

R³-MDT: A Retain-Rerank-Revise Framework Based on Uncertainty and Consistency for Medical Decision Tree Extraction

Boyu Qiu¹ Xiaolong Weng¹ Zehua Wang¹
Buzhou Tang^{1,2,*} Ying Xiong^{1,*} Min Zhang¹

¹Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China

²Peng Cheng Laboratory, Shenzhen, China

24S151107@stu.hit.edu.cn, tangbuzhou@gmail.com

Abstract

Automatic extraction of Medical Decision Trees (MDTs) is an important task for Clinical Decision Support Systems, but remains challenging due to the complex logical structures involved. Recent approaches based on Large Language Models (LLMs) commonly adopt generate-then-revise pipelines with self-correction to improve reasoning accuracy. However, we find that applying revision uniformly across all samples can be harmful, as it often alters already correct and stable predictions, leading to over-correction on simple cases. To alleviate this issue, we propose R³-MDT, an adaptive inference framework that dynamically selects reasoning strategies based on sample difficulty. We use model uncertainty and output consistency as internal signals to estimate difficulty and route samples into Easy, Medium, and Hard categories. For Easy samples with confident and consistent predictions, we directly retain the initial outputs. Medium-difficulty samples, which exhibit ambiguity among multiple plausible candidates, are handled via reranking to select the optimal decision tree. Hard samples, characterized by high uncertainty and structural inconsistency, are revised to rectify intractable generation failures. By adaptively selecting inference strategies based on difficulty levels, R³-MDT avoids unnecessary revisions while effectively correcting underlying reasoning failures. Experiments on the Text2DT benchmark show that R³-MDT alleviates the over-correction problem and achieves state-of-the-art performance, while auxiliary language- and task-shift evaluations indicate the broader transferability of its routing principle.

1 Introduction

Medical Decision Trees (MDTs) are core components of Clinical Decision Support Systems (CDSS), representing interpretable clinical logic in

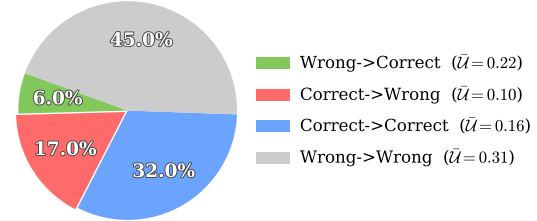


Figure 1: Distribution of inference outcomes under the indiscriminate *Revise* strategy. **Over-correction** (red) predominantly occurs in samples with low uncertainty ($\bar{U}=0.10$), while **Improvement** (green) is concentrated in high-uncertainty regions ($\bar{U}=0.22$).

hierarchical structures (Zhu et al., 2023; Shortliffe and Sepúlveda, 2018). However, manual construction by experts is labor-intensive and difficult to scale (Grosan and Abraham, 2011; Shortliffe and Sepúlveda, 2018), making automated MDT extraction a critical task.

While Large Language Models (LLMs) show advanced reasoning and structured generation capabilities (Chen et al., 2021; Wei et al., 2022; Achiam et al., 2023), their one-pass outputs often require inference-time refinement such as Self-Correction or Reranking (Wang et al., 2022; Ravaut et al., 2022). However, recent studies in Information Extraction (IE) indicate that these strategies can suffer from **over-correction** (Ma et al., 2023), where the model wrongfully alters correct entities during refinement.

To verify this risk, we conducted a pilot experiment on the Text2DT benchmark (Zhu et al., 2023). As shown in Figure 1, applying an indiscriminate generate-then-revise pipeline led to a significant decline in Tree Accuracy. The revision stage corrupted significantly more samples in low-uncertainty regions (17.0%) than it improved in high-uncertainty ones (6.0%), confirming that the key lies not in *how* but *when* to refine.

This observation connects our problem to a classical control and fault-diagnosis perspective: reli-

*Corresponding authors

able systems should not apply corrective actions uniformly, but should first monitor residual-like signals that reveal deviations from expected behavior (Frank, 1990; Blanke et al., 2006). In LLM inference, we treat token-level uncertainty and structural consistency as such diagnostic signals, enabling the system to detect distinct failure modes and decide whether to preserve, select, or revise an output.

We thus propose **R³-MDT**, a **Retain-Rerank-Revise** framework that adaptively routes samples based on difficulty. Utilizing token-level uncertainty and structural consistency as internal probes, we stratify samples into: (1) **Retain** for Easy samples to preserve high-confidence predictions; (2) **Rerank** for Medium samples to select the optimal structure from candidates; and (3) **Revise** for Hard samples to reconstruct decision trees from the source text.

Our main contributions are as follows:

- **Difficulty-aware diagnostic routing mechanism:** We propose a routing mechanism that treats token-level uncertainty and structural consistency as residual-like internal probes for failure mode detection, enabling accurate stratification of samples based on reasoning complexity.
- **The R³-MDT framework:** We introduce an inference-time framework that adaptively allocates refinement strategies (Retain, Rerank, and Revise). This mechanism simultaneously prevents the over-correction of simple samples and rectifies complex errors.
- **State-of-the-art performance:** Extensive experiments on Text2DT (Zhu et al., 2023) — *the sole public benchmark annotated for this task*—demonstrate that our method achieves superior accuracy while maintaining high inference efficiency. Furthermore, auxiliary evaluations under language and task shifts show the broader transferability of our routing principle.

2 Methodology

2.1 Problem Formulation

The task of MDT extraction aims to transform an unstructured clinical text $X = \{x_1, x_2, \dots, x_n\}$ into a structured binary tree T . Formally, an MDT consists of a set of nodes V , where each node $v \in V$ is defined as a tuple: $v = (\text{Role}, \mathcal{T}, \text{Logic})$.

Here, $\text{Role} \in \{\text{CONDITION}, \text{DECISION}\}$ distinguishes internal judgment nodes from leaf treatment nodes. $\mathcal{T} = \{(s, r, o)_k\}_{k=1}^m$ is a set of subject-relation-object triplets capturing the medical semantics (e.g., *Patient-Has-Fever*). $\text{Logic} \in \{\text{AND}, \text{OR}, \text{NULL}\}$ specifies the logical relationship among multiple triplets within a single node.

Following standard practices (Zhu et al., 2023; Vinyals et al., 2014), we linearize the tree structure T into a token sequence $Y = [y_1, y_2, \dots, y_m]$ via pre-order traversal. Our primary objective is to model the conditional probability $P(Y|X)$ using an LLM. To this end, we employ the R³ framework as an inference-time refinement mechanism to dynamically verify and rectify the generated hypothesis, thereby maximizing the overall extraction accuracy.

2.2 Candidate Generator

Given the input text X , this module aims to explore the hypothesis space to provide a diverse set of candidate decision trees. We employ a decoding strategy (e.g., Beam Search) to generate the top- K candidate paths, denoted as $\mathcal{H} = \{Y^{(1)}, Y^{(2)}, \dots, Y^{(K)}\}$, where $Y^{(1)}$ represents the highest-probability hypothesis. This candidate set $\mathcal{H}$ serves as the foundation for the subsequent estimation of token-level uncertainty and structural consistency.

2.3 Metric Evaluator

To evaluate the generated candidates, we introduce two metrics specifically designed to detect local reasoning failures and structural ambiguities that are often masked by high-confidence functional tokens in long sequences.

Uncertainty Estimation. We measure the token-level uncertainty of the top hypothesis $Y^{(1)}$. While perplexity is a common metric, we reject it due to the *length bias* inherent in auto-regressive models (Vashurin et al., 2025; Kuhn et al., 2023). In MDT extraction, long sequences often contain many high-confidence functional tokens that artificially inflate the average probability, masking low-confidence errors at crucial early decision nodes. However, in medical contexts, the reliability of a clinical pathway is governed by its most vulnerable component (i.e., the least confident token), rather than the average log-likelihood. Therefore, we utilize

