Travel on the ICD Tree: Benchmarking Agentic Reasoning for ICD Coding from Chinese Electronic Medical Records

Xinjie Xu[♢*], **Yongqi Fan**[♢*], **Shuangshuang Chen**[♠], **Xinxuan Hu**[♢]
Weibin Guo[♢†], **Qi Ye**[♢†]
[♢]East China University of Science and Technology, Shanghai, China
[♠]Zhongshan Hospital Affiliated to Fudan University, Shanghai, China
{gweibin,yehqi1125}@ecust.edu.cn

Abstract

Accurate International Classification of Diseases (ICD) coding is crucial for hospital management and healthcare data governance. In clinical practice, straightforward cases can often be matched directly to ICD codes via diagnostic text, establishing retrieval-based methods as the baseline. More advanced approaches leverage large language models to rerank these results. However, real-world coding scenarios are typically more complex, demanding reasoning that goes beyond superficial descriptions. For instance, it involves synthesizing key information such as disease subtype, anatomical location, and complications from complex progress notes to accurately identify the primary diagnosis. However, a comprehensive evaluation framework for ICD coding based on complete EMRs is still lacking. To address these challenges, we constructed the Code4Detail dataset, which comprises 560 real clinical records covering 434 common diseases across 19 core chapters of ICD-10. To systematically explore the capability boundaries of large language models under different paradigms, we further propose the Travel on the ICD Tree (ToT-ICD) evaluation framework. Unlike the conventional retrieval-recall approach, ToT-ICD treats ICD coding as a structured exploration process across a hierarchical taxonomy. We design an agentic workflow that integrates similarity retrieval, path-guided navigation, and dynamic backtracking, enabling logical reasoning and decision-making under coding rules.

1 Introduction

The International Classification of Diseases (ICD) coding serves as the foundational standard for healthcare data governance, directly underpinning hospital evaluation, medical insurance settlement, and clinical research (World Health Organization, 2004). In practical clinical applications, accurately performing fine-grained "leaf-node-level" coding requires coders to possess not only medical knowledge but also the ability to conduct logical investigation and rule comparison within the complex hierarchical structure of ICD. With the evolution of Large Language Models (LLMs) (Tu et al., 2023), automated coding technology is advancing from simple text matching to complex clinical reasoning. However, in real-world deployment, researchers still face the dual challenges of lacking high-quality Chinese fine-grained benchmarks and unclear task paradigm selection.

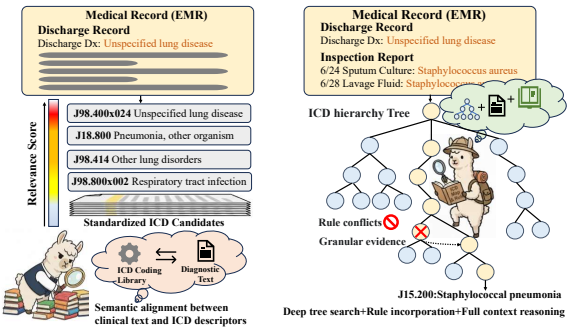

Figure 1: Comparison of two ICD coding paradigms: traditional semantic alignment (left) and the *Travel on the ICD Tree* framework (right).

In-depth analysis reveals that ICD coding tasks exhibit significant **task heterogeneity and difficulty gradients** in practical applications. In cases where medical records feature explicit diagnostic descriptions, the textual content can be directly mapped to standard ICD terminology, making retrieval-based or reranking approaches an efficient solution (as illustrated in the left panel of Figure 1). However, in complex clinical scenarios, coding logic often involves incomplete diagnostic descriptions, ambiguous expressions, or primary diagnosis selection among multiple complications (as depicted in the right panel of Figure 1). These tasks demand that models not only possess seman-

[*]Co-first authors.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 3927–3943

tic matching capabilities but also engage in deep evidence tracing and rule comparison by integrating complete clinical course records.

Yet, this task complexity and heterogeneity have not been adequately represented in existing research and evaluation systems for automated coding, leading to two critical bottlenecks.

First, there is a **lack of suitable evaluation benchmarks**: prevailing benchmarks like the MIMIC series (Johnson et al., 2016, 2023) are English-based and annotated at the category level, missing Chinese leaf-node samples needed for high-difficulty decisions involving anatomical or pathogen details. Second, there is **uncertainty in paradigm suitability**: most work treats coding as flat classification, failing to capture the hierarchical logic of human coders or to compare paradigms across "straightforward" versus "complex" tasks.

From the perspective of professional coding practice, ICD coding is intrinsically a **hierarchical, rule-constrained decision-making process**. Human coders must comprehensively analyze clinical information within medical records, progressively navigate to appropriate diagnostic categories within the ICD taxonomic hierarchy, and adhere to explicit coding rules and exclusion principles throughout this process. Although LLMs have exhibited strong capabilities in medical text comprehension, most existing LLM-based ICD coding methods persist in employing single-step prediction or flat-ranking approaches, failing to fully encapsulate this structured decision process. This raises a pivotal question: How do various LLM-driven coding workflows perform in realistic, complex ICD coding scenarios? To address this question, this paper first constructs **Code4Detail**, a benchmark dataset derived from real-world Chinese electronic medical records for the systematic evaluation of ICD coding workflows. Building upon this benchmark, we propose **Travel on the ICD Tree (ToT-ICD)**, an agentic reasoning framework. We conceptualize ICD coding as a **structured exploration process** across the ICD hierarchy and introduce an agent-based workflow (Yao et al., 2022) that integrates similarity retrieval with hierarchical path navigation and backtracking mechanisms (Wei et al., 2022), while providing the model with relevant ICD coding rules during decision-making. Grounded in the constructed benchmark, we conduct a comprehensive evaluation comparing traditional retrieval-reranking workflows with agent-based workflows across a range of open-source and closed-source large language models.

Experimental results reveal significant performance disparities among different workflows across ICD coding scenarios of varying difficulty levels. The evaluation further demonstrates the potential advantages of agent-based, ICD-tree-guided reasoning workflows in complex clinical coding situations. This research provides novel empirical insights for understanding the capability boundaries of large language models in real-world ICD coding tasks and establishes a reusable evaluation benchmark for subsequent related studies.

Our contributions are summarized as follows:

- **Task Analysis:** We systematically analyze the difficulty variation of ICD coding tasks in real clinical environments, highlighting the insufficient attention paid to complex coding scenarios that rely on reasoning with complete medical records in existing evaluation settings.
- **Reasoning Framework:** We propose the *Travel on the ICD Tree (ToT-ICD)* agentic reasoning framework, which models ICD coding as an exploratory decision-making process within the ICD hierarchical structure.
- **Benchmark and Evaluation:** We construct **Code4Detail**, a real-world Chinese electronic medical record benchmark dataset. Grounded in this benchmark, we conduct a comprehensive comparative analysis of retrieval-reranking methods and agent-based coding workflows across various large language models, revealing the strengths and limitations of different approaches in complex ICD coding scenarios.

