From Answers to Arguments: Toward Trustworthy Clinical Diagnostic Reasoning with Toulmin-Guided Curriculum Goal-Conditioned Learning

Chen Zhan[1*], **Xiaoyu Tan**[2*], **Gengchen Ma**[1], **Yu-Jie Xiong**[1], **Xiaoyan Jiang**[1], **Xihe Qiu**[1†]
[1]School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China.
[2] Tencent Youtu Lab, Shanghai, 200232, China.
chenzhan361@gmail.com, arthurtan@tencent.com, qiuxihe1993@gmail.com
[*] Equal contribution [†] Corresponding author

Abstract

The integration of Large Language Models (LLMs) into clinical decision support is critically obstructed by their opaque and often unreliable reasoning. In the high-stakes domain of healthcare, correct answers alone are insufficient; **clinical practice demands full transparency to ensure patient safety and enable professional accountability.** A pervasive and dangerous weakness of current LLMs is their tendency to produce **"correct answers through flawed reasoning."** This issue is far more than a minor academic flaw; such process errors signal **a fundamental lack of robust understanding, making the model prone to broader hallucinations and unpredictable failures when faced with real-world clinical complexity.** In this paper, we establish a framework for trustworthy clinical argumentation by adapting the Toulmin model to the diagnostic process. We propose a novel training pipeline: **Curriculum Goal-Conditioned Learning (CGCL)**, designed to progressively train LLM to generate diagnostic arguments that explicitly follow this Toulmin structure. CGCL's progressive three-stage curriculum systematically builds a solid clinical argument: (1) extracting facts and generating differential diagnoses; (2) justifying a core hypothesis while rebutting alternatives; and (3) synthesizing the analysis into a final, qualified conclusion. We validate CGCL using **T-Eval**, a quantitative framework measuring the integrity of the diagnosis reasoning. Experiments show that our method achieves diagnostic accuracy and reasoning quality comparable to resource-intensive Reinforcement Learning (RL) methods, while offering a more stable and efficient training pipeline. [1]

1 Introduction

Large Language Models (LLMs) are rapidly reshaping both science and society, and healthcare shows this impact most vividly (Yu et al., 2025; Tan et al., 2024). From automating clinical note summarization and assisting in image interpretation to accelerating drug discovery (Obuchowicz et al., 2025; Liu et al., 2024), LLMs are now entering critical stages of the medical pipeline. Recent systems such as GPT-5 and Gemini-2.5-Pro have posted record-breaking results on medical benchmarks like MedQA and the USMLE, in some cases even surpassing human experts (Nori et al., 2023; Kung et al., 2023; Tu et al., 2025). But standardized exams are not the same as real clinical practice. Clinical decision-making is not about picking the right option from a fixed set. Instead, it is about reasoning under uncertainty, integrating messy and incomplete information, and making judgments where the cost of error is human life (Hager et al., 2024; McCoy et al., 2025). This leads us to the core scientific question: **Do today's LLMs truly have the capacity to function as reliable clinical decision-makers, or are they only test-takers with no license to practice?**

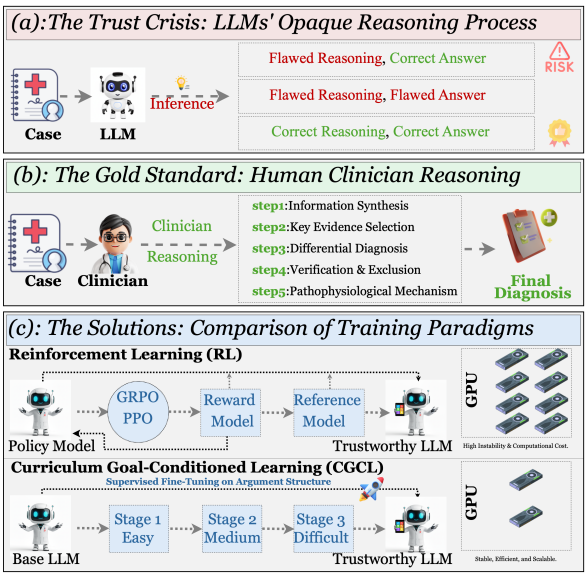


Figure 1: Clinical Diagnostic Reasoning Paradigms.

[1]https://github.com/Leonard-zc/CGCL.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13985–14004
July 2-7, 2026

At the heart of this gap is a dangerous illusion: **models that produce correct answers for the wrong reasons.** Most current evaluation paradigms fixate on whether the final answer is correct, while overlooking the reasoning path that leads there (Turpin et al., 2023; Lanham et al., 2023). In medicine, however, the path matters as much as the destination. Take a patient with fever and elevated white blood cell count. An LLM might flag a bacterial infection and suggest antibiotics seemingly correct, but only by matching patterns in the text. What it may miss are subtle but critical indicators of sepsis, such as rising lactate levels or borderline hypotension, which demand urgent intervention. In this case, the model's "right" answer is anchored in shallow reasoning that would fail under real clinical pressure (Jin et al., 2020). **Such brittle correctness is not a minor academic flaw; it is a systemic hazard. By rewarding outcomes without probing reasoning, current benchmarks inflate our perception of what LLMs can actually do.** This misalignment fuels a growing trust crisis and stands as the central obstacle between impressive lab performance and safe, reliable use at the bedside. Figure 1 summarizes this shift from answer-centric evaluation to the structured clinical reasoning paradigm studied in this paper.