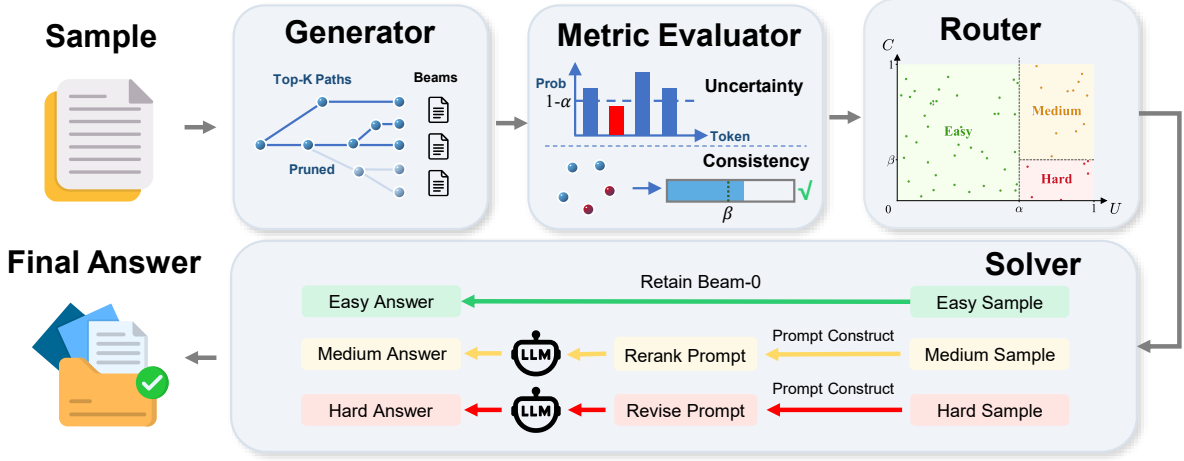


Figure 2: The overall architecture of R^3 -MDT. The framework operates in three stages: (1) **Estimate**, where the Candidate Generator produces hypotheses and the **Metric Evaluator** computes token-level uncertainty ($\mathcal{U}$) and structural consistency ($\mathcal{C}$); (2) **Route**, which stratifies samples into Easy, Medium, and Hard levels based on these metrics; and (3) **Solve**, which adaptively applies Retain, Rerank, or Revise strategies to optimize the trade-off between efficiency and accuracy.

the minimum token probability criterion:

$$\mathcal{U}(X) = 1 - \min_{1 \leq t \leq |Y^{(1)}|} P(y_t^{(1)} | X, y_{<t}^{(1)}) \quad (1)$$

where a higher $\mathcal{U}$ indicates a fatal local flaw that would otherwise be masked by the high confidence of surrounding functional tokens in long sequences. We provide an empirical correlation analysis in Appendix A, confirming that a lower minimum token probability strongly correlates with a higher reasoning error rate, thus validating its use for sample routing.

Consistency Quantification. We verify the structural agreement within the candidate set $\mathcal{H}$ to diagnose the *nature* of the model’s uncertainty. Our hypothesis is that resolvable ambiguity (e.g., oscillating between two plausible clinical concepts) typically manifests as a consistent divergence point across candidates. Conversely, chaotic reasoning results in scattered divergence, indicating a deeper structural failure. For each candidate $Y^{(k)} \in \mathcal{H}$, let m_k be the index of the token with the minimum probability:

$$m_k = \arg \min_{1 \leq t \leq |Y^{(k)}|} P(y_t^{(k)} | X, y_{<t}^{(k)}) \quad (2)$$

We define the consistency score $\mathcal{C}(X)$ as the frequency of the mode within the set of indices $\{m_1, \dots, m_K\}$:

$$\mathcal{C}(X) = \frac{\sum_{k=1}^K \mathbb{I}(m_k = \text{Mode}(\{m_j\}_{j=1}^K))}{K} \quad (3)$$

Intuitively, $\mathcal{C}(X)$ measures the dominance of the primary divergence point, representing the propor-

tion of candidates that align on the most frequent bottleneck location.

To accommodate alignment shifts caused by trivial tokenization artifacts, we employ a **relaxed matching criterion**. Specifically, two indices m_i and m_j are considered consistent if $|m_i - m_j| \leq \delta$. We set $\delta = 2$, which empirically corresponds to the typical span of non-semantic tokens (e.g., delimiters) that cause alignment jitter.

A high $\mathcal{C}$ implies that candidates diverge at the *same* node, forming a “multiple-choice” scenario where the correct path likely exists within $\mathcal{H}$, thus making it suitable for the Rerank strategy. Conversely, a low $\mathcal{C}$ indicates chaotic divergence, suggesting the model is lost in a vast solution space, which necessitates the Revise strategy.

2.4 Sample Router

The Router acts as the triage module of our R^3 framework, mapping each input sample X into a difficulty space defined by its uncertainty $\mathcal{U}(X)$ and consistency $\mathcal{C}(X)$. To adaptively allocate inference resources, we introduce two hyperparameters: an uncertainty threshold α and a consistency threshold β . Based on these metrics, the Router stratifies samples into three distinct difficulty levels:

Easy ($\mathcal{U} < \alpha$): The model exhibits high confidence, requiring no intervention. **Medium** ($\mathcal{U} \geq \alpha, \mathcal{C} \geq \beta$): High local uncertainty but consistent global structure suggests resolvable ambiguity. **Hard** ($\mathcal{U} \geq \alpha, \mathcal{C} < \beta$): High uncertainty and divergent structures indicate deep hallucinations,

necessitating revision.

2.5 Adaptive Solver

Guided by the Router’s stratification, the Solver executes one of three adaptive inference strategies to ensure robust extraction across varying levels of complexity.

Retain Strategy (for Easy samples). For Easy cases, efficiency is paramount. Since the model is already confident, heavy post-processing yields diminishing returns. Therefore, we simply **Retain** the top-ranked hypothesis $Y^{(1)}$ from the generator.

$$\hat{Y} = Y^{(1)} \quad (4)$$

This "fast path" ensures minimal latency for the majority of Easy samples.

Rerank Strategy (for Medium samples). Medium cases exhibit high consistency, suggesting the correct logic resides within the candidate pool $\mathcal{H}$ despite potential errors in the top-ranked $Y^{(1)}$. We employ an Assistant LLM (e.g., Qwen2.5-72B-Instruct) to critically evaluate the top- K candidates via a *Rerank Prompt*. By comparing these options against the input context, the Assistant identifies and selects the optimal decision tree aligned with the medical context, resolving local ambiguities.

Revise Strategy (for Hard samples). Hard samples involve high uncertainty and structural divergence, indicating fundamental reasoning failures in $\mathcal{H}$ where reranking is insufficient. We trigger the Revise strategy, leveraging the Assistant LLM to reconstruct the clinical logic directly from the source text. Guided by a *Revise Prompt* that presents both the original medical text and the erroneous hypothesis, the Assistant generates a fully rectified decision tree, effectively resolving intractable generation failures. Detailed prompts are provided in Appendix B.

3 Experiments

3.1 Experimental Setup

Dataset and Metrics. We experiment on Text2DT dataset (Zhu et al., 2023). To the best of our knowledge, this is currently the only publicly available benchmark specifically annotated for Medical Decision Tree extraction. By strictly adhering to this community standard, we ensure a fair and reproducible comparison with prior SOTA approaches (He et al., 2024; Li et al., 2025). Detailed dataset statistics are in the Appendix D. Following

previous works (Zhu et al., 2023; He et al., 2024), we report: (1) **Triple Precision, Recall, and F1** for semantic extraction; (2) **Node & Path F1** for structural validity; and (3) **Tree Accuracy** and **Tree Levenshtein Ratio (LR)** (Bours, 1965). Given the high stakes of clinical decision-making, we prioritize **Tree Accuracy** as the primary metric, as it enforces a strict exact-match constraint aligning with CDSS safety standards.

Baselines. We compare our framework against a range of strong baselines: discriminative methods BERT-Biaffine (Wu, 2023) and PromptRE (Jiang et al., 2023); generative methods Text2MDT (Zhu et al., 2024) and FT-MDT (Li et al., 2025); and the previous SOTA, GM-MDT (He et al., 2024), which employs specialized linearization. We also include GPT-4o (5-shot) (Hurst et al., 2024; Brown et al., 2020) to benchmark strong few-shot capabilities. Baseline details are provided in Appendix E.

Implementation Details. To ensure fair comparison with GM-MDT, we use ChatGLM-6B (Zeng et al., 2024) as the backbone generator. For the *Rerank* and *Revise* strategies, we employ Qwen2.5-72B-Instruct, Qwen2.5-7B-Instruct (Yang et al., 2024), GLM-4-32B (Zeng et al., 2024), and DeepSeek-V3 (DeepSeek-AI et al., 2024) as assistant models. Full training details are provided in the Appendix F.

3.2 Main Results

As shown in Table 1, R³-MDT achieves a new SOTA in Tree Accuracy. Specifically, with a minimal candidate size of $K = 2$, our method attains **68.00%**, outperforming the Base Generator by **+5.0%** and GM-MDT by **+1.0%**. It is worth noting that while increasing K to 5 yields a marginal further improvement to **68.20%**, the computational cost rises disproportionately. Therefore, we adopt $K = 2$ as our primary implementation to prioritize inference efficiency while maintaining competitive accuracy.

This gain validates our difficulty-aware design: by retaining high-confidence predictions for Easy samples, we preserve correct answers and avoid noise; meanwhile, for Medium/Hard samples, our refinement mechanism effectively rectifies intractable reasoning failures that the backbone cannot resolve.