In addition to the above contributions, we also publicly release our code and data on the repository: `https://github.com/JOHNNY-fans/Travel-on-the-ICD-Tree`.

2 Related Work

2.1 ICD Coding Datasets

High-quality standardized datasets are the primary driver of automated ICD coding research. The MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets, derived from de-identified U.S. critical care units, have long served as the *de facto* gold standard. These benchmarks

have enabled the development and rigorous comparison of numerous deep learning models. However, reliance on these English-centric datasets limits model generalizability to other linguistic and administrative contexts.

Research by Chen et al. (Chen et al., 2017) highlights that Chinese clinical texts possess unique characteristics—such as flexible abbreviations and complex terminology structures—that differ fundamentally from English corpora, rendering direct model transfer ineffective. Moreover, while ICD-11 has been officially released, ICD-10 remains the dominant operational standard for payment and quality monitoring in many regions, as evidenced by continued focus in recent studies (Sun et al., 2024; Zhou et al., 2021). Existing benchmarks often evaluate performance only at coarse "Chapter" or "Category" levels, lacking the "Full Code" (leaf node) granularity required for real-world automated billing (Zhou et al., 2020).

2.2 Evolution of Automated Coding Methods

The methodology for automated coding has evolved through three phases: discriminative learning, the generative shift enabled by foundation models, and agentic reasoning workflows. **Discriminative Deep Learning.** Early approaches treated coding as text classification using machine learning classifiers (Larkey and Croft, 1996). The deep learning era began with the CAML framework (Mullenbach et al., 2018), which utilized CNNs with label-attention mechanisms. To address the hierarchical nature of ICD codes, Perotte et al. (Perotte et al., 2014) introduced hierarchy-based support vector machines, while later works like JAN (Wu et al., 2022) and LgFAT-RGCN (Chen et al., 2023) incorporated label co-occurrence and graph structures to model dependencies. With the rise of PLMs, methods such as PLM-ICD (Huang et al., 2022) and XR-LAT (Liu et al., 2023) leveraged segment pooling for long clinical documents. However, these discriminative methods often struggle with zero-shot generalization and the long-tail distribution of rare diseases. **Foundation Models as Generative Engines.** To overcome the rigidity of fixed label sets, the field shifted toward generative paradigms, accelerated by the rapid maturation of Foundation Models. The Qwen series, specifically Qwen3 (Yang et al., 2025), demonstrates state-of-the-art instruction following and multilingual proficiency, essential for processing non-English records. DeepSeek-V3 (Liu et al., 2024a) utilizes Mixture-of-Experts (MoE) architectures to balance inference efficiency with deep logic, while Gemini (Team et al., 2023) offers extended context windows for ingesting full patient histories. These models enable approaches like GPsoap (Yang et al., 2023) and MedCodER (Baksi et al., 2025), which use prompting to generate diagnoses.However, to ensure the generated diagnoses strictly adhere to standardized terminologies, researchers increasingly integrate Large Language Models (LLMs) with retrieval and reranking mechanisms. Within this landscape, the RRNorm framework (Fan et al., 2024) harnesses LLM-driven terminology component recognition coupled with dense retrieval to achieve high-precision diagnosis normalization. Furthermore, the MedOdyssey benchmark (Fan et al., 2025a) evaluates retrieval robustness in ultra-long clinical narratives via medical “needle-in-a-haystack” tasks, while MedEureka (Fan et al., 2025b) establishes a multi-granular retrieval evaluation suite for healthcare settings. Collectively, these efforts underscore the efficacy of foundation models as powerful semantic matching and generative reasoning engines. **Retrieval, Reasoning, and Agentic Workflows.** Recent research in medical code generation focuses on mimicking human cognitive processes. “Retrieval and Rerank” strategies (Silva et al., 2024; Wang et al., 2024) decompose the task by retrieving candidate codes and using LLMs to re-rank them. More advanced frameworks incorporate explicit reasoning. Chain-of-Thought (CoT) (Wei et al., 2022) allows models to articulate decision logic. Inspired by in-text reasoning (OneRec-Think) (Liu et al., 2025), the MSDiagnosis framework (Hou et al., 2024) introduces a “forward inference and backward refinement” strategy. Subsequently, medical frameworks such as MedCoT (Liu et al., 2024b) further employ hierarchical experts to enhance reasoning rigor. The latest frontier lies in agentic systems: ReAct (Yao et al., 2022) synergizes reasoning and acting, while CLH (Motzfeldt et al., 2025) employs multi-agent collaboration for verification. Nonetheless, transforming unstructured rules into executable constraints for tree search and enabling dynamic backtracking decisions remains a missing core component in current research. Consequently, integrating structured search algorithms—such as Tree of Thoughts (ToT)—with rigorous clinical guidelines represents a crucial next step to empower agents with systematic verification capabilities.

3 Methodology

3.1 Task Definition

Automated ICD Coding (AICD) aims to bridge the semantic gap between unstructured clinical narratives in Electronic Medical Records (EMRs) and standardized medical ontologies. The Code4Detail benchmark in this study focuses on the challenging task of primary diagnosis inference to investigate the reasoning boundaries of Large Language Models (LLMs). This requires models to accurately navigate the vast label space of the Chinese national clinical standard (GB/T 14396-2016) and pinpoint leaf-node codes, moving beyond simple disease category recognition. Formally, we define the task as a hierarchical classification path search problem. Let $\mathcal{D}$ denote the space of unstructured clinical texts (e.g., treatment course, pathology report). Let $\mathcal{T}$ represent the ICD knowledge tree, with $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ being the set of leaf nodes. Given a clinical record $d \in \mathcal{D}$, the model must learn a discriminative function $f : \mathcal{D} \to \mathcal{Y}$ such that the predicted leaf-node code y^* satisfies:

$$y^* = \arg\max_{y \in \mathcal{Y}} P(y \mid d, \mathcal{T}) \tag{1}$$

The target code y^* should accurately reflect the core clinical problem consuming the most healthcare resources during the hospitalization.

To achieve such leaf-node-level precision, models must overcome three core challenges that elevate the task beyond mere semantic matching:

- **Clinical Specificity Confirmation:** The model must perform precise mapping based on explicit clinical evidence (e.g., anatomical laterality), selecting specific leaf-node codes over generalized parent codes. This requires evidence-aligned reasoning beyond simple pattern matching.
- **Adherence to Complex Coding Rules:** The model must function as a clinical decision system, understanding and applying strict coding rules (e.g., cause vs. manifestation, acute vs. chronic) that can depend on treatment context. A critical case is the coding for malignant tumors, where the correct code strictly depends on the primary purpose of the current hospitalization: if the hospitalization is for surgical treatment, then code to Chapter 2, if the hospitalization is solely for subsequent chemotherapy or radiotherapy, then code to Chapter 21.
- **Resistance to Contextual Interference:** The model must distinguish current active conditions from historical information to avoid temporal confusion, and refrain from fabricating unmentioned details to fit a specific code.

