Q2, Q4, and Q8) to each of the five models, resulting in 15 LLMs. Medical LLMs were chosen for their fine-tuning on medical datasets, making them well-suited for ICF classification tasks, which are heavily reliant on medical terminology.

Our testing involved presenting patient descriptions to each LLM and requesting an ICF classification. Given the diversity in model training data and ICF classification methods, we used a structured approach to maintain consistency across tests. Three types of prompts were used:

- **Basic Prompt**: A simple request for ICF classification without background or additional context.
- **ICF structure Prompt**: A detailed prompt explaining the ICF framework, including all possible ICF classes, their codes, and hierarchy.
- **One shot Prompt**: A prompt that first provides an example of a classified patient description, then asks the LLM to classify another patient description, with the ground truth included to guide learning. Additionally, our prompts required LLMs to classify the reason for each medication, such as identifying the relevant ICF class for sensory and pain in cases where medication was prescribed for pain relief. This provided insight into the LLMs' ability to handle medical terminology and classify medication reasons correctly.

Our study focused on the first three levels of the ICF checklist (cf: II-A), specifically testing at Level 2 (broader classifications) and Level 3 (more detailed classifications) to evaluate model performance at different levels of granularity. This structure allowed us to assess how well the models could adapt to varying levels of detail in medical classifications.

In the second phase of testing, we compared the best-performing medical LLMs from the first round to general-purpose LLMs, including ChatGPT [23], Claude [11], and Llama [9]. For these models, we used only the Basic Prompt, simulating a more typical use case where users (such

as doctors) might interact with LLMs without additional structure or examples.

By testing under these controlled conditions, we aimed to evaluate the models’ ability to perform ICF classification consistently and assess their potential for improvement through iterative testing. This methodology allowed for a thorough analysis of the models’ strengths and weaknesses, ensuring that any observed performance differences could be attributed to the models themselves, rather than variations in input structure.

IV. EVALUATION AND RESULTS

A. Medical LLM comparison

Given the substantial number of LLMs involved in this study, our initial objective was to streamline the selection process by identifying and retaining the top performing models. This was achieved through both qualitative and quantitative comparisons, allowing for a thorough assessment of their capabilities.

1) Qualitative Comparison : We evaluated several parameters of the LLMs (Table I) to assess both their computational efficiency and structural integrity. The evaluation covered the following aspects:

- Speed: Measured by generation time (gentime) and the number of tokens produced per second (token/s), which reflects the LLM’s efficiency. These metrics were obtained using Windows 11 software, a 13th Gen Intel Core i9-13980HX CPU, 32GB RAM, and an NVIDIA GeForce RTX 4080 Laptop GPU.
- Structure: A binary variable indicating whether the LLM follows the ICF structure, organizing classes under the four main categories: Body Functions, Body Structures, Activities and Participation, and Environmental Factors.
- Justification: A binary variable assessing whether the LLM provides relevant justification for its classifications, enhancing transparency and reliability (e.g., explaining mobility limitations).
- Hallucination: A binary variable indicating whether the LLM generates information not supported by the provided data, which is critical for maintaining accuracy.

A qualified professional manually evaluated the structure, justification, and hallucination of the ICF classifications provided by the LLMs. This assessment focused on checking if the LLM followed the correct ICF structure, the validity of its justifications, and identifying any hallucinations instances where the LLM generated unsupported information.

Table I displays the average metrics of the first four tests for each LLM, both before and after correction. The generation time is provided in seconds and structure, justification and hallucinations are averaged, which will give a score between 0 and 1.