Furthermore, Table 1 highlights the robustness of R³-MDT across diverse LLM backbones, including Qwen2.5 (7B/72B), GLM-4 (32B), and

Method	K	Tri. Prec.	Tri. Rec.	Tri. F1	Node F1	Path F1	Tree LR	Tree Acc.
GPT-4o (5-shot)	—	79.01	81.55	80.26	74.36	55.72	83.90	44.00
BERT-Biaffine*	—	—	—	90.19	74.80	52.71	—	37.00
PromptRE*	—	—	—	94.39	85.31	69.27	—	55.00
Text2MDT*	—	89.80	90.40	90.10	—	53.20	89.80	49.00
FT-MDT*	—	91.10	91.60	91.30	—	67.90	93.60	55.00
GM-MDT*	—	—	—	95.43	90.48	77.91	—	67.00
Base Generator	—	92.01	89.68	90.83	87.11	71.96	89.22	63.00
R ³ -MDT (w/ Qwen2.5-7B)	2	93.07	90.86	91.95	88.55	73.99	89.71	65.00
	5	93.48	91.17	92.30	88.86	74.64	89.92	65.80
R ³ -MDT (w/ GLM-4-32B)	2	93.38	91.13	92.24	88.88	74.52	90.08	65.80
	5	93.72	91.37	92.53	89.20	75.22	90.22	67.20
R ³ -MDT (w/ DeepSeek-V3)	2	92.85	90.52	91.67	88.76	75.14	89.93	66.40
	5	93.34	91.17	92.24	<u>89.48</u>	<u>76.22</u>	90.25	68.20
R³-MDT (w/ Qwen2.5-72B)	2	<u>93.66</u>	<u>91.47</u>	<u>92.55</u>	89.12	75.87	90.66	<u>68.00</u>
	5	<u>93.66</u>	<u>91.47</u>	<u>92.55</u>	89.15	75.94	<u>90.67</u>	68.20

Table 1: Main results on the Text2DT dataset. Methods marked with * are cited from the original papers. The best results among comparable models are highlighted in **bold** and the second-best are underlined. For R³-MDT, we report the average performance across 5 runs.

DeepSeek-V3. Regardless of the base model’s size or architecture, our framework consistently yields performance gains over the baseline greedy decoding. This consistency indicates that our difficulty-aware routing mechanism is model-agnostic: it successfully identifies and rectifies intrinsic reasoning flaws inherent to auto-regressive generation, rather than merely overfitting to the biases of a specific model.

Crucially, even compared to the most powerful general-purpose LLM, GPT-4o (5-shot), our R³-MDT framework still achieves superior Tree Accuracy. This demonstrates that a strategically adaptive inference pipeline, rather than mere model scale, is paramount for robust MDT extraction.

Regarding component-level metrics (Triple, Node, and Path F1), we attribute GM-MDT’s marginal lead to its specialized Seq2Seq Copy Mechanism, which maximizes local lexical alignment with the reference text. However, high lexical overlap does not guarantee structural validity. GM-MDT relies on a resource-intensive multi-model ensemble to stabilize these local predictions, imposing a massive computational and memory burden. In contrast, R³-MDT prioritizes global structural coherence over rigid lexical matching. Our superior performance on Tree Accuracy—the strictest and most clinically relevant metric—demonstrates that our framework effectively rectifies intractable global failures that ensemble voting cannot resolve, achieving this with significantly lower resource consumption.

Despite the improvement, remaining failures pri-

marily arise from model over-confidence, where the generator assigns high probabilities to incorrect logical structures, thereby bypassing the adaptive routing mechanism. We provide a detailed analysis of these calibration-related errors in Appendix H.

3.3 Ablation Study

To dissect the contribution of each component in R³-MDT, we conduct ablation studies as detailed in Table 2.

Variant	Tri. F1	Node F1	Path F1	Tree Acc.
R³-MDT (Full)	92.55	89.08	76.09	68.00
Base Generator	90.83	87.11	71.96	63.00
+w/ Same Model	88.75	84.87	71.14	63.00
w/o Rerank	91.87	88.31	74.41	66.00
w/o Revise	91.51	87.88	73.64	65.00
All Rerank	92.04	86.22	66.44	56.00
All Revise	92.35	83.22	66.77	49.00

Table 2: Ablation study on Text2DT test set. Metrics include Triple F1, Node F1, Path F1, and Tree Accuracy.

Efficacy of Capability Gap. We first investigate whether the performance gain stems from our adaptive routing or relies on the superior reasoning of the assistant. Using the Base Generator (6B) as its own assistant (w/ Same Model) proves counter-productive: Tree Accuracy stalls at 63.00% while Node F1 and Triple F1 deteriorate. This indicates that smaller models lack the self-diagnostic capacity to identify their own errors. Without a sufficient capability gap (provided by the 72B assistant), the refinement process degenerates into noise introduction, confirming that advanced reasoning ca-

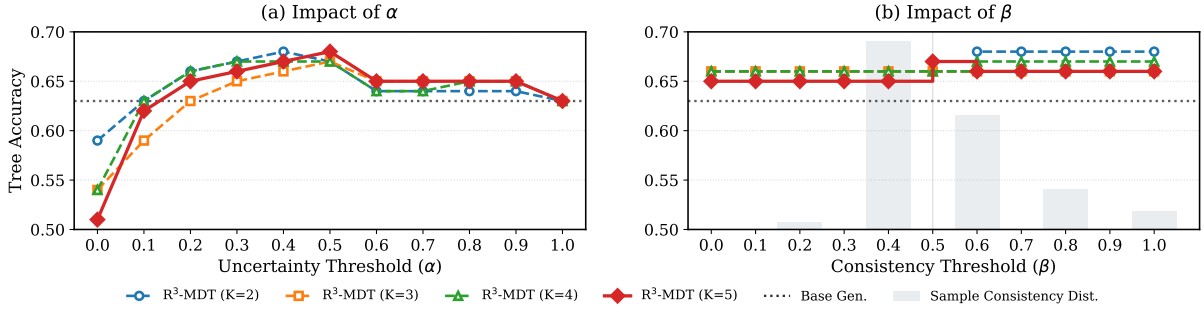


Figure 3: Parameter analysis on the test set. (a) **Impact of Uncertainty (α)**: The model exhibits a robust high-performance plateau across a broad range ($0.3 \leq \alpha \leq 0.5$), consistently outperforming the SOTA baseline. (b) **Impact of Consistency (β)**: The distinct step-wise pattern confirms the model’s reliability in identifying structural bottlenecks. (c) **Impact of Candidate Size (K)**: Performance remains remarkably stable across $K = 2$ to 5 , validating that a minimal candidate set is sufficient for effective routing while maximizing inference efficiency.

pabilities are a prerequisite for resolving complex samples within our framework.

Necessity of Refinement Modules. Removing either the Rerank or Revise modules leads to a performance decline, confirming their complementary roles in handling different failure modes. The **w/o Rerank** variant (66.00%) demonstrates that for Medium samples where the correct logic already exists within the candidate pool, an LLM-based critical selection is more robust than greedy decoding. Conversely, the drop in **w/o Revise** (65.00%) underscores that Hard samples often exhibit deep hallucinations absent from the initial hypothesis space, necessitating a full reconstruction to rectify intractable generation failures.

Impact of Adaptive Routing. The collapse in performance under **All Rerank** (56.00%) and **All Revise** (49.00%) highlights the danger of indiscriminate refinement. Both strategies lead to severe degradation compared to the adaptive baseline, indicating that without the Router as a safety valve, the model aggressively overwrites correct generations and introduces destructive noise. This confirms the Router’s essential role in preserving high-confidence predictions (Retain) while focusing resources on samples that strictly require intervention.

3.4 Parameter Analysis

We conduct a sensitivity analysis to investigate the impact of three critical hyper-parameters in our framework: the uncertainty threshold α , the consistency threshold β , and the number of candidates K . We adopt a control-variable approach: when varying one parameter, we fix the others to their optimal values determined on the development set. Figure 3 illustrates the performance trends.

Impact of Uncertainty Threshold (α). As shown in Fig. 3(a), Tree Accuracy exhibits a robust high-performance plateau across a broad range ($0.3 \leq \alpha \leq 0.5$) rather than a sensitive peak. Performance consistently matches or surpasses the previous SOTA, peaking at 68.00% ($\alpha \approx 0.4$). This insensitivity to hyperparameter variations ensures stable extraction quality across clinical settings without requiring precise tuning. In contrast, an overly aggressive threshold ($\alpha < 0.2$) triggers indiscriminate intervention leading to over-correction, while a conservative setting ($\alpha > 0.6$) fails to rectify complex samples, causing accuracy to converge to the baseline.

