Probing the Intrinsic Effectiveness of Large Language Models for Medical Classifications

Probing the Intrinsic Effectiveness of Large Language Models for Medical Classifications

Riccardo Lunardi
University of Udine
Udine, Italy
riccardo.lunardi@uniud.it

Kevin Roitero
University of Udine
Udine, Italy
kevin.roitero@uniud.it

Vincenzo Della Mea
University of Udine
Udine, Italy
vincenzo.dellamea@uniud.it

Abstract

This paper investigates the effectiveness of Large Language Models (LLMs) in handling medical classifications, focusing on two tasks: converting medical descriptions to codes (text-to-code) and generating descriptions from codes (code-to-text). We deliberately study these tasks in a probing setting, without employing fine-tuning or Retrieval-Augmented Generation (RAG) techniques, as access to external knowledge sources would make the tasks largely trivial and would not reflect the models' intrinsic knowledge acquired during training. Our goal is therefore to evaluate the structured medical knowledge internalized by LLMs during pre-training.

Experiments across several widely used medical classifications reveal substantial limitations: although larger models generally perform better, LLMs correctly identify only a limited fraction of codes and descriptions, while frequently producing hallucinated or incorrect outputs. Performance also varies substantially across classification systems, particularly for more complex ontologies. These findings highlight significant reliability and safety concerns when applying LLMs to medical classification tasks in their current basic form, and underline the need for improved models and methodologies before such systems can be responsibly deployed in medical informatics.

CCS Concepts

• **Information systems** → **Ontologies**; **Language models**; **Information extraction**; • **Computing methodologies** → **Natural language generation**; **Information extraction**; • **Applied computing** → **Health informatics**.

Keywords

Large Language Models, Classifications, Probing, Ontologies, Medical Informatics

ACM Reference Format:
Riccardo Lunardi, Kevin Roitero, and Vincenzo Della Mea. 2026. Probing the Intrinsic Effectiveness of Large Language Models for Medical Classifications. In *17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '26), June 30–July 03, 2026, Rende (CS), Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3807503.3819373

BCB '26, Rende (CS), Italy

ACM ISBN 979-8-4007-2653-8/26/06
https://doi.org/10.1145/3807503.3819373

1 Introduction

Medical classifications such as the International Classification of Diseases (ICD-10 and ICD-11) constitute foundational infrastructures of modern healthcare systems. They are used for epidemiological monitoring, billing, reporting, pharmacological regulation, and biomedical research [18, 34, 38].

With the rapid adoption of Large Language Models (LLMs) in medical contexts, it remains unclear to what extent these models intrinsically encode structured medical classification knowledge. In many practical deployments, systems rely on Retrieval-Augmented Generation (RAG) [23] or other external knowledge access mechanisms to obtain near-exact matches from authoritative databases [1, 10]. While such approaches can substantially improve accuracy, they do not reveal whether the underlying model itself possesses internalized knowledge of these classification systems. In fact, with direct access to official repositories, the conversion between codes and descriptions becomes largely trivial from an information retrieval perspective. In such cases, the problem reduces to a standard information retrieval task over a structured and authoritative repository, rather than requiring the model to rely on its internal representations or previously acquired medical knowledge.

In this work, we therefore adopt a probing perspective and deliberately evaluate LLMs without fine-tuning or RAG. Our goal is to assess the intrinsic effectiveness of pre-trained models in handling medical classifications in zero-shot and few-shot settings. By isolating the models from external knowledge sources, we aim to measure what they have genuinely learned during training, and to identify their limitations when operating purely on internal representations. We emphasize that our objective is not to measure the best achievable performance of LLM-based systems, but rather to isolate and quantify the knowledge encoded within the models.

Specifically, we investigate two complementary tasks: converting textual descriptions from medical classifications into standardized codes (*text-to-code*), and generating official descriptions from provided codes (*code-to-text*). The former task evaluates the model's ability to map complex medical language to structured identifiers, while the latter assesses whether codified knowledge can be faithfully reconstructed into natural language. Specifically, our research addresses the following questions (RQ):

RQ1: How effectively can LLMs convert textual descriptions from medical classifications into their respective standardized codes when operating without external retrieval?
RQ2: To what extent can LLMs reliably generate official descriptive texts from classification-provided codes under the same constraints?

The code, data, and all supplementary material are available at: https://osf.io/z7cge/overview.