This trust crisis is rooted not just in model behavior, but in the very way we evaluate and train them. On the evaluation side, current practices mostly reduce performance to answer accuracy or rely on subjective human ratings (Asgari et al., 2025). Neither provides a structured, quantitative way to assess the quality of reasoning—the logical steps, evidentiary grounding, and robustness of an argument remain largely invisible. On the training side, mainstream methods such as Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) inherit the same bias toward final answers (Kaufmann et al., 2024). The result is a cycle of over-optimizing for correctness while under-optimizing for reasoning. In theory, RL could help address this by shaping richer objectives, but in practice, the barriers are steep: designing reliable reward models is notoriously difficult, training can be unstable, and the computational demands can be substantial in practice. Breaking this bottleneck calls for new paradigms that are both stable and resource-efficient, while directly targeting reasoning as a first-class objective rather than a byproduct (Luo et al., 2024; Lightman et al., 2023).

A natural foundation for clinical reasoning is the **Toulmin model** of argumentation, which emphasizes how claims must be supported by evidence, qualified by uncertainty, and defended against counter-arguments (Toulmin, 2003). This structure closely mirrors the reasoning clinicians use when moving from symptoms and lab results to a defensible diagnosis (Ju et al., 2017; Caroprese et al., 2022). Yet current LLM pipelines neither evaluate nor learn in this structured way, leaving their "reasoning" opaque and often superficial.

To close this gap, we place the Toulmin model at the core of both evaluation and training. On the evaluation side, we introduce **T-Eval**, a scalable framework that quantitatively assesses an LLM's diagnostic reasoning by explicitly measuring the strength of claims, evidence, and rebuttals (Yun et al., 2025). On the training side, we design **Curriculum Goal-Conditioned Learning (CGCL)**, a pipeline that teaches a model to reason in Toulmin's structured form, following the progression of clinical training (Bengio et al., 2009; Schaul et al., 2015; Andrychowicz et al., 2017). In Stage 1 (Fact Gathering), the model plays the role of a junior resident, extracting findings and proposing an initial differential. In Stage 2 (Hypothesis Testing), it advances to a senior resident, justifying its main hypothesis with physiological evidence while refuting alternatives. In Stage 3 (Synthesis & Conclusion), it acts as an attending physician, integrating all evidence into a well-qualified final diagnosis. This curriculum can effectively operationalize Toulmin-style reasoning through imitation learning (e.g., SFT) without relying on RL. In experiments on complex real-world medical cases, CGCL-trained models not only achieve diagnostic accuracy competitive with strong baselines, but also substantially improve structured reasoning quality under T-Eval, with the clearest gains on smaller models.

We make the following key contributions:

(1) T-Eval: A Toulmin-based framework for evaluating clinical reasoning. We introduce the first scalable method to move beyond answer accuracy and directly measure the structure and integrity of diagnostic arguments, capturing how claims are supported, qualified, and defended.

(2) CGCL: A training paradigm that instills Toulmin-style reasoning. We design a three-stage, goal-conditioned curriculum that mirrors medical training, systematically teaching LLMs to gather facts, test hypotheses, and synthesize conclusions in a transparent and stable manner.

(3) Extensive validation on real-world clinical

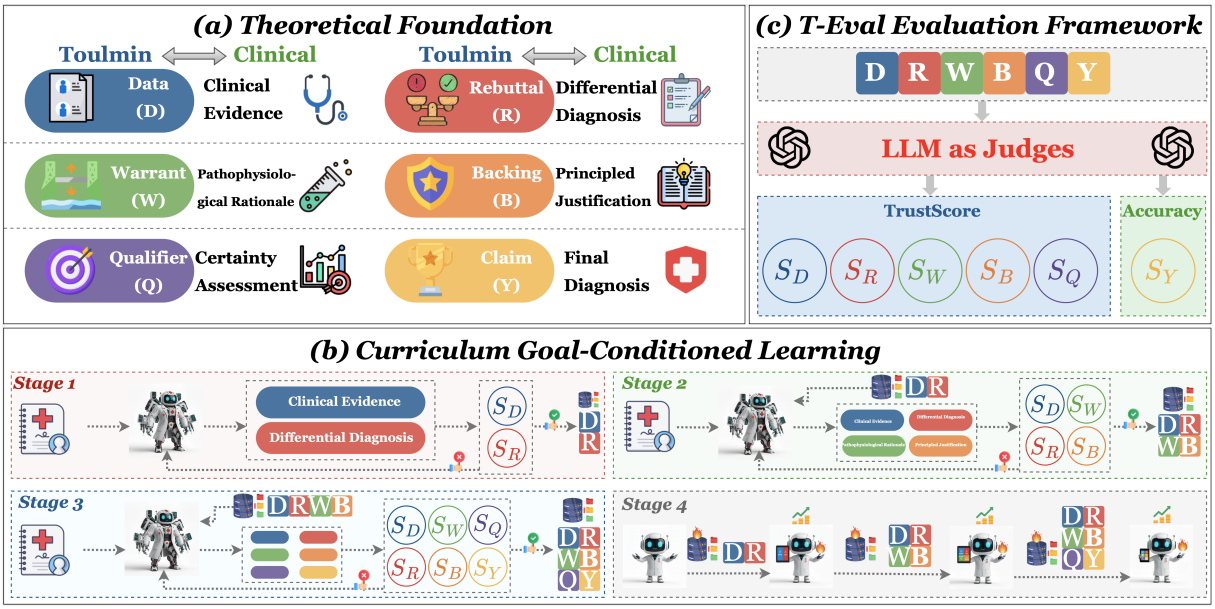


Figure 2: The Synergistic Architecture of CGCL and T-Eval for Trustworthy Clinical Reasoning.

cases. Through comprehensive experiments, we demonstrate that a well-structured curriculum can match the reasoning capabilities of complex RL-based approaches while reducing computational overhead.

2 Methodology

Our framework bridges the gap between answer accuracy and trustworthy reasoning in clinical LLMs. We teach structured reasoning via CGCL, which builds Toulmin-style diagnostic arguments, and assess it with T-Eval, which scores structural integrity. As illustrated in Figure 2, decomposed sub-tasks yield transparent, goal-guided reasoning.