Impact of Consistency Threshold (β). Regarding the consistency threshold β , our framework also maintains high stability, showing a distinct step-wise pattern in Figure 3(b). The background histogram (representing sample consistency distribution at $K = 5$) reveals that structural consistency is largely concentrated in the moderate range rather than being purely polarized. The performance jump occurs precisely when β surpasses this region of ambiguity. This indicates that when candidates diverge significantly, treating the sample as Hard is more effective than treating it as Medium. Consequently, setting a strict threshold ($\beta \geq 0.6$) ensures that structural conflicts are resolved through rigorous rewriting, guaranteeing stable performance.

Impact of Candidate Size (K). Varying K from 2 to 5 yields minimal performance differences, as illustrated in Figure 3. This suggests that the top-2 hypotheses sufficiently capture the structural ambiguities needed for effective routing, rendering larger search spaces redundant. Since the computational cost of reranking scales linearly with K , we adopt $K = 2$ as the default. This setting strikes an

optimal balance, maintaining SOTA performance while minimizing inference latency.

3.5 Efficiency Analysis

Practical clinical deployment demands a rigorous assessment of computational efficiency, specifically regarding Memory Footprint and Inference Latency (see Table 3).

Bottlenecks of Static Ensembles. The previous SOTA, GM-MDT, relies on a static **ensemble** of multiple distinct checkpoints (e.g., three ChatGLM-6B and two Seq2Seq models). This architecture imposes a severe dual burden: (1) Memory Explosion: it requires massive VRAM ($>40\text{GB}$) to hold multiple active models, necessitating data-center grade hardware like A100s; and (2) Computational Redundancy: as model weights differ, input encoder states cannot be shared, forcing redundant re-computation of input representations even for simple samples.

Lightweight Adaptive Framework. In contrast, $\text{R}^3\text{-MDT}$ is designed for resource-constrained environments. By utilizing a **single** local base model (6B), our framework fits comfortably on consumer-grade GPUs (e.g., RTX 3090/4090, $\sim 13\text{GB}$ VRAM). The Assistant LLM is invoked *on-demand* only for complex samples, ensuring minimal latency and zero external API costs for the majority of Easy cases. Detailed traffic statistics are in Appendix G.

Metric	GM-MDT (SOTA)	$\text{R}^3\text{-MDT}$ (Ours)
Paradigm	Static Ensemble	Adaptive Routing
Active Models	$\sim 5 (3 \times 6\text{B} + 2 \times \text{Small})$	1 (6B) + API
Peak VRAM	High ($>40\text{GB}$) [†]	Low ($\sim 13\text{GB}$)
Latency Profile	Fixed High (All)	Adaptive (Fast for Easy)

Table 3: Efficiency comparison. [†]Estimated based on loading 3 distinct GLM-6B models.

3.6 Analysis on Structural Complexity

We stratify the test set by tree length to evaluate resilience. As shown in Table 4, the performance advantage of $\text{R}^3\text{-MDT}$ widens significantly as the decision paths become deeper.

Short Trees (Len 3): Both models achieve 83% accuracy, validating our *Retain* strategy. For these high-confidence inputs, the Router correctly identifies Easy samples and bypasses unnecessary interventions, thereby preserving original correct predictions without introducing the noise typical of indiscriminate revision.

Long Trees (Len 7): As depth increases, the Base Generator suffers from error accumulation, with accuracy dropping to 36%. In contrast, $\text{R}^3\text{-MDT}$ boosts accuracy to 45%, confirming that adaptive refinement effectively counteracts severe performance degradation in high-complexity scenarios where deep decision paths are prone to reasoning failures.

Tree Complexity	Base.	$\text{R}^3\text{-MDT}$.	Gain
Length 3 (Short)	83.33%	83.33%	+0.00%
Length 5 (Moderate)	63.64%	68.83%	+5.19%
Length 7 (Long)	36.36%	45.45%	+9.09%

Table 4: Performance comparison across different decision tree lengths. $\text{R}^3\text{-MDT}$ shows significant gains on complex structures.

3.7 Generalization Analysis

To examine whether $\text{R}^3\text{-MDT}$ extends beyond the original Chinese Text2DT setting, we conduct auxiliary evaluations under both **language shift** (Text2DT-En, translated from Text2DT (Zhu et al., 2023)) and **task shift** (FOLIO (Han et al., 2024), ASDiv (Miao et al., 2020), and Spider (Yu et al., 2018)). For each auxiliary benchmark, we keep the same RETAIN-RERANK-REVISE principle, hold out 10% of the training split as development data, and adapt prompts and thresholds to the task-specific difficulty, output space, and metric. Additional dataset details are provided in Appendix C.

Crucially, both $\mathcal{U}(X)$ and $\mathcal{C}(X)$ operate directly on generated token sequences rather than task-specific structures, making them readily applicable to general sequence generation tasks. Although tasks such as text-to-SQL and mathematical reasoning are not explicitly tree-structured in output form, their reasoning processes often involve implicit branching paths, which provides an intuitive basis for transferring difficulty-aware routing beyond MDT extraction.

As summarized in Table 5, our framework achieves consistent gains across all benchmarks: +6.00% on Text2DT-En (Tree Accuracy), +3.43% on FOLIO (Answer Accuracy), +1.33% on ASDiv (Answer Accuracy), and +2.00% on Spider (Execution Accuracy). These results suggest that uncertainty- and consistency-based routing captures transferable error patterns across different linguistic contexts, output spaces, and reasoning formats, supporting its potential as a general inference-time control mechanism for structured generation.

Task	Dataset	Base	Ours
MDT Extraction	Text2DT-En	25.00%	31.00%
Logic Reasoning	FOLIO	71.08%	74.51%
MWP Solving	ASDiv	87.33%	88.66%
Text-to-SQL	Spider	79.20%	81.20%

Table 5: Preliminary generalization results across one cross-lingual benchmark and three cross-task benchmarks. For Spider, we report only EX Acc.

3.8 Analysis and Case Study

We analyze a representative case of hypomagnesemia treatment to illustrate how R^3 -MDT resolves structural entanglement (Figure 4). The Base Generator exhibits attention inertia, incorrectly merging *oral* and *IV administration* routes into a single branch despite distinct clinical conditions. Our Router effectively detects this structural flaw via a spike in token-level uncertainty and a drop in consistency at the logic bottleneck. Consequently, the Revise strategy is triggered, which successfully decouples the entangled branches and reconstructs the correct decision path, ensuring clinical safety. A granular breakdown of the internal signals for this case (Appendix H.1), along with additional analyses of reranking completeness (Appendix H.2) and model over-confidence limitations (Appendix H.3), are provided in the Appendix H.

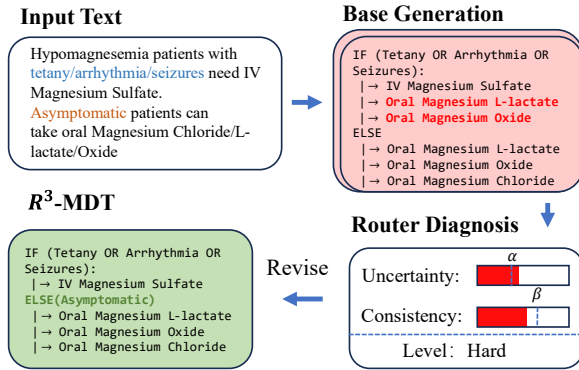


Figure 4: A case study of R^3 -MDT. The **strong semantic context** overshadows the **weak boundary signal**, causing the Base Generator to produce an entangled structure with high **uncertainty**. The Router detects this instability and triggers the **Revise** strategy, successfully decoupling the branches to yield the **correct logic**.

4 Related Work

Medical Decision Tree Extraction. MDT extraction transforms clinical texts into hierarchical structures (Zhu et al., 2023, 2024). Unlike sentence-level Information Extraction (IE) (Lample et al.,

2016; Wadden et al., 2019; Yan et al., 2021), it requires resolving long-range dependencies across entire notes, sharing complexities with document-level event and relation extraction (Yang et al., 2021; He et al., 2023; Yao et al., 2019). While recent paradigms have shifted toward generative sequence-to-set (Tan et al., 2021; He and Tang, 2022), parallel prediction (He and Tang, 2023), and end-to-end linearization (He et al., 2024; Li et al., 2025), these methods typically rely on greedy decoding without structural verification. Such training-centric approaches do not explicitly handle the stochastic nature of generation, leaving models prone to hallucinations. In contrast, our model-agnostic framework actively manages uncertainty to prevent errors during inference.

Reliability and Refinement in LLMs. Ensuring LLM reliability is critical for medical applications. While standard uncertainty estimation relies on perplexity (Malinin and Gales, 2021; Kadavath et al., 2022), our work adopts a minimum-probability criterion to pinpoint local errors at decision nodes. To enhance stability, Self-Consistency (SC) (Wang et al., 2022) marginalizes over reasoning paths but is computationally prohibitive for uniform application. Although inference-time refinement like Reranking (Ravaut et al., 2022) and Self-Correction (Madaan et al., 2023) has gained traction, these methods often suffer from *over-correction* when applied indiscriminately (Huang et al., 2023; Ma et al., 2023). Our R^3 framework bridges these gaps by utilizing uncertainty and consistency as triggers to selectively apply refinement, ensuring both precision and efficiency.