3.2 Code4Detail Benchmark Dataset

To bridge existing benchmark gaps and systematically assess the aforementioned challenges, we develop a rigorous multi-stage pipeline to construct Code4Detail—a Chinese ICD-coding benchmark built from full electronic medical records. The pipeline ensures clinical authenticity, broad representativeness, and calibrated task difficulty.

3.2.1 Data Sourcing & Anonymization

The initial corpus consisted of 11,004 real inpatient records provided by a partner institution. The hospital performed initial de-identification by removing all direct identifiers. We then conducted additional processing to protect sensitive information, applying temporal offsets and redesigning record numbers. Subsequently, we performed semantic-level "narrative reconstruction" using a locally deployed Qwen3-32B model to reconstruct and generate text, thereby eliminating personal details and stylistic features while strictly preserving clinical facts and logic. The entire pipeline was executed on a private computing cluster, thereby guaranteeing data security and regulatory compliance throughout the entire lifecycle.

3.2.2 Representative Sampling Strategy

To ensure the benchmark reflects real-world clinical heterogeneity, we performed stratified sampling from a candidate pool covering 434 ICD categories and 825 subcategories. Categories with only isolated samples were excluded to enhance evaluation robustness. This strategy yields a candidate pool closely aligned with real clinical settings in anatomical systems and disease spectrum distribution, ensuring broad coverage of common conditions.

3.2.3 Cognitive Difficulty Calibration

To prevent the benchmark from degrading into simple keyword matching, we introduced a heuristic filter based on semantic overlap. We computed Jaccard similarity between the history of present illness and the diagnosis on the medical record front page, strategically prioritizing records with overlap below 80% for the test set. Low overlap indicates the diagnosis is not explicitly stated and must be inferred from narrative evidence, thereby forcing

models to perform deep clinical reasoning and effectively evaluate their ability to handle complex decision rules.

3.3 Inference Paradigms

To address the dual demands of "broad coverage" and "deep logic" in ultra-fine-grained ICD coding, we systematically evaluate two core inference paradigms: the Retrieval-Rerank based approach and the Agent-driven hierarchical reasoning paradigm. This section details their design rationale, workflows, and alignment with the core task challenges.

3.3.1 Retrieval-Rerank Paradigm

The Retrieval-Rerank paradigm frames ICD coding as a task of information retrieval and refinement within a semantic space. Its core idea is to first efficiently retrieve candidate codes from the vast standard terminology space, followed by a more refined semantic re-ranking. We implement two representative methods under this paradigm as performance baselines.

Static Semantic Retrieval: This method serves as a baseline to evaluate foundational semantic perception, simplifying coding to a one-step semantic alignment. An LLM first extracts core diagnostic entities from the medical record to form a query vector. A search is then performed by computing similarity between text vector representations, directly returning the most relevant ICD code. It aims to test the model's capability to match anatomical details and pathological features relying solely on lexical and contextual similarity.

Interactive Retrieval & Rerank (R&R): To mitigate potential noise and local optima in direct retrieval, the Retrieval & Rerank paradigm introduces a two-stage verification process. The first stage (Retrieval) rapidly retrieves Top-K candidate codes from the entire label repository. The second stage (Rerank) invokes an LLM to perform an in-depth comparison and fine-grained ranking between the evidence in the original clinical text and the definition of each candidate code within a unified context window. This paradigm's strength lies in its robust global perspective and noise resistance, ensuring reliability when diagnostic text is explicit.

3.3.2 Agent Navigation Paradigm

As a structured complement to retrieval-based methods, we propose the **ToT-ICD (Travel on the ICD Tree)** framework. As shown in Figure 2, this framework reframes ICD coding as an active, sequential navigation process where an LLM-based agent traverses the structured ICD tree topology. The agent acts as a central controller, interacting with an external ICD knowledge base to emulate expert coders' hierarchical decision logic through iterative reasoning.

Atomic Tool Design: The agent interacts with the environment by operating an atomic **Tool Pool (Action Space)**. At each step, the agent selects from five tools based on its current state:

- **get_child_node:** Returns a list of direct child nodes of the current node.
- **select_next_node:** Moves to a selected child node based on clinical evidence such as pathology reports or surgical procedures (e.g., from D13 to D13.2).
- **validate_current_node:** Evaluates consistency between the current path and clinical facts to identify conflicts.
- **backtry_path:** Backtracks to a higher level when a path conflict is detected.
- **finish_selection:** Outputs the final ICD-10 code and the reasoning path upon reaching a validated leaf node.

Retrieval-augmented Navigation Mechanism: The framework contains two agent modes: the **Base Agent** (scans all immediate children) and the **Retrieval-Augmented Agent** (with look-ahead retrieval). The **Base Agent** returns all direct children of the current node in response to the scan-node action. The Retrieval-Augmented Agent concurrently performs a vector search within the current subtree upon receiving a clinical query. The algorithm first aggregates the top 40 nodes most relevant to the query vector, then returns a list of direct child nodes (under the current node) that lead to the most relevant leaf nodes, while also feeding back the Top-3 most relevant leaf nodes (with full code information) to the agent. As shown in Figure 2, this "look-ahead awareness" allows the agent to anticipate the potential endpoints of each branch at high-level decision points (e.g., chapters), effectively reducing path drift.

Decision process: The agent follows an "Observe-Think-Act-Validate" loop (Fig. 2). Initial clues like "abdominal pain" may lead to Chapter