2 Background and Related Work

The automated conversion of medical text to standardized codes has historically relied on rule-based systems, traditional machine learning, and domain-specific encoder models like BioBERT [9] and ClinicalBERT [24, 32]. However, these architectures strictly required task-specific fine-tuning on large annotated datasets (e.g., MIMIC-III [19]). Modern auto-regressive LLMs, conversely, tackle these tasks via zero-shot and in-context learning [11]. Given this paradigm shift, it becomes crucial to determine the extent to which these models intrinsically encode structured domain knowledge, such as the exact mappings of medical classifications during their training phase.

To systematically assess this internalized knowledge level, researchers rely on probing methodologies. Probing techniques in the NLP domain refer a set of methodologies used to understand the properties of a model's internal representations or to assess its capabilities [5, 20, 31]. In this context, such evaluations require using the model strictly in inference mode to assess how well it understands and processes specific concepts [3, 22].

In the healthcare domain, LLM evaluations have largely focused on clinical summarization or question-answering [29]. To the best of our knowledge, this is the first study to explicitly probe the intrinsic knowledge of LLMs across an extensive set of medical classifications, addressing the specific challenge of bidirectional conversion between codes and natural-language descriptions without external augmentations.

3 Methodology

This section outlines the experimental setup, including the selected models, target classifications, tasks, prompting strategies, and evaluation metrics.

3.1 Large Language Models

We employ a diverse set of open-source LLMs spanning multiple architectural families and parameter scales: Qwen2.5-72B-Instruct, Qwen3-8B [15, 28]; Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct [12]; Ministral-3-8B-Instruct-2512, Mistral-Small-3.2-24B-Instruct-2506 [13]; and two medically oriented models, medgemma-4b-it and medgemma-27b-it-text [14].

We deliberately restrict our evaluation to open-source, locally deployable models to ensure full reproducibility and to strictly isolate the models' intrinsic knowledge without the confounding effects of undisclosed retrieval mechanisms or proprietary API updates. Furthermore, local deployment mirrors real-world clinical constraints where transmitting sensitive patient data to external commercial APIs risks violating data privacy frameworks such as the GDPR [2, 16, 30].

Importantly, we do not fine-tune any models nor apply techniques such as RAG. While RAG integrates external retrieval components that can directly query authoritative medical repositories, our objective is to probe the intrinsic knowledge of LLMs, evaluating what is internalized during training without external augmentation.

3.2 Classifications

We evaluate across five widely adopted medical and biomedical classification systems: the International Classification of Diseases, 10th and 11th Revisions (ICD-10 and ICD-11) [26, 36], the International Classification of Functioning, Disability and Health (ICF) [25], the Gene Ontology (GO) [4], and the Anatomical Therapeutic Chemical (ATC) Classification System [17]. This selection spans clinical, functional, molecular, and pharmacological domains, enabling a comprehensive assessment across different structures and granularities.

We emphasize that this heterogeneity is an intentional aspect of our experimental design. Differences in ontology size, identifier structure, and semantic granularity introduce varying levels of task difficulty, allowing us to analyze how LLM performance changes under diverse structural conditions.

3.3 Tasks and Prompting

We evaluate two complementary tasks: *text-to-code* and *code-to-text* conversion. Specifically, in the *text-to-code* task, the model receives the official textual description of a classification entry and must generate the exact corresponding code. In the *code-to-text* task, the model is provided with a standardized code and must generate the official description associated with it. This reverse task assesses whether the model can reconstruct accurate and faithful textual definitions from codified representations, preserving semantic specificity and avoiding hallucinated or partially correct descriptions.

To assess how inference-time reasoning influences effectivness, we apply four prompting strategies: *Zero-shot* [21], *Three-shot* [8], *Self-talk* [33], and *Chain-of-Thought (CoT)* [35]. All strategies operate strictly without external knowledge retrieval. Experiments are run with a temperature of 0 for deterministic outputs, without constrained decoding. Full prompt templates are provided in the supplementary material.

3.4 Evaluation Metrics

To evaluate LLM effectiveness in medical classification tasks, we adopt task-specific metrics tailored to the deterministic nature of standardized coding systems. In the *text-to-code* task, correctness requires an exact match between the generated output and the official classification code. We therefore compute Accuracy as the percentage of exact code matches over the total number of instances. Outputs that contain incorrect, partially correct, non-existent, or malformed codes are therefore counted as wrong.

Conversely, for the *code-to-text* task, we measure the similarity between generated descriptions and official reference texts using ROUGE-L [7, 37]. In addition, we count exact matches between generated and official descriptions as a stricter indicator of faithful reconstruction. We specifically BLEU [27], as is primarily a corpus-level precision-based metric that performs inadequately at the sentence level and is highly sensitive to tokenization and surface n-gram overlap, making it unsuitable for instance-level semantic adequacy in this context.