Adaptive Control and Failure Detection. Our routing design is also related to classical ideas in fault diagnosis and fault-tolerant control, where residual analysis compares observed system behavior with expected behavior to detect abnormal operating modes and trigger corrective actions only when needed (Frank, 1990; Blanke et al., 2006). We instantiate this diagnostic philosophy for LLM-based structured generation: token-level uncertainty provides a local residual-like signal for semantic or decision-node errors, while structural consistency captures global disagreement among candidate trees. The Router then maps the diagnosed state to Retain, Rerank, or Revise, turning refinement from a uniform post-processing step into a failure-mode-aware control policy.

5 Conclusion

We presented R³-MDT, a reliable inference-time framework for Medical Decision Tree extraction. By using token-level uncertainty and structural consistency as diagnostic signals, R³-MDT routes samples to Retain, Rerank, or Revise, preserving stable predictions while focusing refinement on cases that require intervention.

Experiments on Text2DT show that R³-MDT achieves state-of-the-art Tree Accuracy with improved efficiency. Auxiliary evaluations under language and task shifts further suggest that its uncertainty- and consistency-based routing can transfer beyond the original Chinese MDT setting, though these results are best viewed as preliminary evidence rather than exhaustive validation. Future work will further examine this diagnostic routing principle across broader clinical datasets and structured generation tasks.

Limitations

Despite the promising results, our framework has several limitations that warrant further investigation.

First, **Model Over-Confidence**: Our adaptive routing relies heavily on the assumption that the model’s token probabilities accurately reflect its correctness. However, LLMs often suffer from over-confidence (Guo et al., 2017), assigning high probabilities even to hallucinated contents (Ji et al., 2022). This calibration error may cause the Router to incorrectly categorize hard failure cases as Easy, bypassing necessary refinement. Integrating advanced confidence calibration techniques is a crucial direction for future work.

Second, **Dataset Scarcity and Evaluation Nuances**: While we have extended our empirical validation beyond the Chinese Text2DT benchmark through a self-translated English version and three public structured reasoning datasets, we acknowledge the inherent limitations of these auxiliary evaluations. The English MDT dataset serves only as a *silver-standard* proxy, as it was generated via machine translation and lacks expert-level human annotation. Consequently, the performance reported on English clinical guidelines may be influenced by translation artifacts or potential noise in label alignment. Due to the current absence of high-quality, gold-standard English benchmarks specifically for MDT extraction, our cross-lingual generalizability results should be interpreted as preliminary. Sim-

ilarly, the cross-task experiments probe transferability only across selected structured reasoning settings, so they do not establish universal applicability to all task formats. Future efforts will focus on collaborating with medical professionals to curate high-fidelity, natively-annotated datasets in multiple languages and evaluating the routing principle across a wider range of generation tasks.

Finally, **Dependency on Assistant Capability**: While our framework effectively routes hard samples to a more capable assistant model, the ultimate performance on these cases remains intrinsically bounded by the reasoning upper bound of the assistant itself. For extremely complex clinical logic that exceeds the capacity of even the 72B-parameter model, our current "Revise" mechanism cannot generate a correct solution from scratch. Future research could explore integrating external knowledge bases or tool-use (e.g., code interpreters) to break through this reasoning ceiling.

Ethical considerations

Data Privacy and Compliance. The main MDT experiments in this work were conducted using the publicly available Text2DT benchmark dataset. The Text2DT data originates from published medical textbooks and clinical guidelines, containing no personally identifiable information (PII) or private patient records. The auxiliary generalization experiments use public structured reasoning benchmarks or a machine-translated derivative of Text2DT. Therefore, our research involves no violation of patient privacy and adheres to standard data usage protocols.

Clinical Risk and Responsible Use. This work involves the automated extraction of clinical decision logic, a high-stakes application domain. We strongly emphasize that the Medical Decision Trees generated by our R³-MDT framework are intended solely as assistive references for qualified medical professionals. They should never be deployed for autonomous diagnosis or treatment decision-making without rigorous human verification. We disclaim liability for any medical errors resulting from the direct, unverified use of our model’s outputs.

Artifacts and Licensing. The Text2DT dataset and the auxiliary public benchmarks used in this work are publicly available for research use. The open-source models (ChatGLM-6B, Qwen2.5, and

DeepSeek-V3) are employed in accordance with their respective licenses (e.g., Apache 2.0, MIT). We accessed proprietary models (GPT-4o) via official APIs strictly following their terms of service.

Use of AI Assistants. In accordance with the ACL policy, we acknowledge the use of AI assistants in this work. Specifically, these tools were utilized to: (1) refine the linguistic clarity and correct grammatical errors in the manuscript; and (2) assist with partial code implementation, such as debugging and formatting data preprocessing scripts. We emphasize that all scientific claims, experimental designs, and logical reasoning were conceived and verified by the authors. The authors bear full responsibility for the accuracy and originality of the content.

References

- OpenAI Josh Achiam, Steven Adler, S. Agarwal, L. Ahmad, Ilge Akkaya, Florencia Leoni Aleman, D. Almeida, Janko Altenschmidt, S. Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder. 2006. *Diagnosis and Fault-Tolerant Control*. Springer, Berlin, Heidelberg.
- P.A.H. Bours. 1965. [Binary codes capable of correcting deletions, insertions, and reversals](#). *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 63.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, G. Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Neural Information Processing Systems*, 33:1877–1901.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv.org*.
- DeepSeek-AI, A. Liu, B. Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, C. Deng, Chenyu Zhang, and 1 others. 2024. Deepseek-v3 technical report. *arXiv.org*.
- Paul M. Frank. 1990. [Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results](#). *Automatica*, 26(3):459–474.
- Crina Grosan and Ajith Abraham. 2011. Rule-based expert systems. In *Intelligent systems: A modern approach*, pages 149–185. Springer.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhen-ting Qi, Martin Riddell, Luke Benson, Lucy Sun, E. Zubova, Yujie Qiao, Matthew Burtell, and 1 others. 2024. [FOLIO: Natural language reasoning with first-order logic](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031. Association for Computational Linguistics.
- Yuxin He, Jingyue Hu, and Buzhou Tang. 2023. [Revisiting event argument extraction: Can EAE models learn better when being aware of event co-occurrences?](#) In *Annual Meeting of the Association for Computational Linguistics*, pages 12542–12556. Association for Computational Linguistics.
- Yuxin He and Buzhou Tang. 2022. [Setgner: general named entity recognition as entity set generation](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 3074–3085. Association for Computational Linguistics.
- Yuxin He and Buzhou Tang. 2023. [Bispn: Generating entity set and relation set coherently in one pass](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 2066–2077. Association for Computational Linguistics.
- Yuxin He, Buzhou Tang, and Xiaoling Wang. 2024. [Generative models for automatic medical decision rule extraction from text](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 7034–7048. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. [Large language models cannot self-correct reasoning yet](#). In *International Conference on Learning Representations*.
- OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, D. Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, and 1 others. 2022. [Survey of hallucination in natural language generation](#). *ACM computing surveys*, 55(12):1–38.
- Yiwen Jiang, Hao Yu, and Xingyue Fu. 2023. [Medical decision tree extraction: A prompt based dual contrastive learning method](#). In *Communications in Computer and Information Science*, pages 103–116. Springer, Springer Nature Singapore.
- Saurav Kadavath, Tom Conerly, Amanda Askell, T. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, and 1 others. 2022. [Language models \(mostly\) know what they know](#). *arXiv.org*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *International Conference on Learning Representations*.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *North American Chapter of the Association for Computational Linguistics*, pages 260–270. Association for Computational Linguistics.
- Yuheng Li, Jiechao Gao, Wei Han, Wenwen Ouyang, Wei Zhu, and Hui Yi Leong. 2025. [FT-MDT: Extracting decision trees from medical texts via a novel low-rank adaptation method](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 65–76. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. [Large language model is not a good few-shot information extractor, but a good reranker for hard samples!](#) In *Conference on Empirical Methods in Natural Language Processing*, pages 10572–10601. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. [Self-refine: Iterative refinement with self-feedback](#). *Advances in Neural Information Processing Systems* 36, 36:46534–46594.
- Andrey Malinin and Mark Gales. 2021. [Uncertainty estimation in autoregressive structured prediction](#). In *International Conference on Learning Representations*.
- Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. [A diverse corpus for evaluating and developing English math word problem solvers](#). In *Annual Meeting of the Association for Computational Linguistics*, pages 975–984. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, N. Gimelshein, L. Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, volume 32, pages 8024–8035.
- Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. [Summareranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization](#). In *Annual Meeting of the Association for Computational Linguistics*, pages 4504–4524. Association for Computational Linguistics.
- E. Shortliffe and M. Sepúlveda. 2018. [Clinical decision support in the era of artificial intelligence](#). *JAMA*, 320(21):2199.
- Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. [A sequence-to-set network for nested named entity recognition](#). In *International Joint Conference on Artificial Intelligence*, pages 3936–3942. International Joint Conferences on Artificial Intelligence Organization.
- Roman Vashurin, Maiya Goloburda, Preslav Nakov, and Maxim Panov. 2025. [Uncertainty-line: Length-invariant estimation of uncertainty for large language models](#). *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7892–7919.
- Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2014. Grammar as a foreign language. *Neural Information Processing Systems*, 28.
- David Wadden, Ulme Wennberg, Yi Luan, and Hananeh Hajishirzi. 2019. [Entity, relation, and event extraction with contextualized span representations](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 5783–5788. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*, pages 287–301. Association for Computational Linguistics.
- 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](#). In *Neural Information Processing Systems*, volume 35, pages 24824–24837. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, C. Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. [Transformers: State-of-the-art natural language processing](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 38–45. Association for Computational Linguistics.