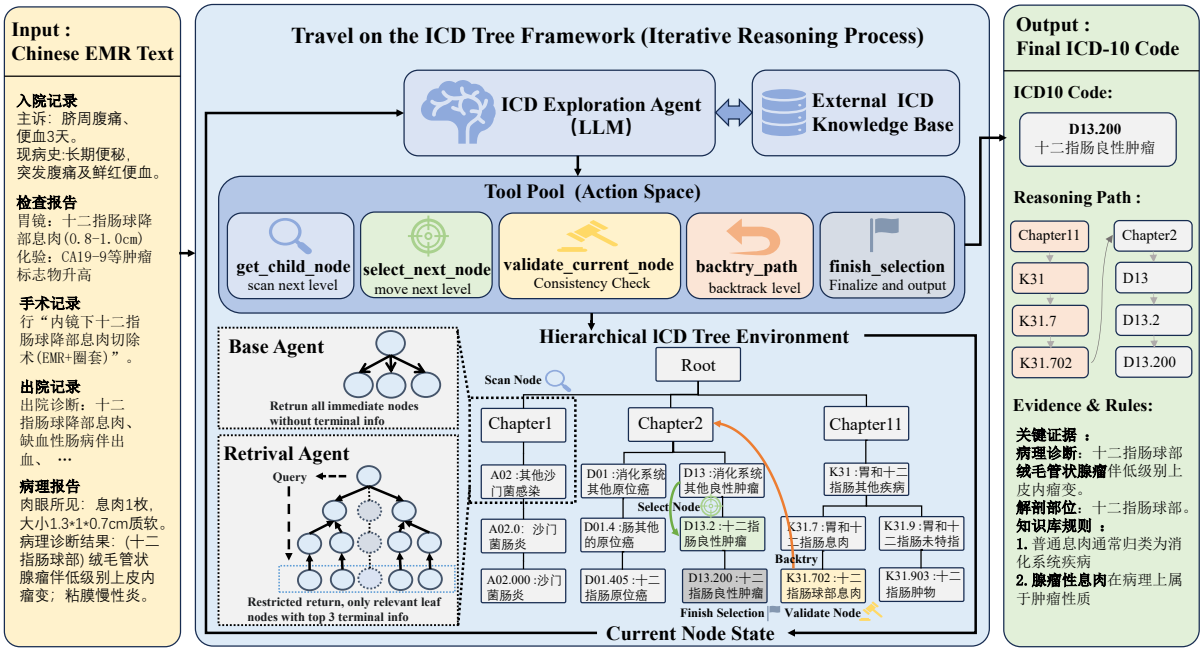

Figure 2: Overview of the *Travel on the ICD Tree* framework. The agent navigates the ICD hierarchy via iterative reasoning and defined actions to transform clinical text into precise ICD-10 codes.

11 (Diseases of the digestive system). Key findings (e.g., “villotubular adenoma” on pathology) then trigger rule validation and backtracking to the correct chapter (e.g., Chapter 2 for neoplasms).

At each step, when the agent selects a node, the system returns the node’s information along with critical, rule-based knowledge. For example, upon selecting Chapter 2 (Neoplasms) from the root, the system provides comprehensive guidance for that chapter—such as its scope (covering all benign and malignant neoplasms), exclusions (e.g., cysts are classified elsewhere), and key distinctions (e.g., neoplastic polyps like adenomas belong here, whereas inflammatory polyps are classified under digestive disease chapters). This continuous integration of specific knowledge supports precise, context-aware decision-making at every step. The implementation details for integrating these ICD rules are provided in Appendix C.

Guided by this process, the agent progressively narrows down the options and accurately assigns the final code (e.g., D13.200). Every step is documented with evidence and rule citations, ensuring full interpretability and strict guideline compliance.

4 Experiments

To evaluate the performance of large language models on complex ICD coding tasks and validate the effectiveness of different inference paradigms, we conducted a systematic experimental study on the Code4Detail benchmark. This section details the experimental setup, analyzes the results, and discusses their implications. Our study is designed to answer the following research questions: 1) How do different paradigms perform across LLMs with varying capability levels? 2) What advantages does the agent-based paradigm offer compared to retrieval-rerank paradigms? 3) How should technical pathways be selected for practical clinical ICD coding?

4.1 Experimental Setup

Benchmark and Paradigms: Experiments are conducted on the Code4Detail dataset. We compare four technical paradigms: 1) **Static Retrieval**: One-step semantic matching based on embedding vectors; 2) **Retrieval & Rerank (R&R)**: Two-stage pipeline with retrieval followed by fine-grained reranking; 3) **Base Agent**: Hierarchical tree navigation based on the ToT-ICD framework; 4) **Retrieval-Augmented Agent (ToT-ICD)**: An enhanced Agent incorporating look-ahead local retrieval.

Model Selection: To systematically analyze the relationship between paradigm efficacy and model capability, we selected seven representative models spanning a capability gradient. These comprise: 1) top-tier closed-source models *DeepSeek-V3.2* (Liu

Table 1: Performance comparison (Accuracy %) of different LLMs under Retrieval-based and Agent-based paradigms on the Code4Detail benchmark. The best and second-best results for each model are highlighted in **bold** and underlined, respectively.

Model	Retrieval-based Paradigms		Agent-based Paradigms	
	Retrieval	Retrieval & Rerank	Base Agent	ToT-ICD
Proprietary / Large-scale Models				
DeepSeek-V3.2	34.64	40.00	39.46	**43.75**
Gemini-2.5-Pro	34.46	36.07	34.82	**41.25**
Qwen3-Plus	34.11	36.25	33.75	**38.75**
Open-source / Mid-scale Models				
Qwen3-32B	33.99	**36.85**	26.25	35.00
Qwen3-30B-A3B	28.93	30.36	18.43	**31.43**
Qwen3-8B	30.89	**33.75**	19.53	29.82
Qwen2.5-14B	22.32	22.50	22.50	**29.80**

et al., 2024a) and *Gemini-2.5-Pro* (Team et al., 2023); 2) variants from the *Qwen3* series (Yang et al., 2025) (*Qwen3-Plus*, *Qwen3-32B*, *Qwen3-30B-A3B*, and *Qwen3-8B*); 3) *Qwen2.5-14B* (Yang et al., 2024) for inter-generational architectural comparison.

Implementation Details: To ensure deterministic generation, the model temperature is set to 0.0 for all experiments. Retrieval paradigms uniformly employ the BGE embedding model with a recall number of Top-k=40. In agent-based paradigms, the ICD tree depth is set to 4 levels to balance exploration completeness and computational efficiency. All paradigms share the same medical record text input and ICD knowledge base. All experiments are conducted as single runs per configuration.

4.2 Result Analysis

Table 1 presents the accuracy (%) of different models under the four paradigms. Overall, the results clearly reveal the performance boundaries between paradigms and their dependency on model capabilities. For details regarding the accuracy rates at each level, please refer to Appendix B **Superiority of the Retrieval-Augmented Agent:** Our proposed **Retrieval-Augmented Agent** paradigm achieves the best or near-best performance across all tested models. Notably, its advantage is most pronounced on top-tier reasoning models (e.g., *DeepSeek-V3.2*). This clearly demonstrates that the efficacy of the agent-based hierarchical reasoning architecture is strongly correlated with the capability of the underlying model: the more powerful the model, the more fully the agent architecture can leverage its strengths in guiding multi-step exploration and complex decision-making. **The "Capability Threshold" of the Base Agent:** The performance of the **Base Agent** paradigm shows severe polarization, highlighting its high demand for model reasoning and instruction-following capabilities. A "catastrophic collapse" on medium and small-scale models indicates that successful hierarchical navigation requires strong multi-step reasoning and state maintenance abilities.

Robustness and Ceiling of Retrieval & Rerank: The **Retrieval & Rerank** paradigm shows stable performance across models but is insensitive to model scale, exhibiting a clear "accuracy ceiling." Its performance, limited by its matching and ranking logic, is hard to surpass with more capable models. Nonetheless, it remains a practical alternative under resource or capability constraints.

Practical Implications and Paradigm Selection: Our results offer clear guidance for technical selection: 1) On models with strong reasoning, prioritize the **Retrieval-Augmented Agent** to unlock structured reasoning potential for maximum accuracy. 2) Under limited model capability, the **Retrieval & Rerank** paradigm serves as a reliable choice due to its stable performance. These findings provide key references for designing automated coding systems in real-world clinical settings.