2.1 Problem Formulation and Toulmin Model Instantiation

2.1.1 Toulmin-structured diagnostic argument.

We represent clinical diagnostic reasoning as a structured argument $A = \{D, R, W, B, Q, Y\}$ under the Toulmin model (see Appendix A). Here, D are case-grounded evidence items, R is a ranked differential with brief rebuttals, W links evidence to hypotheses via pathophysiology, B states supporting clinical principles, Q calibrates uncertainty and missing information, and Y is the final diagnosis.

2.1.2 Problem Formulation

We formalize trustworthy clinical diagnostic reasoning as a **structured text generation** task. Given a patient case presentation $P = \{p_1, p_2, \ldots, p_m\}$, the objective is to train a model M_θ to generate a diagnostic argument A that explicitly instantiates the Toulmin structure through a sequence of intermediate outputs $C^{(k)}$

$$A = f(P) = \{D, R, W, B, Q, Y\} \tag{1}$$

where the argument is constructed progressively through three curriculum stages:

$$\begin{aligned} C^{(1)} &= \{D, R\} \\ C^{(2)} &= C^{(1)} \cup \{W, B\} \\ C^{(3)} &= C^{(2)} \cup \{Q, Y\} \end{aligned} \tag{2}$$

The quality of argument A is quantified by our T-Eval framework, which assesses the integrity of all Toulmin components. The learning objective is to find optimal parameters that maximize the expected argument quality:

$$\theta^* = \arg\max_\theta \mathbb{E}_{P\sim\mathcal{P}} \left[T\text{-}Eval(A) \mid A = M_\theta(P)\right] \tag{3}$$

This formulation directly motivates our curriculum learning approach in Section 2.2, where we sequentially optimize for the generation of each component set $C^{(k)}$.

2.2 Curriculum Goal-Conditioned Learning

The CGCL framework is architected around the principle of **Goal-Conditioned Offline Imitation Learning**. This paradigm sequentially distills expert-level reasoning trajectories into a target

model under a structured curriculum, effectively circumventing the challenges associated with online reinforcement learning, such as reward design and exploration.

The process leverages a powerful, *frozen* **Strategy Model** as an expert demonstrator. For a given clinical case P, it generates a multitude of candidate reasoning steps. These candidates are rigorously evaluated by the **T-Eval** framework, which acts as a principled reward function, assigning quality scores based on the integrity of specific Toulmin components. The highest-scoring candidates are systematically selected and fused into optimal, structured reasoning trajectories $C^{(k)}$ for each curriculum stage k.

A target model M_θ, initialized from a base pretrained language model, is then trained to imitate these curated trajectories through supervised fine-tuning. The training proceeds sequentially across stages $k = 1, 2, 3$. At each stage, the model learns to generate the corresponding optimal trajectory $C^{(k)}$ by minimizing the negative log-likelihood objective:

$$\mathcal{L}^{(k)}(\theta) = \mathbb{E}_{(P,C^{(k)})\sim\mathcal{D}^{(k)}}\Big[-\log p(C^{(k)} \mid P;\theta)\Big] \tag{4}$$

Here, $\mathcal{D}^{(k)}$ denotes the stage-k curriculum dataset consisting of paired clinical cases and the selected/-fused optimal trajectories $C^{(k)}$ constructed by the strategy model and T-Eval.

Crucially, the model parameters are updated iteratively, with the model at stage k being initialized from the parameters of the stage $k-1$ model:

$$\theta^{(k)} = \arg\min_\theta \mathcal{L}^{(k)}(\theta), \quad \text{with } \theta^{(k)} \text{ initialized from } \theta^{(k-1)} \tag{5}$$

This iterative refinement, where $\theta^{(0)}$ represents the base model, ensures that competencies acquired in earlier, foundational stages are preserved and built upon, thereby progressively instilling a robust and transparent diagnostic reasoning capability.

2.2.1 Stage 1: Factual Grounding and Hypothesis Generation

The first stage of the curriculum targets the establishment of foundational clinical reasoning capabilities: comprehensive data extraction and systematic hypothesis generation. The objective is to produce an initial structured output $C^{(1)}$ that integrates two core Toulmin components—the clinical data (D) and a preliminary differential diagnosis (R).

To construct the training dataset $\mathcal{D}^{(1)}$, we employ a trajectory collection process. For each clinical case P, the strategy model M_{strategy} generates candidate sets $\{C_1^i\}$ for data extraction and $\{C_2^j\}$ for differential diagnosis, following respective instructional prompts. Each candidate is evaluated by the T-Eval framework, yielding quality scores $S_D(C_1^i)$ and $S_R(C_2^j)$. The optimal candidates, selected via $C_1^* = \arg\max_{c_1^i} S_D(C_1^i)$ and $C_2^* = \arg\max_{c_2^j} S_R(C_2^j)$, are subsequently fused by the strategy model into a coherent, composite output $C^{(1)}$. This fusion ensures logical continuity between the extracted evidence and the generated hypotheses, forming an optimal Stage 1 reasoning trajectory. The resulting dataset is defined as:

$$\mathcal{D}^{(1)} = \{(P_i, C_i^{(1)})\}_{i=1}^N. \tag{6}$$

The model M_θ is then initialized with the base parameters $\theta^{(0)}$ and fine-tuned on $\mathcal{D}^{(1)}$ to distill this structured reasoning capability. The stage concludes with the optimization:

$$\theta^{(1)} = \arg\min_\theta \mathbb{E}_{(P,C^{(1)})\sim\mathcal{D}^{(1)}}\Big[-\log p(C^{(1)} \mid P;\theta)\Big], \quad \text{with } \theta^{(1)} \text{ initialized from } \theta^{(0)}, \tag{7}$$

yielding a model proficient in transforming an unstructured clinical narrative into a solid foundation for subsequent diagnostic argumentation.