4 Results

In this section, we evaluate the intrinsic capabilities of LLMs to process medical classifications, addressing both the text-to-code (RQ1) and code-to-text (RQ2) tasks.

4.1 RQ1: Effectiveness of LLMs in Text-to-Code

We start by analyzing the effectiveness of the investigated LLMs in converting official medical descriptions into their corresponding standardized codes. Results for the *text-to-code* task across all models, prompting strategies, and classifications are reported in Table 1.

A first clear pattern that emerges is the strong dependence on model scale. Smaller models (4B-8B parameters), including medgemma-4b, Llama-3.1-8B, and Qwen3-8B, generally achieve near-zero accuracy across most classifications. In contrast, larger models demonstrate substantially stronger performance on selected datasets, though absolute accuracy remains far from perfect.

ICD-10 and ATC emerge as the most tractable classifications overall. For ICD-10, Llama-3.1-70B achieves 0.319 accuracy in the zero-shot setting (4,007 correct predictions out of 12,542), and 0.309 under CoT. Mistral-Small-3.2-24B reaches up to 0.195, while the domain-specialized medgemma-27b improves substantially over its 4B counterpart, reaching 0.072, but remains well below the strongest general-purpose models. These results suggest that large-scale models internalize a non-trivial portion of ICD-10 mappings, though still failing on the majority of instances.

However, highly granular or structurally complex classifications prove exceptionally challenging. The GO yields very low accuracies, with Llama-3.1-70B reaching only 0.058 under CoT. ICD-11 and ICF represent near-total failure modes, with Llama-3.1-70B peaking at 0.003 for ICD-11 and 0.063 for ICF. These findings suggest that newer, structurally complex, or less frequently encountered classification systems are very poorly internalized during training.

Prompting strategies do not consistently alter these trends. For large models, zero-shot often performs comparably to or better than more elaborate prompting schemes. For instance, Llama-3.1-70B achieves its strongest ICD-10 and ATC scores in zero-shot or CoT settings, while three-shot prompting sometimes reduces accuracy. For smaller models, differences between prompting strategies are marginal relative to the overall low effectiveness, indicating that inference-time reasoning cannot compensate for missing structured knowledge.

Overall, intrinsic text-to-code effectiveness varies dramatically across model scale and classification type. While large-scale models can achieve moderate scores on ICD-10 and ATC, effectiveness sharply deteriorates for ICD-11, ICF, and GO. Since no retrieval or external augmentation is employed, these results reflect the degree to which structured medical classification knowledge is internalized during training. The findings indicate that such knowledge is only partially encoded and often insufficient for reliable deterministic medical coding without external support.

4.2 RQ2: Effectiveness of LLMs in Code-to-Text

We now turn to the reverse task, namely generating the official textual description given a standardized medical code. Results for the *code-to-text* task are reported in Table 2, where each configuration reports exact matches and average ROUGE-L score. Unlike text-to-code, this task involves natural language generation, so exact matches are complemented with ROUGE-L to capture graded similarity.

As in RQ1, model scale plays a central role. Smaller models (4B-8B), generally achieve very low exact matches across all classifications. Even when ROUGE-L scores are non-zero, they often reflect partial or loosely related descriptions rather than fully correct reconstructions. For instance, medgemma-4b rarely produces more than a handful of exact matches (e.g., 3 correct for ICD-10 in zero-shot), with ROUGE-L scores typically below 0.08.

ICD-10 again emerges as the most tractable classification. Llama-3.1-70B achieves the strongest results, with 2,131 exact matches in zero-shot (ROUGE-L 0.448) and 1,681 exact matches under three-shot (ROUGE-L 0.382). These values are substantially higher than those observed for smaller models and indicate that large-scale models internalize a significant portion of ICD-10 textual definitions. Domain-specialized medgemma-27b reaches at most 203 exact matches (ROUGE-L 0.176), confirming that domain specialization alone does not outweigh scale.

For ATC and GO, patterns diverge from ICD-10. For ATC, effectiveness is lower overall but still non-negligible for larger models: Llama-3.1-70B achieves 229 exact matches in zero-shot (ROUGE-L 0.114), while Mistral-Small-3.2-24B reaches up to 161 exact matches under self-talk. Smaller models remain largely ineffective, often generating paraphrased or generic descriptions that do not match official entries. GO shows a contrasting pattern: Llama-3.1-70B produces 1,580 exact matches in zero-shot, yet average ROUGE-L (0.177) remains substantially lower than ICD-10, reflecting GO's semantic granularity and susceptibility to hallucinated partial definitions.