Zihong Wu. 2023. [Research on decision tree method of medical text based on information extraction](#). In *Communications in Computer and Information Science*, pages 127–133. Springer, Springer Nature Singapore.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. [A partition filter network for joint entity and relation extraction](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 185–197. Association for Computational Linguistics.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. [Document-level event extraction via parallel prediction networks](#). In *Annual Meeting of the Association for Computational Linguistics*, pages 6298–6308. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv.org*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. [Docred: A large-scale document-level relation extraction dataset](#). In *Annual Meeting of the Association for Computational Linguistics*, pages 764–777. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 2018. [Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task](#). In *Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921. Association for Computational Linguistics.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, and 1 others. 2024. [ChatGLM: A family of large language models from glm-130b to glm-4 all tools](#). *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, arXiv:2406.12793.

Wei Zhu, Wenfeng Li, Xing Tian, Pengfei Wang, Xiaoling Wang, Jin Chen, Yuanbin Wu, Yuan Ni, and Guotong Xie. 2024. [Text2MDT: extracting medical decision trees from medical texts](#). *arXiv.org*, pages 89–102.

Wei Zhu, Wenfeng Li, Xiaoling Wang, Wendi Ji, Yuanbin Wu, Jin Chen, Liang Chen, and Buzhou Tang.

2023. [Extracting decision trees from medical texts: an overview of the text2dt track in chip2022](#). In *China Health Information Processing Conference*, pages 89–102. Springer, Springer Nature Singapore.

A Uncertainty-Error Correlation Analysis

To justify the use of minimum token probability as a routing signal, we analyze its relationship with the Base Generator’s error rate. We partition the test set into five equal-frequency bins based on the minimum probability. As shown in Figure 5, the error rate increases monotonically as the minimum probability decreases, reaching 65.0% in the lowest probability bin. This strong correlation demonstrates that our uncertainty metric effectively captures local reasoning failures that require intervention.

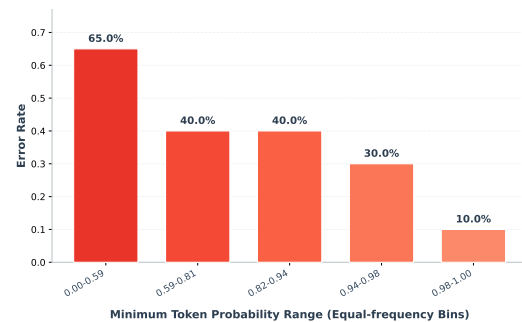


Figure 5: **Correlation between minimum token probability and reasoning error rate.** The test set is stratified into five equal-frequency bins based on the minimum probability. The clear monotonic decrease in error rate as the probability increases confirms that *min_prob* serves as a reliable internal signal to identify samples requiring intervention, effectively validating our difficulty-aware routing mechanism.

B Prompt Designs

Since the Text2DT dataset consists of Chinese medical guidelines, all prompts utilized in our experiments are in Chinese. To facilitate reproducibility and understanding for non-Chinese speakers, we provide the original Chinese prompts alongside their English translations in this section. We utilized fixed prompt templates where dynamic inputs are denoted by square brackets (e.g., [Medical Text]).

B.1 GPT-4o Few-Shot Prompt Template

Instruction (Chinese):

请形式化地描述以下医疗文本所表达的诊疗决策过程，目标输出格式为“若...[，若...]，则...，否则，若...[，若...]，则...，否则...”。其中，省略部分由单个或多个关系三元组构成，关系类型包括“临床表现”，“用药”，“治疗方案”，“用法”，“基本情况”，“慎用”。

下面是一些例子：文本：[Example Input 1] 诊疗决策过程：[Example Output 1] ... (4 more examples omitted for brevity) ...

现在请形式化地描述以下医疗文本所表达的诊疗决策过程：文本：[Target Medical Text] 诊疗决策过程：

Instruction (English Translation):

Please formally describe the diagnosis and treatment decision process expressed in the following medical text. The target output format is "If... [，if...]，then....，else，if... [，if...]，then....，else...". The omitted parts consist of single or multiple relation triplets. Relation types include "Clinical Manifestation", "Medication", "Treatment Plan", "Usage", "Basic Condition", and "Contraindication".

Here are some examples: Text: [Example Input 1] Decision Process: [Example Output 1] ... (4 more examples omitted for brevity) ...

Now please formally describe the diagnosis and treatment decision process expressed in the following medical text: Text: [Target Medical Text] Decision Process:

Note: In the actual prompt, we provide **5 fixed examples** (5-shot) covering various decision logic patterns (e.g., recursive logic, parallel logic). A representative example (Example 1) involves a patient with Henoch-Schönlein purpura, where the decision tree branches based on symptoms like "pulmonary-renal syndrome" or "rapid progression" to determine treatments like "plasma exchange" or "immunosuppressants".

Table 6: The prompt template used for the **GPT-4o (5-shot)** baseline. Due to space constraints, we show the instruction and the structure of the few-shot examples.

B.2 Prompt for Cross-Lingual Dataset Translation

Role Setting:

You are a professional medical translator.

Task:

Translate the 'input' and 'target' fields from Chinese to English.

Rules:

1. **Relation Mapping:** You MUST use these exact terms for relations inside parentheses: [{re_instructions}].
 2. **Logic Connectors:** Use 'then', 'else', 'and', 'or' to represent the tree logic.
 3. **Format:** The 'target' field structure is "If (S, R, O)... then (S, R, O)...".
 4. **Output Format:** ONLY return a valid JSON object with 'input' and 'target' keys.
-

Table 7: The prompt template used for translating the Text2DT dataset from Chinese to English to construct the cross-lingual generalization benchmark.

B.3 Rerank Prompt Template

Instruction (Chinese):

你是一个医学专家, 请根据医疗文本, 在候选文本集合中选出正确的决策文本树, 先输出序号, 再进行解释。如果不确定, 尽量输出序号0。决策文本树格式为"若...[, 若...], 则..., 否则, 若...[, 若...], 则..., 否则..."。其中, 省略部分由单个或多个关系三元组构成, 关系类型包括"临床表现", "用药", "治疗方案", "用法", "基本情况", "慎用"。

注意事项如下: (1) 决策文本树以是否最符合医疗文本为标准。(2) 三元组尾实体与原始医疗文本完全一致, 可以截取, 严禁改写。(3) 实体清洗规则: 实体应仅保留核心医学概念, 必须去除冗余的动词、副词或连接词前后缀。(4) 在语义准确的前提下, 优先选择实体更简短、更干净的候选文本。(5) 请在思考过程中保持简练, 不要过度展开, 直接聚焦于关键逻辑。

医疗文本: [Medical Text] 候选文本集合: [Candidate Set]

Instruction (English Translation):

You are a medical expert. Based on the medical text, please select the correct decision tree text from the set of candidate texts. Output the index first, followed by an explanation. If unsure, try to output index 0. The format of the decision tree text is "If... [, if...], then..., else, if... [, if...], then..., else...". The omitted parts consist of single or multiple relation triplets. Relation types include "Clinical Manifestation", "Medication", "Treatment Plan", "Usage", "Basic Condition", and "Contraindication".

Precautions are as follows: (1) The standard for the decision tree text is whether it best aligns with the medical text. (2) The tail entity of the triplet must be exactly consistent with the original medical text; truncation is allowed, but rewriting is strictly forbidden. (3) Entity cleaning rules: Entities should retain only core medical concepts and must remove redundant verbs, adverbs, or prefixes/suffixes of conjunctions. (4) Under the premise of semantic accuracy, prioritize candidate texts with shorter and cleaner entities. (5) Please keep your thinking process concise, do not expand excessively, and focus directly on the key logic.