This stage instills the fundamental clinical discipline of separating objective observation from interpretation, mirroring the training of medical students who are first taught to meticulously gather facts before forming diagnostic hypotheses.

2.2.2 Stage 2: Argumentative Justification and Critical Refutation

Building upon the foundational outputs of Stage 1, the second curriculum stage aims to develop the model's capacity for deep, causal justification and critical evaluation of competing diagnoses. This stage focuses on generating the extended output $C^{(2)}$, which incorporates the Warrant (W) for the primary diagnosis and the Backing (B) for refuting alternatives, integrated with the prior output $C^{(1)}$.

The data collection protocol for Stage 2 extends the established methodology. For each case P, the strategy model is prompted to generate candidate warrants $\{C_3^p\}$ and candidate backing rationales $\{C_4^q\}$, conditioned on the optimal Stage 1 output $C^{(1)}$. These candidates are evaluated by T-Eval, receiving scores $S_W(C_3^p)$ and $S_B(C_4^q)$, respectively.

The optimal candidates C_3^* and C_4^* are selected based on these scores. A critical fusion step is then performed, wherein the strategy model synthesizes $C^{(1)}$, C_3^*, and C_4^* into a unified and logically coherent argument $C^{(2)}$. This composite output represents a complete diagnostic justification up to this point. The dataset for this stage is constructed as an augmentation of the previous one:

$$\mathcal{D}^{(2)} = \mathcal{D}^{(1)} \cup \{(P_i, C_i^{(2)})\}_{i=1}^{N}. \tag{8}$$

The distillation objective for Stage 2 is to fine-tune the Stage 1 model to now generate the more complex output $C^{(2)}$. The optimization leverages the parameters $\theta^{(1)}$ as the starting point, ensuring retention of Stage 1 capabilities while learning new skills:

$$\theta^{(2)} = \arg\min_{\theta} \mathbb{E}_{(P,C^{(2)})\sim\mathcal{D}^{(2)}} \Big[-\log p(C^{(2)} \mid P; \theta) \Big], \quad \text{with } \theta^{(2)} \text{ initialized from } \theta^{(1)}. \tag{9}$$

This progression from fact collection to critical reasoning reflects the natural advancement in clinical training, where trainees evolve from passive information gatherers to active, analytical diagnosticians capable of defending their conclusions.

2.2.3 Stage 3: Synthesis and Qualified Conclusion

The final stage cultivates the expert-level skill of diagnostic synthesis and calibration, requiring the model to integrate all preceding analysis into a definitive yet appropriately qualified conclusion. A key innovation is the mandatory evidence-based revision mechanism, which instills intellectual honesty and meta-cognitive awareness, namely the ability to recognize and correct one's own diagnostic errors based on conflicting evidence. The target output is the complete diagnostic argument $C^{(3)}$, which formally integrates all components from previous stages along with the final qualified claim (Q, Y) and, when applicable, an evidence-based revision rationale Δ. The complete argument can thus be represented as $C^{(3)} = C^{(2)} \cup \{Q, Y, \Delta \cdot \mathbb{I}_{\text{rev}}\}$, where $\mathbb{I}_{\text{rev}} \in \{0, 1\}$ indicates whether the final diagnosis revises the initial top candidate.

The trajectory collection for Stage 3 completes the reasoning process. Conditioned on the full contextual argument $C^{(2)}$, the strategy model generates candidate final diagnoses and qualifiers $\{C_5^r\}$. T-Eval assesses these candidates with emphasis on claim correctness (S_Y) and qualifier quality (S_Q), where the latter is activated when the final diagnosis revises the initial top candidate. The optimal candidate C_5^* is fused with $C^{(2)}$ to produce the final trajectory $C^{(3)}$, forming the ultimate training dataset:

$$\mathcal{D}^{(3)} = \mathcal{D}^{(2)} \cup \{(P_i, C_i^{(3)})\}_{i=1}^{N}. \tag{10}$$

The model distillation objective for Stage 3 fine-tunes the Stage 2 model to generate the complete argument $C^{(3)}$ from the clinical case P:

$$\theta^{(3)} = \arg\min_{\theta} \mathbb{E}_{(P,C^{(3)})\sim\mathcal{D}^{(3)}} \Big[-\log p\big(C^{(3)} \mid P; \theta\big) \Big], \quad \text{with } \theta^{(3)} \text{ initialized from } \theta^{(2)}. \tag{11}$$

The resulting model $M_{\theta^{(3)}}$ embodies the culmination of the CGCL pipeline, capable of executing end-to-end diagnostic reasoning that transparently exhibits all Toulmin components through structured reasoning trajectories.

2.3 T-Eval: A Toulmin-Based Evaluation Framework

2.3.1 Evaluation Dimensions and Scoring Metrics

The T-Eval framework provides a principled methodology for quantifying the quality of diagnostic reasoning by directly operationalizing the Toulmin components established in Section 2.1. We define six evaluation dimensions that correspond to the core elements of our clinical argumentation mapping: **Data Score** (S_D): Measures completeness and accuracy of **Clinical Evidence** extraction. **Warrant Score** (S_W): Assesses medical plausibility of **Pathophysiological Rationale**. **Backing Score** (S_B): Quantifies appropriate use of **Principled Justification**. **Rebuttal Score** (S_R): Evaluates thoroughness of **Differential Diagnosis** analysis. **Qualifier Score** (S_Q): Assesses the appropriateness and justification of the diagnostic **Certainty Assessment**. **Claim Score** (S_Y)**:** Measures the correctness of the final diagnosis.