5 Error Analysis

To better understand failure mechanisms in fine-grained clinical coding, we systematically analyzed error cases from the top-performing DeepSeek-V3 across three paradigms. Using a six-dimensional taxonomy ranging from macro-logic to micro-attributes, we aim to uncover the cognitive bottlenecks of LLMs in processing complex medical

Table 2: Phenotypic Error Taxonomy of Different ICD Coding Paradigms on Code4Detail.

Paradigm	Total	CDSE	CHC	OSH	FIM	AAS	ILD
Retrieval (R&R)	334	164 (48.2%)	24 (7.2%)	68 (20.4%)	45 (13.5%)	35 (10.5%)	1 (0.3%)
Base Agent	339	160 (47.2%)	25 (7.4%)	53 (15.6%)	57 (16.8%)	41 (12.1%)	3 (0.9%)
Retrieval Agent	313	153 (48.9%)	20 (6.4%)	57 (18.2%)	44 (14.1%)	37 (11.8%)	2 (0.6%)

Table 3: Operational Failure Mode Analysis

Paradigm	Failure Mode	Count	Ratio
Retrieval (R&R)	Retrieval Stage Missing	188	56.0%
	Retrieval Logic Error	148	44.0%
Retrieval Agent	Initial Judgment Error	142	45.4%
	Narrative Misunderstand	112	35.8%
	Navigation/Drift Error	31	9.8%
	Validation Conflict	27	8.6%
	System/Process Error	3	1.0%

data. Additionally, we present two representative case studies in Appendix D. Based on the logical chain of clinical decision-making, we categorize errors into the following six types:

1. **Core Disease Selection Error (CDSE):** The model predicts a disease entity or affected organ fundamentally different from the ground truth.

2. **Clinical History Confusion (CHC):** The model misclassifies past medical history or admission background information as an active diagnosis for the current hospitalization.

3. **Over-specification Hallucination (OSH):** While the primary disease category is correct, the model fabricates location or subtype details not present in the medical record.

4. **Fine-grained Information Missing (FIM):** The medical record contains explicit evidence for specific details, but the model selects an overly broad code such as "unspecified" or "other."

5. **Attribute Association Shift (AAS):** The model identifies disease details but incorrectly associates them (e.g., predicting a left-sided lesion for a right-sided condition).

6. **Insufficient Level Depth (ILD):** The model's general direction is correct, but stops at a chapter or 3-digit category level without drilling down to a leaf-node code.

Analysis of the six-dimensional distribution in Table 2 reveals two key phenomena:

1. **The Fidelity-Hallucination Trade-off:** Specifically, The Retrieval & Rerank (R&R) paradigm shows a notably higher rate of *Over-specification Hallucination* (20.4%), stemming from vector retrieval's clear bias towards richly described labels for semantic fitting. Agent-based paradigms, conversely, effectively curb such fabrication via evidence-anchoring mechanisms.

2. **Contrast in Logical Robustness:** In addition, the Retrieval Agent achieves the lowest error rates in both *Clinical History Confusion* (6.4%) and *Fine-grained Information Missing* (14.1%), indicating that its look-ahead mechanism clearly harmonizes the breadth of retrieval with the depth of navigational reasoning in an integrated fashion.

To further investigate the internal mechanisms of each paradigm, we analyze vulnerable points in its decision logic. Table 3 compares the operational failure modes of the R&R paradigm and the Retrieval Agent paradigm. **Inherent Bottleneck of R&R: Retrieval Blind Spots.** Analysis of 336 R&R failure cases shows that over half of the errors (56.0%) occur because the ground truth is absent from the initial retrieval set. This indicates that the correct answer is lost at the initial retrieval stage: specific modifiers are easily overshadowed by the primary disease name, imposing a fundamental performance ceiling on the R&R paradigm.

Decision Chain Analysis of Agent Navigation. For the ToT-ICD framework, we deconstruct failure cases by decision stage. Compared to the Base Agent, the Retrieval Agent shows improvements in key areas: 1) **Correction of Path Errors:** The Base Agent has a 9.8% Navigation error rate. The Retrieval Agent reduces this to 8.6% via its look-ahead perception. 2) **Improved Decision Consistency:** The Base Agent exhibits a 10.9% Validation Conflict error rate. By injecting concrete underlying evidence as "decision anchors," the Retrieval Agent reduces this to 8.6%, significantly boosting the model's confidence.

Table 4: Impact of Top-k retrieval count on the overall accuracy of the Retrieval & Rerank (R&R) paradigm.

Top-k	**10**	**20**	**30**	**40**	**50**
Accuracy (%)	33.99	35.24	36.96	**40.00**	37.39

6 Discussion

To evaluate ICD coding paradigms' practical application, this section examines computational overhead and hyperparameter sensitivity, addressing cost-accuracy trade-offs and retrieval optimization; additional details are provided in the Appendix.

6.1 Cost-Accuracy Trade-off in Practice

We evaluated accuracy versus computational cost across paradigms. The Retrieval & Rerank (R&R) baseline is efficient (2 steps, 32K tokens/case), while agents incur higher costs: Base Agent (14 steps, 323K tokens) and ToT-ICD (18 steps, 408K tokens). This trade-off is justified in high-stakes medical coding, where accuracy directly affects reimbursement and DRG assignment. Even marginal accuracy improvements prevent costly claim denials, penalties, and audits, making inference costs negligible. Given that error-related financial and legal risks far exceed computational expenses, high-precision agentic workflows are both necessary and feasible for clinical deployment.

6.2 Optimal Retrieval Size in R&R

We conducted the analysis of retrieval size impact based on DeepSeek-V3.2, the best-performing model. In the R&R paradigm, the initial retrieval count (Top-k) is a critical hyperparameter (default 40). As shown in Table 4, accuracy improves as k increases from 10 to 40, attributed to higher recall from a larger candidate pool. However, when k exceeds 40 (e.g., $k = 50$), performance declines to 37.39%. This indicates that an excessively large candidate set introduces noise that interferes with the LLM's reranking judgment, leading to performance degradation.

7 Conclusion

This paper addresses fine-grained ICD coding challenges by introducing the **Code4Detail** benchmark and **ToT-ICD** framework. Experiments reveal distinct, complementary strengths between the two paradigms. The **Retrieval-Rerank** method proves highly robust (33.75% accuracy on Qwen3-8B), effectively handling explicit cases but plateauing with model scaling. Conversely, the **ToT-ICD Agent** unlocks superior precision when powered by strong reasoners, peaking at 43.75% with DeepSeek-V3.2. Consequently, we recommend retrieval paradigms for stable, resource-constrained deployments and agentic workflows for complex, high-precision clinical scenarios.

Limitations

Despite the proposed benchmark and novel paradigms, this work has several limitations.

Limited Dataset Coverage. The Code4Detail benchmark is constructed by sampling real-world medical records. While its disease spectrum aims to reflect common clinical practice, coverage of low-incidence rare diseasaremains insufficient.