ICD-11 and ICF constitute the most challenging classifications. Even Llama-3.1-70B produces zero exact matches for ICD-11 across all prompting strategies, despite non-trivial ROUGE-L scores (e.g., 0.052 zero-shot), indicating partially overlapping but imprecise descriptions. ICF shows similarly near-zero exact matches, though some configurations achieve moderate ROUGE-L values (e.g., 0.257 for Mistral-Small-3.2-24B under three-shot).

Prompting strategies exhibit mixed effects. For large models, three-shot prompting sometimes increases the number of exact matches (e.g., Llama-3.1-70B for ICD-10), while in other cases zero-shot performs comparably or better. Self-Talk and CoT do not consistently improve reconstruction accuracy and occasionally reduce exact-match counts despite similar ROUGE-L values. For smaller models, differences between prompting strategies are generally marginal relative to the overall low effectiveness.

Overall, intrinsic code-to-text effectiveness mirrors the patterns observed in RQ1. Large models partially internalize widely used classifications like ICD-10, but effectiveness deteriorates sharply for ICD-11, ICF, and GO. These results confirm that consistent deterministic reconstruction remains limited without external knowledge access.

4.3 Prediction Overlap in Code and Text Generation

To further understand the relationship between the two tasks, we analyze the overlap of correct predictions between the text-to-code and code-to-text tasks. Given the generally lower performance in code-to-text generation, a reasonable assumption might be that its successful predictions are merely a subset of the text-to-code task.

Table 1: Metrics for text-to-code.

Dataset	Tech.	medgemma-4b	medgemma-27b	Llama-3.1-8B	Llama-3.1-70B	Ministral-3-8B	Mistral-S-3.2-24B	Qwen3-8B	Qwen2.5-72B
ATC (N=6,799)	Zero-Shot	19 (.003)	244 (.036)	139 (.020)	1,902 (.280)	206 (.030)	884 (.130)	4 (.001)	175 (.026)
	3-Shot	15 (.002)	159 (.023)	141 (.021)	959 (.141)	441 (.065)	521 (.077)	22 (.003)	115 (.017)
	Self-Talk	15 (.002)	251 (.037)	42 (.006)	841 (.124)	374 (.055)	506 (.074)	0 (.000)	421 (.062)
	CoT	17 (.003)	187 (.028)	166 (.024)	715 (.105)	398 (.059)	946 (.139)	24 (.004)	440 (.065)
GO (N=42,271)	Zero-Shot	5 (.000)	0 (.000)	235 (.006)	2,316 (.055)	500 (.012)	730 (.017)	0 (.000)	56 (.001)
	3-Shot	5 (.000)	22 (.001)	7 (.000)	1,611 (.038)	437 (.010)	1,036 (.025)	1 (.000)	15 (.000)
	Self-Talk	5 (.000)	85 (.002)	120 (.003)	1,683 (.040)	468 (.011)	1,012 (.024)	24 (.001)	689 (.016)
	CoT	7 (.000)	50 (.001)	262 (.006)	2,472 (.058)	455 (.011)	1,132 (.027)	34 (.001)	718 (.017)
ICD10 (N=12,542)	Zero-Shot	133 (.011)	891 (.071)	485 (.039)	4,007 (.319)	874 (.070)	2,448 (.195)	10 (.001)	385 (.031)
	3-Shot	32 (.003)	858 (.068)	332 (.026)	3,244 (.259)	752 (.060)	2,393 (.191)	639 (.051)	595 (.047)
	Self-Talk	140 (.011)	908 (.072)	258 (.021)	2,206 (.176)	1,254 (.100)	848 (.068)	120 (.010)	2,089 (.167)
	CoT	144 (.011)	856 (.068)	610 (.049)	3,872 (.309)	1,360 (.108)	2,387 (.190)	133 (.011)	2,209 (.176)
ICD11 (N=30,966)	Zero-Shot	1 (.000)	7 (.000)	0 (.000)	91 (.003)	9 (.000)	26 (.001)	0 (.000)	6 (.000)
	3-Shot	1 (.000)	8 (.000)	4 (.000)	64 (.002)	6 (.000)	26 (.001)	3 (.000)	5 (.000)
	Self-Talk	1 (.000)	8 (.000)	0 (.000)	83 (.003)	5 (.000)	14 (.000)	0 (.000)	6 (.000)
	CoT	2 (.000)	7 (.000)	1 (.000)	82 (.003)	3 (.000)	17 (.001)	0 (.000)	4 (.000)
ICF (N=1,374)	Zero-Shot	0 (.000)	2 (.001)	2 (.001)	55 (.040)	4 (.003)	42 (.031)	0 (.000)	1 (.001)
	3-Shot	0 (.000)	7 (.005)	5 (.004)	87 (.063)	7 (.005)	61 (.044)	4 (.003)	16 (.012)
	Self-Talk	0 (.000)	1 (.001)	0 (.000)	27 (.020)	8 (.006)	27 (.020)	0 (.000)	20 (.015)
	CoT	0 (.000)	0 (.000)	1 (.001)	53 (.039)	11 (.008)	46 (.033)	0 (.000)	13 (.009)