Each dimension score $S_c^{(i)}$ is rated on a 1–5 Likert scale. We first normalize it to $\tilde{S}_c^{(i)} = (S_c^{(i)} - 1)/4 \in [0, 1]$. The overall **T-Eval TrustScore** is then computed as the average normalized reasoning quality across the five Toulmin components, scaled to $[0, 100]$:

$$\text{T-Eval TrustScore} = \frac{100}{5N} \sum_{i=1}^{N} \sum_{c\in\{D,W,B,R,Q\}} \tilde{S}_c^{(i)}. \tag{12}$$

MedCaseReasoning					
Method		**Qwen2.5-7B-Instruct**		**Qwen2.5-3B-Instruct**	
Category	Variant	**Accuracy** (↑)	**TrustScore** (↑)	**Accuracy** (↑)	**TrustScore** (↑)
Prompt	Vanilla	22.37	-	20.61	-
	Think & Answer	25.68	62.43	23.47	61.32
	Trust-Think & Answer	26.31	65.52	23.42	63.28
SFT	SFT-GT	31.72	63.74	28.16	62.38
	SFT-CoT	32.08	60.79	28.27	59.83
	SFT-CGCL(Stage3)	30.12	70.78	26.58	68.67
RL	DPO	32.24	71.18	28.59	69.27
	GRPO	**32.81**	**73.12**	28.90	71.25
Ours	**CGCL**	32.63	72.84	**28.95**	**71.30**
Method		**LLaMA3.1-8B-Instruct**		**LLaMA3.2-3B-Instruct**	
Category	Variant	**Accuracy** (↑)	**TrustScore** (↑)	**Accuracy** (↑)	**TrustScore** (↑)
Prompt	Vanilla	22.46	-	20.57	-
	Think & Answer	25.73	62.58	23.51	61.29
	Trust-Think & Answer	26.38	65.47	23.43	63.36
SFT	SFT-GT	31.79	63.81	28.12	62.49
	SFT-CoT	32.14	60.86	28.33	59.91
	SFT-CGCL(Stage3)	30.07	70.92	26.64	68.59
RL	DPO	32.29	71.22	28.63	69.31
	GRPO	**32.83**	**73.24**	**28.95**	71.30
Ours	**CGCL**	32.67	72.89	28.90	**71.35**

Table 1: Overall Performance Comparison on MedCaseReasoning Benchmark. Paired significance tests and 95% confidence intervals for the key CGCL comparisons are reported in Appendix D.5 (Table 9).

Relation to prior TrustScore. The term TrustScore has been used in prior work (Zheng et al., 2024a) as a reference-free trustworthiness metric for general LLM responses. In contrast, our metric aggregates rubric-based judgments over Toulmin components tailored to clinical diagnostic reasoning. To avoid ambiguity, we refer to our metric as T-Eval TrustScore throughout. We define diagnostic **Accuracy** using the T-Eval claim score $S_Y \in \{1, \ldots, 5\}$. Specifically, we normalize the claim score as $\tilde{S}_Y = (S_Y - 1)/4 \in [0, 1]$ and aggregate over cases:

$$Accuracy = \frac{100}{N} \sum_{i=1}^{N} \tilde{S}_Y^{(i)}. \tag{13}$$

This yields a calibrated 0–100 measure of claim correctness derived from rubric-guided judging, consistent with our component scoring. We compute TrustScore over $\{D, R, W, B, Q\}$ to measure reasoning quality *orthogonally* to end-task correctness; the final claim is evaluated separately via the T-Eval claim score S_Y (reported as Accuracy in Eq. 13).

2.3.2 Implementation with Calibrated LLM Judges

This implementation enables scalable and reproducible assessment via rubric-guided prompting and a multi-judge consensus mechanism (three independent LLM instances with aggregation and outlier handling). We emphasize that the main T-Eval scores are produced automatically by LLM judges.

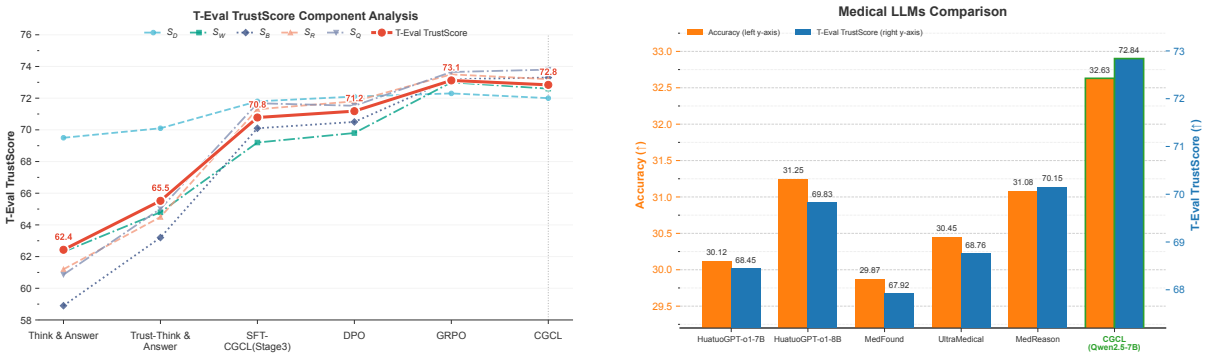
Clinician validation. We validated the automated T-Eval TrustScore against expert judgment by sampling cases from the MedCaseReasoning test set. Board-certified clinicians, blinded to the source, rated reasoning traces using the same Toulmin-aligned 1–5 Likert rubric. We averaged these ratings to form a Clinician TrustScore and calculated its correlation (Spearman's ρ) with the T-Eval TrustScore, alongside inter-rater reliability (ICC). See Appendix D.7 for full study protocols.