Performance Bottleneck Persists. Even with advanced agent paradigms, absolute accuracy leaves room for improvement. *Core Disease Selection Error* remains the predominant failure mode, highlighting that anchoring the core clinical problem from complex narratives is still a fundamental challenge.

Scope and Transferability. Our experiments focus on Chinese ICD-10. Since ICD-11 employs a multi-axial structure and clinical conventions vary across languages, our findings require targeted validation before transferring to other contexts.

Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers for their constructive suggestions. We also thank the Shanghai Municipal Health Commission for their valuable assistance.

Ethical Considerations

The Code4Detail benchmark fully complies with ethical standards. To ensure privacy, all data underwent strict de-identification, including date shifting, ID reconstruction, and semantic text regeneration via locally deployed LLMs. All processing occurred within a secure internal network, guaranteeing the absence of personal or sensitive information.

Annotation was conducted by collaborating medical records professionals. All annotators participated voluntarily with informed consent and were fairly compensated for their time. The resulting dataset is strictly reserved for academic research purposes.

References

Krishanu Das Baksi, Elijah Soba, John J Higgins, Ravi Saini, Jaden Wood, Jane Cook, Jack I Scott, Nirmala Pudota, Tim Weninger, Edward Bowen, and 1 others. 2025. Medcoder: A generative ai assistant for medical coding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 449–459.

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):e0173410.

Zhenghan Chen, Changzeng Fu, Ruoxue Wu, Ye Wang, Xunzhu Tang, and Xiaoxuan Liang. 2023. Lgfat-rgcn: Faster attention with heterogeneous rgcn for medical icd coding generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5428–5435.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2025a. Medodyssey: A medical domain benchmark for long context evaluation up to 200k tokens. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 32–56.

Yongqi Fan, Nan Wang, Kui Xue, Jingping Liu, and Tong Ruan. 2025b. Medeureka: A medical domain benchmark for multi-granularity and multi-data-type embedding-based retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2825–2851.

Yongqi Fan, Yansha Zhu, Kui Xue, Jingping Liu, and Tong Ruan. 2024. Rrnorm: A novel framework for chinese disease diagnoses normalization via llm-driven terminology component recognition and reconstruction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9162–9175.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Guangya Yu, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: A benchmark for evaluating large language models in multi-step clinical diagnosis. *arXiv preprint arXiv:2408.10039*.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. Plm-icd: Automatic icd coding with pretrained language models. *arXiv preprint arXiv:2207.05289*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Leah S Larkey and W Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024b. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2023. Automated icd coding using extreme multi-label long text transformer-based models. *Artificial Intelligence in Medicine*, 144:102662.

Zhanyu Liu, Shiyao Wang, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng, Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu, and 1 others. 2025. Onerec-think: In-text reasoning for generative recommendation. *arXiv preprint arXiv:2510.11639*.

Andreas Motzfeldt, Joakim Edin, Casper L Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. 2025. Code like humans: A multi-agent solution for medical coding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22612–22627. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Hugo Silva, Vítor Duque, Mário Macedo, and Mateus Mendes. 2024. Aiding icd-10 encoding of clinical health records using improved text cosine similarity and plm-icd. *Algorithms*, 17(4):144.

Yaoqian Sun, Lei Sang, Dan Wu, Shilin He, Yani Chen, Huilong Duan, Han Chen, and Xudong Lu. 2024. Enhanced icd-10 code assignment of clinical texts: A summarization-based approach. *Artificial Intelligence in Medicine*, 156:102967.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, P. A. Mansfield, Sushant Prakash, Renee C Wong, Sunny Virmani, and 13 others. 2023. Towards generalist biomedical ai. *ArXiv*, abs/2307.14334.

Xindi Wang, Robert Mercer, and Frank Rudzicz. 2024. Multi-stage retrieve and re-rank model for automatic medical coding recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4891.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

World Health Organization. 2004. *International Statistical Classification of Diseases and related health problems: Alphabetical index*. World Health Organization.

Yuzhou Wu, Zhigang Chen, Xin Yao, Xuechen Chen, Zeren Zhou, and Jinkai Xue. 2022. Jan: Joint attention networks for automatic icd coding. *IEEE journal of biomedical and health informatics*, 26(10):5235–5246.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2023. Multi-label few-shot icd coding as autoregressive generation with prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5366–5374.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629.

Lingling Zhou, Cheng Cheng, Dong Ou, and Hao Huang. 2020. Construction of a semi-automatic icd-10 coding system. *BMC medical informatics and decision making*, 20(1):67.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957.

Table 5: Controlled comparison of correctly predicted cases (out of 30) before and after semantic rewriting to assess potential distribution bias. Values are shown as *Before* / *After* / Δ.

Model	Retrieval			Retrieval & Rerank			Base Agent			ToT-ICD		
	Before	After	Δ	Before	After	Δ	Before	After	Δ	Before	After	Δ
DeepSeek-V3.2	11	11	0	11	12	+1	16	17	+1	16	18	+2
Qwen3-Plus	10	11	+1	13	14	+1	15	15	0	13	14	+1
Qwen3-32B	9	10	+1	11	14	+3	8	10	+2	13	13	0
Qwen3-8B	8	10	+2	12	14	+2	8	8	0	7	8	+1

A Preservation of Data Distribution

During the construction of the Code4Detail benchmark, we utilized a locally deployed Qwen3-32B model to perform semantic-level rewriting of de-identified EMRs, aiming to remove stylistic features that might compromise privacy while preserving core clinical facts. To verify whether this process introduced distribution bias, we conducted a controlled comparison on a subset of 30 randomly selected cases. As shown in Table 5, all differences (Δ) across model-paradigm combinations fall within a narrow range of [0, +3]. This confirms that the semantic rewriting process does not cause information loss or detrimental distribution shifts; slight performance gains likely stem from the reduction of textual noise.

B Detailed Analysis of Performance Decay Across ICD Hierarchy

Table 6 presents a granular breakdown of model performance across the four hierarchical levels of the ICD-10 taxonomy. While a general decline in accuracy is expected as classification granularity shifts from broad Chapters (Level 1) to specific Full Codes (Level 4), the data highlights specific structural bottlenecks in model reasoning.

B.1 Identification of the Primary Performance Bottleneck

The most significant performance degradation across nearly all tested models occurs during the transition from **Level 1 (Chapter)** to **Level 2 (Category)**.

- **The Steepest Drop Phenomenon:** For high-performing models such as `DeepSeek-V3.2` (*Retrieval Agent*), the accuracy drops by 16.25% ($81.61\% \rightarrow 65.36\%$). In smaller-scale models like `Qwen2.5-14B`, this gap exceeds 22%, representing the largest single-step decline in the hierarchy.
- **Structural Complexity:** Unlike the transition from Level 3 to Level 4, which often involves specific clinical modifiers, the leap from Level 1 to Level 2 requires the model to narrow down from a general physiological system (e.g., Circulatory System) to a distinct disease category.