LLM	gentime	token/s	struct.	justif.	hallucin.
MeditronQ2	139.502	8.4	0	0.125	1
MeditronQ4	144.15	6.4	0.125	0	0.875
MeditronQ8	172.733	5.2	0.125	0.25	0.875
MedicinechatQ2	82.437	8.4	0.5	0.75	0.125
MedicinechatQ4	43.632	7.5	0.5	0.375	0.875
MedicinechatQ8	74.167	5.2	0.75	0.375	0.375
MedalpacaQ2	32.281	5	0	0	0.625
MedalpacaQ4	129.442	4.1	0.375	0.25	0.125
MedalpacaQ8	376.856	2.8	0.25	0	0.25
DrsamanthaQ2	64.106	9.4	0.375	0.5	0.625
DrsamanthaQ4	64.216	7.4	0.5	0.5	0.375
DrsamanthaQ8	90.457	5.2	0.5	0.375	0.5
BiomistralQ2	28.736	10.3	0	0	0.75
BiomistralQ4	51.82	7.6	0	0.4	0.8
BiomistralQ8	15.551	5.3	0.125	0	0.625

TABLE I

MEDICAL LLMs AVERAGE QUALITATIVE METRICS.

2) Quantitative Comparison : The quantitative comparison focused on assessing the accuracy of ICF classifications generated by various LLMs. This was done by comparing the LLM-generated classifications to those created by a professional using key metrics:

- True Positive (TP): An ICF class that is present in both the correct ICF classification and the LLM-generated ICF with both text AND class code.
- False Negative (FN): An ICF class that is present in the correct ICF but missing in the LLM-generated ICF.
- False Positive (FP): An ICF class that is present in the LLM-generated ICF but not in the correct ICF.

Given the extensive number of ICF classes and the likelihood that many would not appear in most patient descriptions, we did not calculate True Negatives (TN). Including TNs could have skewed our calculations and introduced bias. Based on the TP, FN, and FP values, we calculated several performance metrics for each LLM, including:

- Accuracy: The proportion of correctly identified ICF classes with their labels among the total predicted classes.
- Precision: The proportion of positive identifications that were actually correct.
- Recall: The proportion of actual positives that were correctly identified by the LLM.
- F-Score: The harmonic mean of precision and recall, providing a single measure of a model’s performance.

To guide the models in understanding the ICF structure, we initially provided a prompt containing a detailed explanation of the ICF framework (ICF structure). We then evaluated the models on both Level 2 and Level 3 ICF classifications in Table II. Following this, we conducted the same tests using a prompt that included, in addition to ICF structure, an example of a correct ICF applied to one patient descriptions (Table III).

Model	icf level	accuracy	precision	recall	fscore
MeditronQ2	2	0.068	0.214	0.083	0.120
	3	0.055	0.500	0.055	0.100
MeditronQ4	2	0.068	0.214	0.083	0.120
	3	0.083	0.500	0.083	0.142
MeditronQ8	2	0.068	0.214	0.083	0.120
	3	0.055	0.500	0.055	0.100
MedalpacaQ2	2	0.105	1.000	0.105	0.190
	3	0.105	1.000	0.105	0.190
MedalpacaQ4	2	0.068	0.214	0.083	0.120
	3	0.055	0.500	0.055	0.100
MedalpacaQ8	2	0.090	0.250	0.111	0.153
	3	0.096	0.517	0.133	0.168
MedicinechatQ2	2	0.068	0.214	0.083	0.120
	3	0.055	0.500	0.055	0.100
MedicinechatQ4	2	0.068	0.214	0.083	0.120
	3	0.055	0.500	0.055	0.100
MedicinechatQ8	2	0.055	0.214	0.080	0.120
	3	0.055	0.500	0.055	0.100
DrsamanthaQ2	2	0.068	0.214	0.083	0.120
	3	0.100	0.750	0.105	0.181
DrsamanthaQ4	2	0.097	0.264	0.145	0.175
	3	0.128	0.750	0.133	0.226
DrsamanthaQ8	2	0.125	0.300	0.173	0.219
	3	0.101	0.750	0.105	0.183
BiomistralQ2	2	0.068	0.214	0.083	0.120
	3	0.055	0.500	0.055	0.100
BiomistralQ4	2	0.068	0.214	0.083	0.100
	3	0.055	0.500	0.055	0.100
BiomistralQ8	2	0.099	0.269	0.145	0.178
	3	0.055	0.500	0.055	0.100