Table 2: Metrics for code-to-text (Average ROGUE-L F1 score).

Dataset	Tech.	medgemma-4b	medgemma-27b	Llama-3.1-8B	Llama-3.1-70B	Ministral-3-8B	Mistral-S-3.2-24B	Qwen3-8B	Qwen2.5-72B
ATC (N=6,799)	Zero-Shot	0 (.009)	12 (.028)	2 (.028)	229 (.114)	28 (.028)	129 (.078)	0 (.011)	28 (.048)
	3-Shot	5 (.001)	42 (.022)	9 (.034)	158 (.087)	46 (.035)	121 (.048)	2 (.005)	18 (.043)
	Self-Talk	0 (.000)	40 (.035)	2 (.028)	195 (.105)	12 (.041)	161 (.075)	0 (.001)	55 (.042)
	CoT	0 (.001)	38 (.034)	13 (.021)	307 (.111)	29 (.039)	210 (.093)	0 (.000)	75 (.055)
GO (N=42,271)	Zero-Shot	0 (.010)	21 (.103)	12 (.010)	1,580 (.177)	211 (.120)	570 (.127)	13 (.063)	155 (.066)
	3-Shot	3 (.000)	15 (.079)	35 (.088)	949 (.137)	226 (.079)	458 (.090)	3 (.000)	125 (.082)
	Self-Talk	0 (.000)	10 (.087)	15 (.018)	677 (.122)	77 (.111)	39 (.029)	0 (.001)	203 (.119)
	CoT	0 (.004)	8 (.075)	84 (.118)	837 (.139)	84 (.108)	482 (.135)	0 (.000)	113 (.111)
ICD10 (N=12,542)	Zero-Shot	4 (.078)	136 (.181)	19 (.117)	2,131 (.448)	83 (.155)	939 (.329)	2 (.001)	414 (.278)
	3-Shot	4 (.017)	190 (.176)	16 (.088)	1,681 (.382)	188 (.207)	556 (.240)	24 (.139)	250 (.264)
	Self-Talk	0 (.003)	203 (.186)	16 (.062)	1,813 (.418)	110 (.189)	396 (.165)	2 (.001)	636 (.300)
	CoT	0 (.006)	190 (.183)	21 (.081)	1,997 (.432)	143 (.199)	889 (.332)	0 (.000)	623 (.299)
ICD11 (N=30,966)	Zero-Shot	0 (.034)	0 (.050)	0 (.022)	0 (.052)	0 (.027)	0 (.047)	0 (.003)	0 (.047)
	3-Shot	3 (.021)	3 (.036)	2 (.031)	2 (.016)	5 (.047)	3 (.043)	3 (.031)	3 (.047)
	Self-Talk	0 (.012)	1 (.041)	0 (.035)	1 (.052)	0 (.026)	0 (.045)	0 (.000)	1 (.044)
	CoT	0 (.015)	0 (.044)	2 (.031)	0 (.043)	0 (.034)	0 (.048)	0 (.000)	0 (.046)
ICF (N=1,374)	Zero-Shot	0 (.045)	0 (.020)	0 (.042)	13 (.121)	0 (.036)	22 (.195)	0 (.041)	1 (.083)
	3-Shot	3 (.066)	3 (.040)	3 (.056)	21 (.157)	3 (.040)	40 (.257)	3 (.046)	4 (.066)
	Self-Talk	0 (.050)	0 (.039)	0 (.005)	0 (.080)	0 (.050)	15 (.179)	0 (.004)	0 (.102)
	CoT	0 (.010)	0 (.005)	0 (.000)	9 (.154)	0 (.051)	19 (.191)	0 (.000)	1 (.072)

However, our analysis reveals a low overlap between the two. For instance, the best-performing Llama-3.1-70B model on ICD-10, over 55% of its correct zero-shot code-to-text predictions (1,189 out of 2,131) are exclusive to that direction. Similar disjoints are present across other models and datasets, meaning that models often fails to map the exact same concept when prompted in reverse.