3 Experiments

3.1 Experimental Setup

Datasets. We conduct our experiments on MedCaseReasoning[2] (Wu et al., 2025b), a publicly available complex medical diagnostic reasoning benchmark. This open-access dataset comprises over 14,000 clinical diagnostic cases sourced from more than 800 medical journals across 30+ medical specialties. For evaluation, we use the official 897-case test split of MedCaseReasoning, which

[2] https://huggingface.co/datasets/zou-lab/MedCaseReasoning



(a) T-Eval TrustScore Component Analysis.

(b) Comparison with Medical LLMs.

Figure 3: Performance and trustworthiness analysis.

contains challenging diagnostic scenarios.

Baselines. We compare CGCL against a comprehensive set of baseline approaches, categorized into four groups:

Prompt-based. Vanilla Direct answer generation without intermediate reasoning steps. Think & Answer Standard chain-of-thought prompting where models reason within `<think>` tags before producing final answers in `<answer>` tags. Trust-Think & Answer Our enhanced reasoning approach that employs comprehensive argumentation to support step-by-step diagnostic reasoning.

SFT-based. SFT-GT Supervised fine-tuning using ground-truth answers only. SFT-CoT Supervised fine-tuning with common chain-of-thought reasoning traces. SFT-CGCL (Stage3) Learns the complete Toulmin structure in a single stage without progressive curriculum learning.

RL-based. DPO (Rafailov et al., 2023) Direct Preference Optimization using contrastive pairs of correct vs. incorrect reasoning trajectories. GRPO (Shao et al., 2024) Group Relative Policy Optimization with trajectory-level correctness rewards.

Medical LLMs. We include comparisons with state-of-the-art medical and reasoning-specialized models including HuatuoGPT-o1 (Chen et al., 2024), MedFound (Liu et al., 2025), UltraMedical (Zhang et al., 2024), and MedReason (Wu et al., 2025a) to establish comprehensive performance benchmarks. For reproducibility, we provide algorithmic details (Appendix C), training recipes, and hyperparameters (Appendix D), and prompt templates and judge rubrics (Appendix E).

Evaluation Metrics. We employ two primary evaluation metrics: **Accuracy** and **T-Eval TrustScore**. Diagnostic Accuracy measures claim correctness derived from the T-Eval claim score S_Y (Eq. 13), while T-Eval TrustScore summarizes reasoning quality over $\{D, R, W, B, Q\}$ (Eq. 12).

Implementation Details. Our experiments are conducted across diverse model architectures and scales: LLaMA3.1-8B, LLaMA3.2-3B (Dubey et al., 2024), Qwen2.5-7B/3B (Qwen et al., 2025). The strategy model for trajectory generation employs DeepSeek-R1 (Guo et al., 2025). For training-based approaches, models are fine-tuned on the training splits of each benchmark and evaluated on the respective test sets. Prompt-based methods are evaluated directly in a zero-shot setting.

Compute and training cost. To substantiate the resource-efficiency claim, we report GPU-hours, and peak GPU memory for CGCL, DPO, and GRPO under matched compute budgets and batch settings (Table 6; Appendix D.6).

3.2 Main Results

Performance Comparison. Table 1 summarizes the performance of CGCL and all baselines across model sizes. Overall, CGCL yields consistently higher reasoning quality than prompting and standard SFT baselines while maintaining competitive diagnostic accuracy. Compared with GRPO, CGCL is comparable across scales: it is marginally higher on 3B models (+0.05 TrustScore) and slightly lower on 7B/8B models. We further report paired significance tests and confidence intervals for these small gaps in Appendix D.5.

T-Eval TrustScore Analysis. The component-level analysis in Figure 3a reveals CGCL's balanced improvements across Toulmin components, with leading performance in backing quality (S_B = 73.3) and qualifier quality (S_Q = 73.8). Compared to prompt-based methods, CGCL achieves significant T-Eval TrustScore improvements, with notable

gains in warrant and backing qualities. The performance advantage over SFT-CGCL (Stage3) (72.84 vs. 70.78) underscores the importance of progressive curriculum. Since the SFT-CGCL(Stage3) baseline is trained on the exact same set of high-quality trajectories as the full CGCL model but without staging, the observed gain is directly attributable to the curriculum structure rather than data quality alone. While GRPO maintains slight edges in data, warrant, and rebuttal qualities on larger models, CGCL's superior performance on smaller models and in critical reasoning components validates its efficiency and balanced reasoning capabilities.

3.3 Analysis

Comparison with Medical LLMs. Figure 3b compares CGCL to recent medical and reasoning-specialized LLMs on MedCaseReasoning under the same prompting and scoring pipeline. Across both 3B and 8B backbones, CGCL achieves the best or competitive T-Eval TrustScore while maintaining similar diagnosis Accuracy. These results indicate that explicitly training Toulmin-structured argumentation can improve the organization and justification of diagnostic reasoning even when starting from a general-purpose backbone. We emphasize that this comparison is limited to MedCaseReasoning and our evaluation protocol, and does not measure broader clinical knowledge coverage or real-world safety. To further assess generalization beyond MedCaseReasoning, Appendix D.8 reports zero-shot OOD evaluation on the external MEDFOUND benchmark. The overall ranking trend remains consistent: CGCL outperforms prompting and standard SFT baselines and remains competitive with RL-based methods.

Clinician validation. Clinicians rank CGCL highest in the human evaluation, with a clinician TrustScore of 71.0 ± 0.7 and diagnosis accuracy of 53.3 (Table 2). Compared to GRPO (61.0 ± 0.8), CGCL yields a +10.0 gain in clinician TrustScore, mirroring the main T-Eval ranking and supporting that the improvements are not artifacts of LLM judging. Inter-rater reliability is moderate ($\text{ICC}(2, k) = 0.63$), and T-Eval correlates with clinician assessments with Spearman's $\rho = 0.74$, suggesting that T-Eval is a reasonable proxy for expert judgment while leaving room for future improvements in human-aligned evaluation.