Medical Text: [Medical Text] Candidate Text Set: [Candidate Set]

Table 8: The prompt template used for the **Rerank** strategy.

B.4 Revise Prompt Template

Instruction (Chinese):

角色设定: 你是一位严谨的医学专家, 擅长将临床指南转化为结构化的决策树。请基于提供的文本, 修正并生成完全符合所给医学文本的决策树。树状决策流程包含条件节点和决策节点两种, 基本格式为若[条件节点], 则[决策节点], 否则[决策节点]。决策节点为子树, 可以为空, 也可以包含条件节点继续延伸, 其中节点由类型为(头实体, 关系类型, 尾实体)的三元组以与/或逻辑连接构成, 逻辑结构按顺序展示即可, 不必为了防止歧义而进行嵌套。

任务要求: 输出格式为"若...[, 若...], 则..., 否则, 若...[, 若...], 则..., 否则..." 输入: 原始医学文本和包含错误的医学决策树文本(自然语言形式)。输出: 修正后的、符合医学指南、自然语言格式的决策树, 需满足: 1. 决策树结构完整, 原始医学文本提到的条件节点, 必须全部提取, 严禁遗漏。2. 决策节点中的三元组关系类型包括"临床表现", "用药", "治疗方案", "用法", "基本情况", "慎用"。3. 决策节点的三元组关系内部不包含逻辑关系, 若有, 则进行拆分。4. 决策节点中的实体必须与原始医学文本完全一致, 严禁改写。5. 实体应仅保留核心医学概念, 必须去除冗余的动词、副词或连接词前后缀。6. 在语义准确的前提下, 优先选择实体更简短、更干净的候选文本。7. 请在思考过程中保持简练, 不要过度展开, 直接聚焦于关键逻辑。

原始医疗文本: [Medical Text] 错误决策树文本集合: [Incorrect Tree Text]

Instruction (English Translation):

Role Setting: You are a rigorous medical expert skilled in converting clinical guidelines into structured decision trees. Please revise and generate a decision tree that fully aligns with the given medical text based on the provided text. The tree decision process includes condition nodes and decision nodes. The basic format is "If [Condition Node], then [Decision Node], else [Decision Node]". Decision nodes are subtrees, which can be empty or contain condition nodes extending further. Nodes consist of triplets in the form of (Head Entity, Relation Type, Tail Entity) connected by AND/OR logic. The logical structure should be displayed sequentially without nesting to prevent ambiguity.

Task Requirements: Output format: "If... [, if...], then..., else, if... [, if...], then..., else..." Input: Original medical text and the medical decision tree text containing errors (in natural language form). Output: A revised decision tree in natural language format that complies with medical guidelines, satisfying the following: 1. The decision tree structure must be complete; all condition nodes mentioned in the original medical text must be extracted, with no omissions strictly enforced. 2. Relation types in decision nodes include "Clinical Manifestation", "Medication", "Treatment Plan", "Usage", "Basic Condition", and "Contraindication". 3. Triplets within decision nodes should not contain logical relations; if present, split them. 4. Entities in decision nodes must be exactly consistent with the original medical text; rewriting is strictly forbidden. 5. Entities should retain only core medical concepts and must remove redundant verbs, adverbs, or prefixes/suffixes of conjunctions. 6. Under the premise of semantic accuracy, prioritize candidate texts with shorter and cleaner entities. 7. Please keep your thinking process concise, do not expand excessively, and focus directly on the key logic.

Original Medical Text: [Medical Text] Incorrect Decision Tree Text Set: [Incorrect Tree Text]

Table 9: The prompt template used for the **Revise** strategy.

C Generalization Benchmark Details

Dataset	Task	Train	Eval.	Metric
Text2DT-En	Cross-lingual MDT extraction	300	100	Tree Acc.
ASDiv	Math word-problem solving	1,776	529	Answer Acc.
Spider	Text-to-SQL parsing	8,659	2,147	Execution Acc.
FOLIO	Natural-language logical reasoning	1,004	204	Answer Acc.

Table 10: Auxiliary benchmark details for the generalization analysis. For each auxiliary benchmark, 10% of the training split is held out as a development set for prompt adaptation and routing-threshold tuning.

The auxiliary benchmarks cover one language-shift setting and three task-shift settings. For Text2DT-En, we translate the original Text2DT inputs and targets into English while preserving the MDT output structure. For the three cross-task benchmarks, we adapt R³-MDT to representative structured reasoning settings: mathematical word-problem solving, text-to-SQL parsing, and natural-language logical reasoning. In all auxiliary experiments, we reserve 10% of the training split as a development set for prompt adaptation and routing-threshold tuning; the reported results are computed only on the held-out evaluation split.

ASDiv contains grade-school arithmetic word problems, where the model must recover the underlying calculation and final answer. Spider evaluates executable SQL generation from natural-language questions paired with database schemas, testing whether the routing principle transfers to structured program generation. FOLIO evaluates logical reasoning over natural-language premises, requiring the model to map premises into a logical reasoning format and determine whether the conclusion is true, false, or unknown.

D Dataset Statistics

Detailed statistics of the Text2DT dataset are listed in Table 11.

E Baseline Descriptions

We compare our framework against three categories of strong baselines:

Discriminative Methods. **BERT-Biaffine** treats MDT extraction as a dependency parsing task, utilizing a biaffine classifier to predict edges between text spans. **PromptRE** leverages prompt learning to formulate the task as a cloze-style relation extraction problem.

Generative Methods. **Text2MDT** (Zhu et al., 2023) establishes the benchmark using Seq2Seq models (e.g., BART) with a pipeline approach. **FT-MDT** (Li et al., 2025) adapts LLMs via

Item	Count
Texts	500
Train/Dev/Test Splits	300/100/100
Avg. Text Length	66.5
Relation Classes	6
Relations per Text	6.39
Relation Name	Count (Proportion)
Clinical Manifestation	1374 (42.51%)
Medication	910 (28.15%)
Treatment Plan	561 (17.36%)
Dosage and Usage	222 (6.87%)
Contraindication	83 (2.57%)
Basic Condition	82 (2.54%)
Tree Structure (Pre-order)	Count (Proportion)
CDD	134 (26.80%)
CDCDD	253 (50.60%)
CCDDD	47 (9.40%)
CDCDCDD	45 (9.00%)
CCDCDDD	17 (3.40%)
CCDDCDD	2 (0.40%)
CDCDCDCDD	2 (0.40%)

Table 11: Statistics of the Text2DT Dataset (“C”/“D” represents a “condition”/“decision” node)

Path-Integrated LoRA (PI-LoRA) for parameter-efficient fine-tuning. **GM-MDT** (He et al., 2024) represents the previous SOTA. It employs a specialized linearization strategy and integrates a **Seq2Seq Copy Mechanism** to enhance entity extraction, utilizing a multi-model ensemble strategy for structural selection.

General-Purpose LLM. **GPT-4o (5-shot)** serves as a strong zero-training baseline. We employ In-Context Learning (ICL) with 5 representative examples to benchmark the capability of state-of-the-art closed-source models.

F Implementation Details

We implemented our framework using PyTorch¹ (Paszke et al., 2019) and the HuggingFace Transformers library² (Wolf et al., 2019). Local model

¹<https://pytorch.org/>

²<https://huggingface.co/docs/transformers/>

experiments were conducted on a single server equipped with 1 A100-80GB GPU.

F.1 Model Specifications

We utilize ChatGLM-6B³ as the backbone generator for the initial extraction. For the Rerank and Revise stages, we employ Qwen2.5-72B-Instruct⁴ as the assistant model due to its superior instruction-following capabilities.

F.2 Training Setup (Base Generator)

The backbone model was fine-tuned on the Text2DT training set using the LoRA technique (Hu et al., 2021). We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a cosine learning rate schedule. The detailed hyperparameters are listed in Table 12.

Hyperparameter	Value
<i>Training Dynamics</i>	
Batch Size	8
Learning Rate	2e-4
Epochs	10
Max Sequence Length	512
Warmup Ratio	0.1
Weight Decay	1e-5
Optimizer	AdamW
Scheduler	Linear
<i>PEFT (LoRA) Settings</i>	
LoRA Rank (r)	8
LoRA Alpha	32
LoRA Dropout	0.1
Target Modules	Attn. & FFN Layers

Table 12: Hyperparameters used for fine-tuning the ChatGLM-6B backbone generator.

F.3 Inference Configuration

During the inference phase, we applied the following settings to ensure reproducibility:

Candidate Generation. For the base generator, we utilized Beam Search to generate the candidate set. **Beam Size (K):** Set to 2 (as discussed in Section 3.4).