This asymmetry suggests that the models' internal representations of medical coding knowledge are highly directional. We hypothesize that this behavior might be a manifestation of the Reversal Curse [6] in auto-regressive language models, likely reflecting the contextual sequencing of medical data in web corpora. Consequently, despite the lower overall accuracy of the code-to-text task, it uncovers complementary intrinsic knowledge rather than redundant information. This lack of overlap makes bidirectional generation an interesting and valuable signal, suggesting that future methodologies could benefit significantly from integrating both generative directions to maximize coverage.

5 Error Analysis

Evaluating model effectiveness solely through exact-match accuracy provides a limited view, as it does not capture the severity or structure of errors. In hierarchical classification settings, incorrect predictions may still be close to the correct code (indicating related concepts) or far apart (indicating substantial deviations), and errors may not be uniformly distributed across the label space. To better characterize these aspects, we analyze the structural distance between predicted and ground-truth codes: hierarchical distance for tree-based ontologies (ICD-10, ICD-11, ICF, ATC) and graph distance for GO[1]. Correct predictions are excluded to avoid biasing

[1]We refer to the GO basic version, which forms a directed acyclic graph

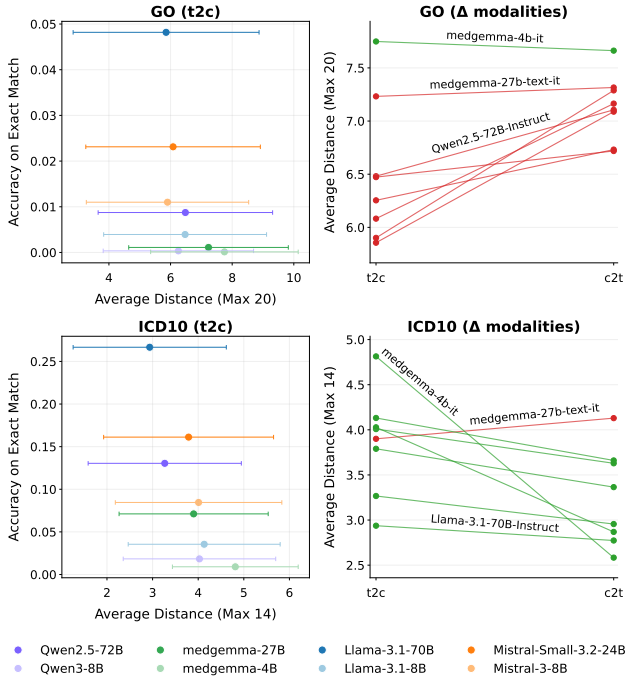

Figure 1: Average error distance across models and tasks (GO and ICD-10).

the analysis, and for reference, maximum distances are 14 (ICD-10), 13 (ICF), 6 (ATC), and 20 (GO and ICD-11).

Figure 1 reports the distance-based analysis for GO (top) and ICD-10 (bottom), shown as representative contrasting examples of contrasting patterns. The left column presents, for the text-to-code direction, the relationship between average error distance and exact-match accuracy per model, while the right column illustrates how the average distance shifts between the text-to-code and code-to-text tasks (t2c and c2t, respectively). Mean error distance remains relatively stable within each classification (5.9–7.7 for GO; 2.9–4.8 for ICD-10), but within-model standard deviation is consistently high, indicating errors are distributed across multiple hierarchy depths rather than concentrated at a specific level. Consistent with the results from RQ1 and RQ2, Llama-3.1-70B-Instruct generally exhibits slightly lower average distances, indicating that its errors tend to remain closer to the correct codes even when predictions are incorrect.

Turning to the comparison across tasks, a consistent asymmetry emerges. For GO, models systematically exhibit higher average distances in the code-to-text direction, indicating that errors in natural language generation tend to correspond to structurally more distant codes. In contrast, for ICD-10, most models show a decrease in average distance when moving to code-to-text, suggesting that generating descriptions from codes leads to structurally closer approximations despite lower exact-match rates. Directional trends across the remaining classifications are mixed: ICF follows a pattern similar to GO, with errors becoming structurally more distant in the code-to-text direction, while ATC and ICD-11 show no consistent directional shift across models. Notably, across all 320 configuration combinations, a Mann–Whitney U test with Bonferroni correction confirms that model errors are systematically closer to the ground truth than arbitrary random predictions: 290 out of 320 configurations (90.6%) are statistically significant ($\alpha = 0.05$), confirming that these directional patterns reflect ontology structure rather than random failure.