TABLE II

CLASSIFICATION PERFORMANCE BEFORE EXAMPLE

Model	icf level	accuracy	precision	recall	fscore
MeditronQ2	2	0.052	0.166	0.066	0.095
	3	0.06	0.5	0.06	0.107
MeditronQ4	2	0.052	0.166	0.066	0.095
	3	0.16	0.75	0.185	0.273
MeditronQ8	2	0.05	0.166	0.066	0.095
	3	0.06	0.5	0.06	0.107
MedalpacaQ2	2	0.052	0.166	0.066	0.095
	3	0.122	1	0.122	0.218
MedalpacaQ4	2	0.105	0.25	0.177	0.190
	3	0.173	0.380	0.247	0.288
MedalpacaQ8	2	0.185	0.358	0.277	0.311
	3	0.290	0.772	0.435	0.422
MedicinechatQ2	2	0.175	0.289	0.288	0.285
	3	0.064	0.2	0.08	0.114
MedicinechatQ4	2	0.129	0.333	0.177	0.228
	3	0.268	0.777	0.372	0.401
MedicinechatQ8	2	0.124	0.309	0.177	0.220
	3	0.06	0.5	0.06	0.107
DrsamanthaQ2	2	0.14	0.205	0.233	0.218
	3	0.06	0.5	0.06	0.107
DrsamanthaQ4	2	0.152	0.333	0.233	0.261
	3	0.332	0.833	0.435	0.460
DrsamanthaQ8	2	0.135	0.291	0.233	0.238
	3	0.226	0.75	0.31	0.357
BiomistralQ2	2	0.052	0.166	0.066	0.095
	3	0.06	0.5	0.06	0.107
BiomistralQ4	2	0.157	0.309	0.288	0.269
	3	0.082	0.533	0.122	0.150
BiomistralQ8	2	0.052	0.166	0.066	0.095
	3	0.06	0.5	0.06	0.107

TABLE III

CLASSIFICATION PERFORMANCE AFTER EXAMPLE

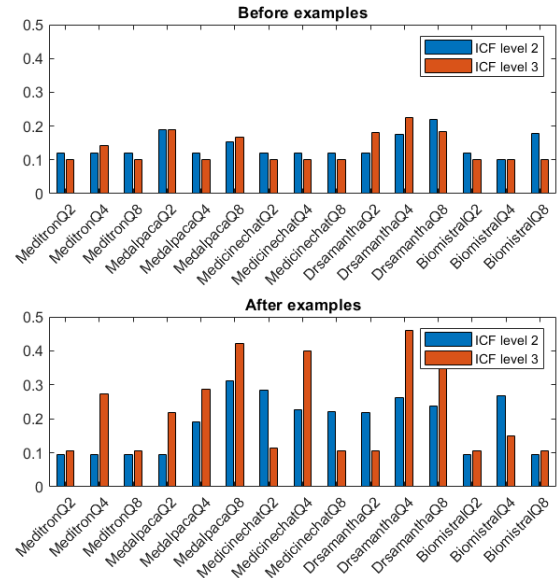


Fig. 3. Classification before and after showing correct ICF examples on F-score on the medical LLMs.

B. Benchmarking Across General-Purpose LLMs

Since all LLMs demonstrated similar performance in terms of speed and adherence to the general ICF structure, we discontinued further quantitative comparisons. Models exhibiting frequent hallucinations were excluded due to their poor performance. After considering both performance and qualitative aspects, we retained the top five models which can be identified in Figure 3: Dr. Samantha Q8, Dr. Samantha Q4, MedAlpaca Q8, Medicine Chat Q2, and Medicine Chat Q4. These selected models were then tested across both datasets (Mimic and Rehab dataset). We also added some popular general-purpose LLMs (Claude, Chat-GPT and Llama). We used for both medical and general LLMs basic prompts only requesting ICF classification (without giving any ICF structure or any correct ICF example). Additionally, the true positives were calculated based solely on the class code OR name, not code AND name to simplify the task and make it less strict. Table IV shows the different results for automatic filling of the ICF at level 2. For the two best models (Claude and Chat-GPT) we also added the standard deviation for accuracy and F-score for the two datasets.