Clinician Validation Performance		
Method	Clinician TrustScore ↑	Dx Acc. (%) ↑
Trust-Think & Answer	51.5	46.7
SFT-CoT	54.5	48.3
GRPO	61.0	51.7
CGCL (Ours)	**71.0**	**53.3**

Table 2: Clinician validation on $N_{\text{clin}} = 50$ randomly sampled test cases.

Ablation Studies. The curriculum stage ablation in Table 3 validates the progressive nature of our training approach. Stage 1 alone, focusing on factual grounding and hypothesis generation, achieves moderate performance (65.32 T-Eval TrustScore). Adding Stage 2, which introduces argumentative justification and critical refutation, provides a substantial 4.52 point T-Eval TrustScore improvement. The complete three-stage curriculum yields the best results, with an additional 3.00 point gain, demonstrating that synthesis and qualification capabilities are essential for optimal diagnostic reasoning. This monotonic improvement across stages confirms that each curriculum phase builds upon previously acquired skills, and compressing the training into fewer stages compromises reasoning quality.

Curriculum Stage Ablation		
Curriculum Stage	Accuracy (↑)	TrustScore (↑)
Stage 1 only	28.45	65.32
Stage 1+2 only	30.67	69.84
CGCL (full)	**32.63**	**72.84**

Table 3: Ablation study of curriculum learning stages.

4 Conclusion

We propose CGCL and the T-Eval evaluation framework, which enhance the transparency and trustworthiness of large language models in clinical diagnostic reasoning using the Toulmin model of argumentation. Experiments on MedCaseReasoning demonstrate that CGCL consistently improves T-Eval TrustScore compared to prompting and standard SFT baselines, while achieving accuracy comparable to strong RL-based approaches, particularly with smaller models. Ablation studies validate the effectiveness of our progressive curriculum design, showing continuous improvements in reasoning quality across the three-stage training process. Our results suggest that, compared to medical LLMs, structured reasoning training can partly compensate for the lack of domain-specific pre-training on MedCaseReasoning. Looking ahead, we aim

to explore richer supervision through multi-expert trajectory distillation and extend the framework to incorporate multi-modal clinical data. This will further enhance the model's practicality and reliability in real-world clinical environments, while maintaining its strong theoretical foundation for trustworthy clinical decision support systems.

Limitations

While the proposed CGCL framework demonstrates significant improvements in clinical diagnostic reasoning, several limitations should be addressed. First, the performance of the framework is inherently constrained by the quality and diversity of the trajectory generation model. This may not fully capture the complete range of clinical reasoning patterns. Second, although our evaluation is comprehensive, it primarily focuses on diagnostic accuracy and structured reasoning quality, without testing the model in real-time clinical workflows or integrating it with existing clinical decision support systems. Third, the current implementation relies on synthetic trajectory generation and distillation, which may not fully replicate the nuanced decision-making processes of human clinical experts across all medical specialties. While the Toulmin schema and T-Eval penalize inconsistent arguments, models can still arrive at correct final diagnoses despite flawed intermediate reasoning. A targeted qualitative audit to identify failure modes such as 'correct answer, wrong reason' at intermediate steps will be an important next step. The system is intended as decision support and must be used under clinician supervision; it is not a substitute for professional medical judgment. Finally, the computational requirements of the multi-stage training process, particularly trajectory generation and fusion, present challenges for rapid deployment in resource-constrained clinical settings. These limitations point to key directions for future research and further refinement of our approach.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. *Advances in neural information processing systems*, 30.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Luciano Caroprese, Eugenio Vocaturo, and Ester Zumpano. 2022. Argumentation approaches for explanaible ai in medical informatics. *Intelligent Systems with Applications*, 16:200109.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Sibel Erduran, Shirley Simon, and Jonathan Osborne. 2004. Tapping into argumentation: Developments in the application of toulmin's argument pattern for studying science discourse. *Science education*, 88(6):915–933.

Hamzah Noori Fejer, Ali Hadi Hasan, and Ahmed T Sadiq. 2022. A survey of toulmin argumentation approach for medical applications. *International Journal of Online & Biomedical Engineering*, 18(2).

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. 2019. Learning to reach goals via iterated supervised learning. *arXiv preprint arXiv:1912.06088*.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. Pmlr.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Hyunjung Ju, Ikseon Choi, and Bo Young Yoon. 2017. Do medical students generate sound arguments during small group discussions in problem-based learning?: an analysis of preclinical medical students' argumentation according to a framework of hypothetico-deductive reasoning. *Korean journal of medical education*, 29(2):101.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational linguistics*, 45(4):765–818.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Xiao-huan Liu, Zhen-hua Lu, Tao Wang, and Fei Liu. 2024. Large language models facilitating modern molecular biology and novel drug development. *Frontiers in Pharmacology*, 15:1458739.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, and 1 others. 2025. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, and 1 others. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, Peter Brodeur, and 1 others. 2025. Assessment of large language models in clinical reasoning: A novel benchmarking study. *NEJM AI*, 2(10):AIdbp2500120.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Rafał Obuchowicz, Julia Lasek, Marek Wodziński, Adam Piórkowski, Michał Strzelecki, and Karolina Nurzynska. 2025. Artificial intelligence-empowered radiology—current status and critical review. *Diagnostics*, 15(3):282.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. 2018. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pages 4344–4353. PMLR.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. 2015. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320. PMLR.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Songtao Tan, Xin Xin, and Di Wu. 2024. Chatgpt in medicine: prospects and challenges: a review article. *International journal of surgery*, 110(6):3701–3706.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36:e5.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, and 1 others. 2025a. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025b. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Erlan Yu, Xuehong Chu, Wanwan Zhang, Xiangbin Meng, Yaodong Yang, Xunming Ji, and Chuanjie Wu. 2025. Large language models in medicine: Applications, challenges, and future directions. *International Journal of Medical Sciences*, 22(11):2792.