R³-MDT Framework. The adaptive routing thresholds were determined based on performance on the validation set:

- **Uncertainty Threshold (α):** 0.4

³<https://huggingface.co/THUDM/chatglm-6b>

⁴<https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>

- **Consistency Threshold (β):** 0.6

For the assistant model (Qwen2.5-72B-Instruct) in Rerank and Revise stages, we set the temperature to 0.7 to ensure deterministic and rigorous outputs. The maximum generation length was restricted to 2048 tokens.

F.4 Baselines Implementation

For GPT-4o, we accessed the model via the OpenAI API (version gpt-4o-2024-05-13) with the temperature set to 0.01.

G Additional Efficiency Statistics

To complement the deployment analysis in Section 3.5, we provide a breakdown of the generated token traffic in Table 13.

Method	Avg.Tokens	Traffic Ratio	Tree Acc.
Base Generator	205.78	1.00×	63.00
GM-MDT	>1000*	>4.86×	67.00
R³-MDT (Ours)	418.46	2.03×	68.00

Table 13: Average generated token volume per sample.

*Estimated based on the output sum of the 5-model ensemble used in GM-MDT.

Note on Computational Cost. It is important to acknowledge that the per-token computational cost of the 72B Assistant LLM (used in R³-MDT) is higher than that of the 6B models used in GM-MDT. However, GM-MDT’s ensemble approach requires generating complete output sequences from 5 distinct models for *every* sample. In contrast, R³-MDT significantly reduces the total volume of generated text (Traffic Ratio 2.03× vs. >4.86×) by restricting the heavy refinement process to hard samples only. This reduction in generation traffic helps offset the unit-cost difference of the larger model, while the removal of the multi-model ensemble requirement significantly lowers the barrier to deployment.

H Qualitative Analysis and Case Studies

This appendix provides a qualitative analysis of R³-MDT’s behavior. We categorize the results into successful rectifications of complex logic and systematic failures arising from model over-confidence, providing insights into the framework’s boundaries.

H.1 Success Case: Adaptive Revision for Structural Complexity

This subsection provides the detailed analysis of the Hypomagnesemia case presented in Section

3.7 (Figure 4). It demonstrates how the **Revise** strategy corrects deep structural hallucinations that are absent from the initial hypothesis space.

- **The Structural Entanglement:** The guideline distinguishes between IV magnesium for symptomatic patients and oral magnesium for asymptomatic ones. The Base Generator suffers from "attention inertia," erroneously nesting the oral treatment nodes inside the IV branch, creating a dangerous clinical logic.
- **Detection via Dual Signals:**
 - **Uncertainty ($\mathcal{U}$):** A significant spike ($\mathcal{U} = 0.78 > \alpha$) is detected at the logic bottleneck.
 - **Consistency ($\mathcal{C}$):** The candidate trees diverge significantly in their hierarchical arrangement ($\mathcal{C} = 0.25 < \beta$).

The Router identifies this as a **Hard** sample, successfully triggering the **Revise** module.

- **Revision Outcome:** As shown in Table 14, the Assistant LLM re-evaluates the global context to decouple the administration routes, resulting in a clinically safe parallel structure (ELSE branch).

Input Text (Translated): Hypomagnesemia patients with tetany/arrhythmia/seizures need IV Magnesium Sulfate. Asymptomatic patients can take oral Magnesium Chloride/Lactate/Oxide.
Base Generation (Entangled Structure): IF (Tetany OR Arrhythmia OR Seizures): -> IV Magnesium Sulfate -> AND Oral Magnesium Oxide (<i>Structural Hallucination</i>) -> AND Oral Magnesium Chloride
Router Diagnosis: Uncertainty: High → Ambiguous Consistency: Low → Divergent Decision: Hard → Revise
R³-MDT Output (Revised & Decoupled): IF (Tetany OR Arrhythmia OR Seizures): -> IV Magnesium Sulfate ELSE (Asymptomatic): (✓ Correct Logic) -> OR Oral Magnesium Sulfate -> OR Oral Magnesium Oxide -> OR Oral Magnesium Chloride

Table 14: Granular analysis of a successful Revise strategy in resolving structural entanglement.

H.2 Success Case: Reranking for Information Completeness

Standard greedy decoding or beam search often favors shorter sequences, leading to the omission of

intermediate treatment steps. The Rerank strategy proves effective in mitigating this issue by evaluating candidate completeness against the source text.

Table 15 presents a case involving missing triple.

- **The Ambiguity:** The guideline specifies a two-step process for patients with reflux: first "control acute episodes," then consider "preventive antibiotics."
- **Candidate 1 (Top-1 Generation):** It captures the correct logic branches but omits the precondition node ("Control acute episodes"), jumping directly to the medication. This is a common "shortcut" error in Seq2Seq models.
- **Candidate 2 (Selected by Reranker):** It fully preserves both the treatment step and the medication.
- **Outcome:** Although Candidate 1 might have a higher generation probability (due to shorter length), the Reranker correctly identifies that Candidate 2 aligns better with the source constraint "after controlling acute episodes," thus promoting it as the final answer.

Input Text (Translated): For patients with Urinary Tract Infection (UTI): Preventive antibiotics are not recommended for routine use in first-time infections; but for patients with dilated vesicoureteral reflux, preventive antibiotic therapy is recommended after controlling acute episodes .
Candidate 1 (Original Top-1, Rejected): IF (UTI, Clinical, First infection): -> (UTI, Contraindication, Preventive antibiotics) ELSE IF (UTI, Clinical, Dilated VUR): -> (UTI, Medication, Preventive antibiotics) (Missed "Control acute episodes" step)
Candidate 2 (Original Top-2, Selected): IF (UTI, Clinical, First infection): -> (UTI, Contraindication, Preventive antibiotics) ELSE IF (UTI, Clinical, Dilated VUR): -> (UTI, Treatment, Control acute episodes) (✓) -> AND (UTI, Medication, Preventive antibiotics)

Table 15: Analysis of Rerank strategy in improving information completeness.

H.3 Failure Case: The Trap of Model Over-Confidence

A critical limitation of our framework is its reliance on token probability as a proxy for correctness. While generally effective, this assumption fails when the base generator exhibits **Over-Confidence**

assigning high probabilities to erroneous predictions, particularly in semantic classification. In such cases, the Router misclassifies the sample as Easy and triggers the **Retain** strategy, bypassing necessary corrections.

Table 16 illustrates a representative example involving a Triple Error.

- **The Semantic Nuance:** The guideline for *Sweet Syndrome* mentions "Immunosuppressant therapy" alongside specific drugs like "Potassium iodide". In the Text2DT schema, generic therapies should be labeled as *Treatment Plan*, while specific drugs are *Medication*.
- **Model Error:** The Base Generator correctly extracts the structure but misclassifies "Immunosuppressant" as **Medication**.
- **Why it failed?** The model is swayed by the context of the subsequent drugs (Potassium iodide, Colchicine), creating a "contextual inertia." It confidently generates the token for "Medication" with high probability. Consequently, the calculated Uncertainty ($\mathcal{U}$) is low, and the Consistency ($\mathcal{C}$) is perfect (since the structure is stable). The Router mistakenly tags this as an Easy sample and Retains the error.

While this case manifests as a semantic span error (misclassifying a generic therapy as a specific medication), it represents a systemic vulnerability of uncertainty-based routing. The danger of model over-confidence is universal: whether the error is a local semantic label or a global structural branch, the high-confidence assignment effectively 'blinds' the Router. This creates a silent failure where the framework retains an incorrect output, bypassing both Rerank and Revise strategies. This highlights that improving the calibration of LLMs is as crucial as improving their reasoning depth for reliable MDT extraction.

Input Text (Translated): For patients with Sweet Syndrome: those with mild symptoms require no special treatment; those with severe symptoms can undergo immunosuppressant therapy; potassium iodide and colchicine are also common treatments.
R³-MDT Output (Retained Error): IF (Sweet Syn., Clinical, Mild): -> (Sweet Syn., Treatment, No special treatment) ELSE IF (Sweet Syn., Clinical, Severe): -> (Sweet Syn., Medication , Immunosuppressant) <i>(Wrong Relation)</i> -> OR (Sweet Syn., Medication, Potassium iodide) -> OR (Sweet Syn., Medication, Colchicine)
Router Diagnosis: Uncertainty: Low → Confident Consistency: High → Stable Decision: Easy → Retain (Mistake!)
Ground Truth Logic: ... ELSE IF (Sweet Syn., Clinical, Severe): -> (Sweet Syn., Treatment Plan , Immunosuppressant therapy) -> OR (Sweet Syn., Medication, Potassium iodide) ...

Table 16: Failure analysis: A case of over-confidence where incorrect semantic classification bypasses the routing mechanism.