V. DISCUSSION

In this preliminary study, we explored the application of different large language models for the automated creation of the International Classification of Functioning, Disability, and Health. By conducting both quantitative and qualitative evaluations, we observed that while quantization predominantly impacts the speed of the LLMs, it does not necessarily affect their performance in our specific task. It is important to underscore that despite the ICF's rich informational and descriptive nature, there is a notable lack of AI research dedicated to this area. Our findings show the potential of

Model	Data	accuracy	precision	recall	fscore
MedalpacaQ8	Mimic	0.13	0.29	0.17	0.21
	Rehab	0	0	0	0
MedicinechatQ2	Mimic	0.03	0.06	0.08	0.06
	Rehab	0	0	0	0
MedicinechatQ4	Mimic	0.02	0.04	0.04	0.04
	Rehab	0	0	0	0
DrsamanthaQ8	Mimic	0.06	0.08	0.13	0.09
	Rehab	0	0	0	0
DrsamanthaQ4	Mimic	0.05	0.09	0.11	0.10
	Rehab	0	0	0	0
Claude	Mimic	0.17	0.22	0.46	0.29
	Rehab	0.47	0.69	0.60	0.64
Chatgpt	Mimic	0.19	0.26	0.42	0.31
	Rehab	0.43	0.65	0.54	0.56
llama	Mimic	0.10	0.16	0.21	0.17
	Rehab	0.007	0.01	0.01	0.01

TABLE IV

PERFORMANCE OF THE 5 BEST MEDICAL LLMs AND GENERAL PURPOSE LLMs TO FILL ICF BASED ON A BASIC PROMPT.

LLMs in enhancing the efficiency and accuracy of ICF creation. However, this study represents an initial foray into this domain, highlighting the need for further research to fully harness the capabilities of AI in this context. While the ICF is a valuable tool for medical care, its adoption is limited due to the time required for manual completion. Automating or semi-automating the process using patient descriptions would significantly facilitate its wider use. This approach aims to reduce clinicians’ workload while maintaining the necessary human oversight for quality and safety. A first observation is that the medical LLMs need important prompt engineering to provide results. With simple prompts, as the one used for Table IV, we can see null results on the Rehab datasets for the medical LLMs. This is probably due to the structure of the Rehab dataset, where each patient has long descriptions written by doctors in French, while MIMIC has short, keyword-based descriptions in English. The prompt engineering is thus very important, suggesting that incorporating one-shot prompts when an example is provided alongside the request can lead to slight improvements in LLM performance (Tables II and III). While the effect is not drastic, it helps guide the models in structuring the outputs more consistently with the ICF framework.

A second point is that by providing only the ICF structure (Figure 3 upper image) will give similar results for level 2 and 3 for most of the LLMs. The bottom image shows that, after getting a correct ICF classification example, some medical LLMs are even better in predicting level 3 than level 2, while level 3 is more complex. Those are probably the LLMs which better take context into account, as more classes will provide more context.

A third important finding is that general-purpose LLMs (especially Claude and ChatGPT) outperformed medical LLMs in both datasets for basic prompts and level 2 of ICF (Table IV). Also, general LLMs give better results on the Rehab dataset, which is contrary to the results from medical LLMs. If we compare the three general LLMs, Claude and ChatGPT are the best on the average results.

When taking into account the standard deviation, Claude has a smaller standard deviation for the Rehab dataset, making it more interesting. All these findings lead to the hypothesis that general LLMs are much better than fine-tuned ones at capturing context. Indeed, the Rehab dataset descriptions of patients are much longer and more precise, based on natural text providing a better context, while MIMIC is mostly about structured clinical notes with technical words but limited context. These results also suggest that ICF is still a very general task and that domain-specific fine-tuning alone is insufficient for handling complex, reasoning-intensive medical tasks. This underscores the need for further investigation into prompt engineering, dataset adaptation, and potential model fine-tuning to optimize performance. The first prompt engineering tests where general LLMs were also provided with ICF structure and an example of good ICF classification did not show improvement on the Rehab dataset but showed important improvements on the MIMIC dataset. However, these tests need to be refined on more data. Medical LLMs will be better for specific tasks they are fine-tuned for, and not for any medical case, especially when the use case is quite general like ICF.