B.2 Implications for Model Reliability

This sharp decline at Level 2 suggests that LLMs encounter a "semantic resolution" limit. While the models possess sufficient global medical knowledge to identify the correct anatomical chapter, they struggle to resolve the specific category when faced with multiple clinically similar codes. This indicates that:

1. **Feature Overlap:** The models may lack the fine-grained discriminative power to distinguish between neighboring categories within the same chapter.
2. **Effectiveness of Retrieval-Augmentation:** The superior performance of the *Retrieval Agent* compared to the *Base Agent* at Levels 2-4 demonstrates that narrowing the search space is critical. This confirms that the bottleneck at Level 2 is primarily a selection challenge that can be mitigated by external knowledge retrieval.

C Knowledge Binding in Tree-based ICD Coding

Different from traditional vector-database-based Retrieval-Augmented Generation (RAG), our proposed method, termed *Travel on the ICD Tree*, directly binds medical record informatics knowledge and coder's guideline rules to specific nodes within the ICD tree structure.

For example, within Chapter 9 of ICD-10, we bind priority rules that resolve diagnostic ambiguity. The rule `Myocardial infarction` → `Angina pectoris` → `Atherosclerosis` → `Chronic`

Table 6: Detailed performance (Accuracy %) across four ICD hierarchy levels. Level 1 to Level 4 represent Chapter, Category, Subcategory, and Full Code respectively. The best and second-best results for each model at each level are highlighted in **bold** and underlined, respectively.

Model	Paradigm	Level 1 (Chapter)	Level 2 (Category)	Level 3 (Subcategory)	Level 4 (Full Code)
DeepSeek-V3.2	Retrieval	76.25	56.79	41.79	34.64
	Retrieval & Rank	78.75	63.57	48.75	40.00
	Base Agent	80.54	64.29	46.96	39.46
	Retrieval Agent	**81.61**	**65.36**	**51.07**	**43.75**
Gemini-2.5-Pro	Retrieval	80.36	57.50	41.96	34.46
	Retrieval & Rank	**81.43**	63.75	46.61	36.07
	Base Agent	79.82	62.50	45.36	34.82
	Retrieval Agent	80.54	**65.18**	**50.71**	**41.25**
Qwen3-Plus	Retrieval	76.43	54.64	41.07	34.11
	Retrieval & Rank	78.57	62.68	46.79	36.25
	Base Agent	**78.93**	61.25	43.57	33.75
	Retrieval Agent	**78.93**	**63.04**	**47.86**	**38.75**
Qwen3-32B	Retrieval	75.13	55.81	42.22	33.99
	Retrieval & Rank	**79.07**	**63.86**	**47.58**	**36.85**
	Base Agent	73.57	52.86	33.93	26.25
	Retrieval Agent	76.61	57.68	42.14	35.00
Qwen3-30B-A3B	Retrieval	68.93	48.39	36.43	28.93
	Retrieval & Rank	73.04	56.96	40.18	30.36
	Base Agent	72.81	49.19	29.34	18.43
	Retrieval Agent	**75.71**	**58.21**	**41.61**	**31.43**
Qwen3-8B	Retrieval	73.57	52.32	38.21	30.89
	Retrieval & Rank	**75.54**	**59.29**	**44.64**	**33.75**
	Base Agent	69.71	43.55	26.70	19.53
	Retrieval Agent	72.32	53.75	39.46	29.82
Qwen2.5-14B	Retrieval	51.43	35.89	27.68	22.32
	Retrieval & Rank	62.14	44.11	31.07	22.50
	Base Agent	68.75	42.50	28.39	22.50
	Retrieval Agent	**71.45**	**48.83**	**34.83**	**29.80**

`ischemic heart disease` is attached to circulatory disease nodes.

Our experimental evaluation reveals a performance bottleneck: regardless of the methodology employed, *Core Disease Selection Error* remains the challenging failure mode. This underscores the fundamental difficulty of precisely identifying the principal clinical condition from complex medical narratives.

However, the dynamic binding of coding rules to tree nodes provides a promising path to address this challenge. By continuously expanding and refining these rules, we leverage large language models' powerful reasoning capabilities to navigate nuanced clinical scenarios. The system functions as an adaptive prompting framework, where rule sets are progressively accumulated and injected into the model's decision context during tree traversal.

A representative case illustrates the effectiveness of this mechanism:

Clinical Course: The patient received antiplatelet therapy, statins for plaque stabilization, and coronary angiography on August 6, 2024, showing diffuse LAD lesions with 80% stenosis. A stent was implanted in the proximal LAD lesion. Postoperative recovery was uneventful.

Discharge Diagnosis: 1. Coronary atherosclerotic heart disease, exertional angina pectoris

Procedure: Single-catheter coronary angiography

Without the integrated priority rule, models frequently assign the primary code to atherosclerosis based on surface-level patterns. However, through rule-guided reasoning within the *Travel on the ICD Tree* framework, the system correctly identifies the clinical priority, assigning the principal code to angina pectoris. This demonstrates how structured rule injection helps overcome the persistent challenge of core disease selection.

This node-specific knowledge binding represents a dynamic prompting methodology that evolves with coding practice. By systematically expanding rules, we enable models to better navigate the intricacies of ICD coding, gradually overcoming the fundamental challenge of accurately identifying the core disease from multifaceted medical records.

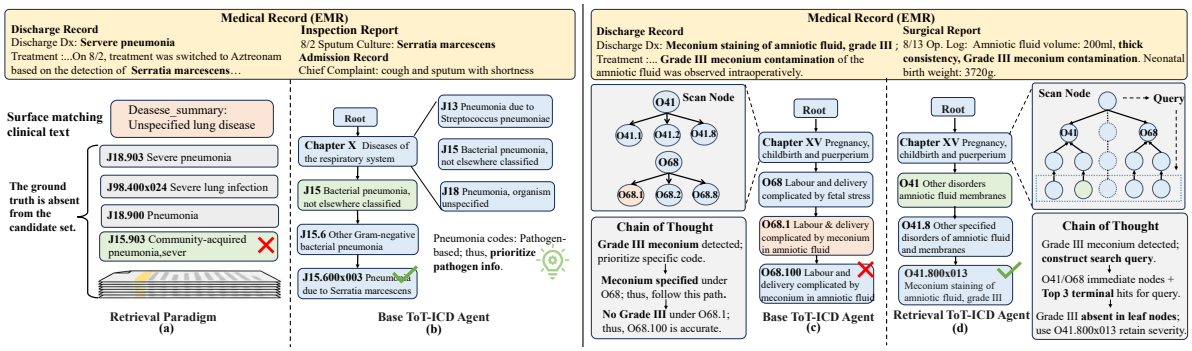

Figure 3: Case studies demonstrating the framework's iterative improvements. Case 1 (left) shows the Base Agent capturing pathogen details missed by traditional retrieval, while Case 2 (right) illustrates how the Retrieval-enhanced Agent further refines precision by identifying specific severity grades (e.g., Grade III).