Examining coverage across all available configurations, approximately 66.6% of ICD-10 codes and 48.0% of ATC codes are correctly predicted by at least one configuration in either direction, indicating meaningful internalization across our evaluated model set. In contrast, over 88% of GO codes, 81% of ICF codes, and 99.2% of ICD-11 codes are never correctly identified, with ICD-11's near-total failure consistent with its recent introduction and limited training corpus representation.

Taken together, these findings provide a consistent picture. While incorrect predictions tend to remain structurally proximate to the correct answer rather than reflecting arbitrary failures, this structured behavior does not prevent systematic and persistent blind spots, most pronounced for ICD-11, ICF, and GO. The coverage gaps are not simply a consequence of real-world code rarity, nor are they fully addressable through inference-time choices such as prompting strategy or task direction. Instead, they point to fundamental absences in the medical knowledge encoded during training, gaps that would require targeted data and training interventions to address.

6 Impact and Limitations

This study provides a systematic probing analysis of the intrinsic effectiveness of state-of-the-art LLMs in handling structured medical classification knowledge. By deliberately excluding fine-tuning and retrieval-augmented mechanisms, our results offer a lower-bound estimate of what current language models internalize. The findings have direct implications for the safe deployment of LLMs in healthcare settings. The findings highlight a severe operational and safety risk: across tasks, even large-scale models frequently fail to produce exact deterministic mappings, particularly for ICD-11, ICF, and GO.

A central impact of our findings concerns the role of model scale. While scaling improves performance, it does not guarantee robust internalization of formal coding systems. These results suggest that internalization of formal coding systems may be partially conditioned by data availability during training: classifications such as ICD-10 may appear more frequently in public corpora, while ICD-11, ICF, and fine-grained GO entries may be underrepresented, resulting in fragmented or incomplete internal representations. However, our analysis indicates that real-world code frequency alone offers only a weak signal for predicting model success, pointing to additional factors beyond mere exposure.

Nonetheless, this study has several limitations. We evaluated a non-exhaustive set of open-source models: proprietary architectures or domain-adaptive variants may behave differently. Additionally, our strict reliance on exact code matching and ROUGE-L similarity, while necessary for deterministic tasks, may not fully capture partial semantic correctness.

Finally, our experimental design intentionally excludes retrieval augmentation and task-specific fine-tuning. While this choice is

methodologically necessary to probe intrinsic knowledge, it does not reflect how production systems are typically engineered. In practical deployments, RAG-based pipelines or supervised fine-tuning on curated medical corpora may substantially improve effectiveness. However, our results indicate that without such external support, current LLMs do not reliably internalize structured medical classifications.

7 Conclusions and Future Work

In this work, we evaluated the intrinsic capabilities of LLMs to process structured medical classifications in text-to-code and code-to-text tasks. Our findings demonstrate that while scaling up model parameters improves engagement with widely used systems like ICD-10 and ATC, effectiveness remains highly variable and often insufficient for safety-critical applications. Furthermore, models fail almost completely on granular or newly introduced classifications like GO and ICD-11.

Importantly, the discrepancy between moderate ROUGE-L similarity and low exact-match accuracy in the code-to-text task highlights a broader limitation: LLMs may generate semantically related or plausible descriptions while failing to reconstruct the precise official definitions required in regulated medical contexts. This gap between approximate semantic competence and deterministic correctness reinforces the need for caution when integrating LLMs into clinical workflows.

Future work will explore several directions. First, targeted domain-adaptive training strategies could improve the internal representation of structured medical classifications, especially for under-represented systems such as ICD-11 and ICF. Additionally, hybrid approaches that combine structured symbolic knowledge with neural representations, such as constrained decoding, ontology-aware architectures, or retrieval-grounded verification layers, could bridge the gap between flexibility and reliability. Finally, a deeper investigation into the relationship between scaling laws, training data composition, and structured knowledge internalization may clarify why certain classifications are more effectively encoded than others.

Acknowledgments

This work was supported in part by the ESF+ 2021/2027 Regional Program of the Autonomous Region Friuli Venezia Giulia (PPO 2023, Program No. 22/23 – LINE A: PhD programmes) and the Region FVG Project "Supporting the diagnosis of rare diseases (MR) through artificial intelligence" (2023-2026, Project A: F53C22001770002).