These findings also show the importance of the dataset in choosing the best LLM to automatically fill the ICF. The rehabilitation dataset is in French, whereas the MIMIC dataset is in English. Since ICF class names and ground truth labels are standardized in English, medical LLMs, despite their specialized training, may not have been sufficiently adapted to process French medical descriptions effectively. General-purpose LLMs, which are typically trained on large multilingual corpora, may have had an advantage in handling these inputs. Furthermore, the way the description is provided — human-generated detailed description with full patient context or structured data with a few technical terms lacking context — greatly impacts LLM performance.

Additionally, the structure of ICF classification itself presents a challenge. Unlike traditional medical tasks that involve straightforward classification or retrieval, ICF requires a nuanced understanding of functional descriptions, activity limitations, and environmental influences. This complexity may not be well captured in the datasets used to train medical LLMs, suggesting that domain adaptation alone is insufficient. Instead, fine-tuning on specific ICF-related datasets may be necessary to improve their performance in this task.

Finally, the relatively small size of the datasets used in this study represents a limitation. A larger and more diverse dataset would likely provide a better evaluation of the LLMs’ real capabilities. Future work could explore the generation of synthetic medical text data for example, by augmenting existing patient descriptions to increase dataset size and diversity while preserving clinical realism. This could help in better training and evaluating models for complex classification frameworks like ICF.

Our results challenge the assumption that medical LLMs are always preferable for medical classification tasks. While they are designed for medical applications, their limitations

in handling complex, multi-layered descriptions highlight the need for further research. Future work should explore whether fine-tuning medical LLMs on functional classification tasks would bridge this gap, or whether general-purpose LLMs with their broader knowledge and reasoning capabilities are inherently better suited for tasks like ICF classification.

VI. CONCLUSION

Our study highlights the potential of LLMs in automating ICF filling from textual data while challenging general assumptions about medical-specific models. General-purpose LLMs clearly outperformed medical LLMs across two distinct medical datasets, emphasizing the need for adaptability, reasoning, and multilingual capabilities over domain-specific pretraining alone.

Future research should explore fine-tuning and hybrid approaches to enhance performance. Moving forward, we plan to optimize a single LLM through advanced prompting and potential fine-tuning, with a focus on real-world clinical integration. Our findings pave the way for AI-driven solutions to make ICF more practical, accessible in healthcare and to facilitate its use as recommended by the WHO. Our approach aims in addressing potential challenges related to trust, regulatory compliance, and user adoption. Our results pave the way for future studies to refine and expand the application of AI in health classification systems, ultimately making the ICF a more accessible and effective tool in medical practice.