D Case Studies

To delve into the decision-making mechanisms of different technical paradigms in real clinical scenarios, we present a comparative analysis of two highly representative cases of ultra-fine-grained coding (Fig. 3).

Case 1: Specific Evidence Extraction and Hierarchical Drill-down. A patient was admitted with "severe pneumonia," with treatment records specifying the pathogen *Serratia marcescens*. The retrieval paradigm, based on semantic summarization of the record, failed to capture the pathogen detail in the lab report due to the high salience of "severe pneumonia." As a result, global retrieval matched only the semantically similar code J15.903, while the correct code J15.600x003 was absent from the candidate set (Fig. 3a).

In contrast, the Base Agent framed coding as a state-machine search in a constrained environment. After entering the respiratory diseases chapter, it recognized the pathogen-driven nature of pneumonia coding, actively backtracked via chain-of-thought to locate the key evidence "*Serratia marcescens*," and used the selection tool to map the generalized diagnosis precisely to leaf-node code J15.600x003 (Fig. 3b).

Case 2: Rule Conflict in High-level Decision and Proactive Correction. An obstetric record documented "amniotic fluid, grade III meconium staining," corresponding to two ICD paths: O68 (fetal distress with meconium) or O41 (disorders of amniotic fluid). The Base Agent was misled by keywords during high-level decisions: the term "meconium" in category O68.1 led it down that branch. Limited to viewing only immediate child nodes, it found that leaf node O68.100 matched the "meconium" feature but lacked severity grading, thus falling into a local optimum and causing path deviation (Fig. 3c).

The Retrieval-enhanced Agent overcame this via its look-ahead retrieval tool. At the chapter level, it concurrently retrieved the subtree and obtained the Top-3 terminal nodes most similar to "grade III meconium." The look-ahead information showed that leaf nodes under O68 lacked severity precision, whereas a highly specific leaf node, O41.800x013, existed under O41. Using this insight, the agent actively avoided the O68 branch and navigated directly to the O41 path, achieving leaf-node-level specificity (Fig. 3d).

These cases demonstrate that leveraging the breadth of global retrieval to guide hierarchical drill-down can effectively preserve key severity information in medical records, thereby improving coding precision.

E Detailed Design of Agent Tools

The proposed **Travel on the ICD Tree (ToT-ICD)** framework utilizes specific tool definitions to guide the agents. The detailed configurations are presented in Figures 4 and 5.

Figure 4 illustrates the standard topology-based definition used by the Base Agent. In contrast, the Retrieval-Augmented Agent introduces semantic search capabilities. As shown in Figure 5, the `get_child_node` tool is modified to include a `query` parameter. This modification implements **"Look-ahead Awareness,"** enabling the agent to anticipate potential leaf-node endpoints at high-level decision points (e.g., Chapter level), thereby effectively mitigating path drift.

Base ToT-ICD Agent Tools

```
  {
      "name": "get_child_node",
      "description": "获取指定节点下的所有子节点信息。node_id 必须从系
统已记录的 node_list 中选择（node_list 包含用户此前通过本工具或路径选择
所查看过的节点及其子节点）。",
      "parameters": {
         "type": "object",
         "properties": {
            "node_id": {
            "type": "string",
            "description": "指定节点的node_id，必须属于系统当前维护的
node_list 范围内"
            }
         },
         "required": ["node_id"],


      }
   },
   {
      "name": "select_next_node",
      "description": "该步骤通过选定子节点来更新当前节点，从而进入编码
树的下一层级。",
      "parameters": {
         "type": "object",
         "properties": {
            "selected_node_id": {
               "type": "string",
               "description": "所选节点ID必须来自当前节点的子节点列表，
不可选择列表外的节点。"
            },
            "evidence_quote": {
               "type": "string",
               "description": "【关键证据】支持从当前父节点进入该子节点
的电子病历原文片段。"
            },
            "rule_quote": {
               "type": "string",
               "description": "【规则依据】支持本次选择的ICD编码规则或
优先原则。"
            }
         },
         "required": ["selected_node_id","evidence_quote","rule_quote"]
      }
   }
```

Figure 4: **Base ToT-ICD Agent Tools Design.**

Base ToT-ICD Agent Tools

```
{
      "name": "validate_current_node",
      "description": "基于电子病历内容，评估当前选择的编码节点是否合理。",
      "parameters": {
         "type": "object",
         "properties": {},
         "required": [],
      }
   },
   {
      "name": "backtry_path",
      "description": "该步骤通过选定回退层级来更新当前节点，从而返回到编码
树的指定层级。",
      "parameters": {
      "type": "object",
      "properties": {
            "level": {
            "type": "integer",
            "description": "指定应回退到的目标层级（0=根节点，1=章节，2=类
目，3=亚目）",
            "minimum": 0,
            "maximum": 3
            }
         },
         "required": ["level"],
      }
   },
   {
      "name": "finish_selection",
      "description": "结束编码选择流程。仅当当前层级为最终编码（level=4）
且已通过至少一次节点校验（即 validate_current_node）时允许调用。",
      "parameters": {
         "type": "object",
         "properties": {
            "node_id": {
            "type": "string",
            "description": "最终选定节点的编号",
            },
            "name": {
            "type": "string",
            "description": "最终选定节点的名称",
            }
         },
      "required": ["node_id","name"],
      }
   }
```

Figure 4: **Base ToT-ICD Agent Tools Design (continued).**

Base ToT-ICD Agent :Scan Node Tool

```
{
     "name": "get_child_node",
     "description": "获取指定节点下的所有子节点信息。node_id 必须从系统已记录的 node_list 中选择（node_list 包含用户此前通过本工具或路径选择所查看过的节点及其子节点）。",
     "parameters": {
       "type": "object",
       "properties": {
         "node_id": {
         "type": "string",
         "description": "指定节点的node_id，必须属于系统当前维护的 node_list 范围内"
          }
       },
       "required": ["node_id"],


     }
  }
```

Retrieval ToT-ICD Agent:Scan Node Tool

```
  {
    "name": "get_child_node",
     "description": "获取指定节点下的子节点及相似ICD编码的聚合信息。该工具根据用户提供的 query，在当前 node_id 所属分类范围内进行名称相似度检索，返回与 query 最相关的子节点列表、每个分类的描述以及其下最相似的3个末端编码（最终节点）信息。",
     "parameters": {
       "type": "object",
       "properties": {
         "node_id": {
           "type": "string",
           "description": "指定节点的 node_id。必须从系统当前已展示或记录的 node_list 中选择（例如 'root'、'第一章' 或具体的类目编码）。"
         },
         "query": {
           "type": "string",
           "description": "临床诊断描述或搜索关键词。系统将根据此信息在 ICD 编码库中执行相似度召回，以辅助定位最准确的编码。"
         }
       },
       "required": ["node_id", "query"]
     }
  }
```

Figure 5: **Tool Schema Differences.** The Retrieval Agent adds a `query` parameter to support semantic look-ahead.