References

[1] Swapna Abhyankar, Dina Demner-Fushman, Fiona M Callaghan, et al. 2014. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *Journal of the American Medical Informatics Association* 21, 5 (01 2014).
[2] Sallam Abualhaija, Marcello Ceci, Nicolas Sannier, et al. 2025. LLM-assisted extraction of regulatory requirements: A case study on the GDPR. In *2025 IEEE 33rd RE*. IEEE.
[3] David Arps, Younes Samih, Laura Kallmeyer, et al. 2022. Probing for constituency structure in neural language models. *arXiv:2204.06201* (2022).
[4] M Ashburner, C A Ball, J A Blake, et al. 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 1 (May 2000).
[5] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* 48, 1 (2022).
[6] Lukas Berglund, Meg Tong, Maximilian Kaufmann, et al. 2024. The Reversal Curse: LLMs trained on "A is B"fail to learn "B is A". In *ICLR*, Vol. 2024.
[7] Kathrin Blagec, Georg Dorffner, Milad Moradi, et al. 2022. A global analysis of metrics used for measuring performance in natural language processing. *arXiv:2204.11574* (2022).
[8] Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. In *Adv. NeurIPS*, Vol. 33. Curran Associates, Inc.
[9] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, et al. 2020. BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR. In *Proc. COLING*. International Committee on Computational Linguistics, Barcelona, Spain (Online).
[10] Yunzhi Chen, Huijuan Lu, and Lanjuan Li. 2017. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PLoS One* 12, 3 (March 2017).
[11] Julian Coda-Forno, Marcel Binz, Zeynep Akata, et al. 2023. Meta-in-context learning in large language models. In *Adv. NeurIPS*, Vol. 36.
[12] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models.
[13] Alexander H. Liu et al. 2026. Ministral 3.
[14] Andrew Sellergren et al. 2025. MedGemma Technical Report.
[15] An Yang et al. 2025. Qwen3 Technical Report.
[16] Georgios Feretzakis, Evangelia Vagena, et al. 2025. GDPR and large language models: Technical and legal obstacles. *Future Internet* 17, 4 (2025).
[17] WHO Collaborating Centre for Drug Statistics Methodology. 2001. ATC Index.
[18] Gregor Jezernik, Mario Gorenjak, and Uroš Potočnik. 2022. Gene Ontology Analysis Highlights Biological Processes Influencing Non-Response to Anti-TNF Therapy in Rheumatoid Arthritis. *Biomedicines* 10, 8 (2022).
[19] Alistair E W Johnson, Tom J Pollard, Lu Shen, et al. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 1 (May 2016).
[20] Najoung Kim, Roma Patel, et al. 2019. Probing what different NLP tasks teach machines about function word comprehension. *arXiv:1904.11544* (2019).
[21] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, et al. 2022. Large Language Models are Zero-Shot Reasoners. In *NeurIPS*, Vol. 35. Curran Associates, Inc.
[22] Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. *arXiv:2104.05882* (2021).
[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. NeurIPS* 33 (2020).
[24] Xiaohong Liu, Hao Liu, Guoxing Yang, et al. 2025. A generalist medical language model for disease diagnosis assistance. *Nature Medicine* 31, 3 (01 March 2025).
[25] World Health Organization. 2001. International classification of functioning, disability and health : ICF.
[26] World Health Organization. 2004. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.
[27] Kishore Papineni, Salim Roukos, Todd Ward, et al. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. COLING*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.
[28] Qwen. 2025. Qwen2.5 Technical Report.
[29] Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023).
[30] Ali Vaezi. 2025. Legal Challenges in the Deployment of Large Language Models: A Comparative Analysis under the GDPR and EU AI Act. *SSRN5285174* (2025).
[31] Eric Wallace, Yizhong Wang, Sujian Li, et al. 2019. Do NLP models know numbers? probing numeracy in embeddings. *arXiv:1909.07940* (2019).
[32] Guangyu Wang, Xiaohong Liu, Zhen Ying, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine* 29, 10 (01 Oct. 2023).
[33] Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv:2203.11171* (2022).
[34] Valerie JM Watzlaf, Jennifer Hornung Garvin, Sohrab Moeini, et al. 2007. The effectiveness of ICD-10-CM in capturing public health diseases. *Perspectives in health information management* 4, 1 (2007).
[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Adv. NeurIPS* 35 (2022).
[36] 72 World Health Assembly. 2019. Eleventh revision of the International Classification of Diseases. (2019).
[37] An Yang, Kai Liu, Jing Liu, et al. 2018. Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. *arXiv:1806.03578* (2018).
[38] Haochen Zhao, Guihua Duan, Peng Ni, et al. 2023. RNPredATC: A deep residual learning-based model with applications to the prediction of drug-ATC code association. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 20, 5 (Sept. 2023).