REFERENCES

- [1] Biomistral the first dedicated medical llm. Accessed: 2024-09-09.
- [2] Brief review — medalpaca — an open-source collection of medical conversational ai models and training data. Accessed: 2024-09-09.
- [3] Build medical ai using open source meditrion llm. Accessed: 2024-09-09.
- [4] Building a healthcare ai chatbot using biomistral-7b. Accessed: 2024-09-09.
- [5] A guide to building medical chatbot using medalpaca. Accessed: 2024-09-09.
- [6] Lm studio. Accessed: 2024-06-13.
- [7] Medical large language models: Bridging technology and healthcare. Accessed: 2024-09-09.
- [8] Meditrion: An llm suite especially suited for low-resource medical settings leveraging meta llama. Accessed: 2024-09-09.
- [9] Meta AI. Llama, 2023. Large language model.
- [10] Corinna Fukushima Albert Haque. Automatic documentation of icd codes with far-field speech recognition. *CoRR*, abs/1804.11046, 2018.
- [11] Anthropic. Claude, 2023. Large language model.
- [12] Jorie M Butler, Bryan Gibson, Lacey Lewis, Gayle Reiber, Heidi Kramer, Rand Rupper, Jennifer Herout, Brenna Long, David Massaro, and Jonathan Nebeker. Patient-centered care and the electronic health record: exploring functionality and gaps. *JAMIA Open*, 3(3):360–368, 10 2020.
- [13] Alarcos Cieza, Roger Hilfiker, Somnath Chatterji, Nenad Kostanjsek, Bedirhan T. Üstün, and Gerold Stucki. The international classification of functioning, disability, and health could be used to measure functioning. *Journal of Clinical Epidemiology*, 62(9):899–911, 2009.
- [14] Alarcos Cieza and Gerold Stucki. The international classification of functioning disability and health: its development process and content validity. *European journal of physical and rehabilitation medicine*, 44(3):303–313, 2008.
- [15] Prateek Munjal-Tathagata Raha Nasir Hayat Ronnie Rajan Ahmed Al-Mahrooqi Avani Gupta Muhammad Umar Salman Gurpreet Gosal Bhargav Kanakiya Charles Chen Natalia Vassilieva Boulbaba Ben Amor Marco AF Pimentel Shadab Khan Clément Christophe, Praveen K Kanithi. Med42 – evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. *AAAI 2024 Spring Symposium - Clinical Foundation Models*, 2024.
- [16] Furu Wei Daixuan Cheng, Shaohan Huang. Adapting large language models via reading comprehension. *ICLR 2024 Conference*, 2024.
- [17] George L Engel. The need for a new medical model: a challenge for biomedicine. *Science*, 196(4286):129–136, 1977.
- [18] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressen. Medalpaca – an open-source collection of medical conversational ai models and training data. *arXiv*, 2023.
- [19] James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. Icd-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21:1–10, 2021.
- [20] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- [21] Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, and David R Karger. Medknowts: Unified documentation and information retrieval for electronic health records. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, page 1169–1183, New York, NY, USA, 2021. Association for Computing Machinery.
- [22] Saad Z Nagi. A study in the evaluation of disability and rehabilitation potential: concepts, methods, and procedures. *American Journal of Public Health and the Nations Health*, 54(9):1568–1579, 1964.
- [23] OpenAI. Chatgpt (gpt-4), 2023. Large language model.
- [24] World Health Organization et al. Icf: International classification of functioning, disability and health. 2001.
- [25] Peter Rosenbaum and Debra Stewart. The world health organization international classification of functioning, disability, and health: a model to guide clinical thinking, practice and research in the field of cerebral palsy. *Seminars in Pediatric Neurology*, 11(1):5–10, 2004. Current Perspectives in Cerebral Palsy.
- [26] sethuyer. Dr samantha 7b. Accessed: 2024-06-13.
- [27] O Svestkova. International classification of functioning, disability and health of world health organization (icf). *Prague Med Rep*, 109(4):268–274, 2008.
- [28] Gilles Tagne, Mathias Blandeau, Othman Lakhali, Eugénie Avril, Laura Wallard, Steeve Mbakop, and Rochdi Merzouki. System of systems approach and user-centered design to improve the autonomy of people with reduced mobility. In *2023 18th Annual System of Systems Engineering Conference (SoSe)*, pages 1–7, 2023.
- [29] Jilda N Vargus-Adams and Annette Majnemer. International classification of functioning, disability and health (icf) as framework for change: revolutionizing rehabilitation. *Journal of Child Neurology*, 29(8):1030–1035, 2014.
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2023.
- [31] Emmanuel Morin Pierre-Antoine Gourraud Mickael Rouvier Richard Dufour Yanis Labrak, Adrien Bazoge. Biomistral: A collection of open-source pretrained large language models for medical domains. *Proceedings of the 62st Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers (ACL 2024)*, 2024.
- [32] Angelika Romanou-Antoine Bonnet Kyle Matoba Francesco Salvi Matteo Pagliardini-Simin Fan Andreas Köpf Amirkeivan Mohtashami Alexandre Sallinen Alireza Sakhaeirad Vinitra Swamy Igor Krawczuk Deniz Bayazit Axel Marmet Syrielle Montariol Mary-Anne Hartley Martin Jaggi Antoine Bosselut Zeming Chen, Alejandro Hernández Cano. Meditrion-70b: Scaling medical pretraining for large language models. *arXiv*, 2023.
- [33] Liang Zhang, Syoichi Tashiro, Masahiko Mukaino, and Shin Yamada. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. *Journal of rehabilitation medicine*, 55:jrm13373, 09 2023.