Zhenghang Yuan, Lichao Mou, Qi Wang, and Xiao Xiang Zhu. 2022. From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. *IEEE transactions on geoscience and remote sensing*, 60:1–11.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, and 1 others. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. *arXiv preprint arXiv:2506.11474*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine.

Danna Zheng, Danyang Liu, Mirella Lapata, and Jeff Z Pan. 2024a. Trustscore: Reference-free evaluation of llm response trustworthiness. *arXiv preprint arXiv:2402.12545*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Problem Formulation and Toulmin Model Instantiation

A.1 The Toulmin Model of Argumentation

The Toulmin model provides a robust framework for analyzing informal arguments by decomposing them into six fundamental components: (1) **Claim (Y)**: the final assertion or conclusion; (2) **Data (D)**: the facts and evidence used to support the claim; (3) **Warrant (W)**: the logical bridge connecting the

Table 4: Notation, data objects, and evaluation scores used in this work.

Symbol	Description
I. Clinical data & model objects	
P	**Clinical case**: unstructured narrative containing symptoms, signs, labs, and imaging findings.
M_{strategy}	**Strategy model**: frozen LLM used to generate candidate reasoning components.
$\{C_{i,j}^{Z}\}_{j=1}^{K_Z}$	**Candidate set**: candidates for component Z, where $Z \in \{D, R, W, B, Q, Y\}$.
II. Toulmin-aligned reasoning components	
D	**Data**: objective clinical evidence extracted from P.
R	**Rebuttal**: differential diagnoses capturing alternative hypotheses and counter-arguments.
W	**Warrant**: pathophysiological rationale connecting evidence (D) to hypotheses.
B	**Backing**: principled justification (guidelines, medical knowledge) supporting W.
Q	**Qualifier**: uncertainty calibration and conditions under which the claim holds.
Y	**Claim**: final diagnosis (ultimate clinical judgment).
III. Curriculum learning & evaluation	
$C^{(k)}$	**Stage-k trajectory** with progressive structure: • Stage 1: $C^{(1)} = (D, R)$ • Stage 2: $C^{(2)} = (D, R, W, B)$ • Stage 3: $C^{(3)} = (D, R, W, B, Q, Y)$
$\mathcal{D}^{(k)}$	**Curriculum dataset**: $\mathcal{D}^{(k)} = \{(P_i, C_i^{(k)})\}_{i=1}^{N}$, containing fused optimal stage-k trajectories.
S_D	**Data score**: Likert (1–5) rating of evidence completeness, accuracy, and grounding in P.
S_R	**Rebuttal score**: Likert (1–5) rating of differential coverage and quality of counter-arguments.
S_W	**Warrant score**: Likert (1–5) rating of mechanistic plausibility linking D to leading hypotheses.
S_B	**Backing score**: Likert (1–5) rating of principled justification quality (clinical rules/knowledge) supporting W, penalizing fabricated claims.
S_Q	**Qualifier score**: Likert (1–5) rating of uncertainty calibration, missing-information awareness, and conditions/limitations stated in Q.
S_Y	**Claim score**: Likert (1–5) rating of the final diagnosis Y, primarily reflecting correctness and consistency with the case evidence.
$\tilde{S}_Z$	**Normalized score**: $\tilde{S}_Z = (S_Z - 1)/4 \in [0, 1]$ for $Z \in \{D, R, W, B, Q, Y\}$.
TrustScore	**T-Eval TrustScore**: average of normalized component scores (over the evaluated components), scaled to $[0, 100]$ (Eq. 12).
Accuracy	**Diagnostic accuracy:** aggregated claim correctness derived from S_Y (Eq. 13).

data to the claim; (4) **Backing (B)**: the general principles or authority that reinforces the warrant; (5) **Qualifier (Q)**: the modality or degree of certainty associated with the claim; and (6) **Rebuttal (R)**: the consideration and refutation of counter-arguments or alternative claims.

A.2 Clinical Diagnostic Reasoning as Structured Argumentation

We propose that the clinical diagnostic process is fundamentally an exercise in structured argumentation. To operationalize the Toulmin model for our CGCL framework, we establish the following precise mapping from its components to core clinical reasoning concepts:

Data (D): **Clinical Evidence** : The objective facts extracted from the patient's presentation, including symptoms, signs, laboratory results, and imaging findings. This constitutes the foundational evidence for the diagnostic argument.

Rebuttal (R): **Differential Diagnosis** : The systematic consideration and ranked listing of alternative diagnostic hypotheses. The process of comparing and prioritizing these competing claims embodies the rebuttal component by arguing against the exclusive validity of any single hypothesis.

Warrant (W): **Pathophysiological Rationale** : The logical bridge and mechanistic explanation that connects the clinical evidence to a specific diagnostic hypothesis. It provides the causal reasoning, typically grounded in disease mechanisms, for why the evidence supports the claim.

Backing (B): **Principled Justification** : The established medical knowledge, clinical guidelines, or diagnostic criteria that authorize and validate the warrants. This component provides the authoritative foundation for both supporting the primary diagnosis and refuting alternatives.

Qualifier (Q): **Certainty Assessment** : The explicit evaluation of diagnostic confidence and,

when applicable, the rationale for any revision from initial hypotheses. This component reflects the calibrated nature of clinical decision-making and acknowledges diagnostic uncertainty.

Claim (Y): Final Diagnosis : The final diagnostic determination that synthesizes all available evidence and reasoning. This represents the definitive clinical judgment based on the complete analytical process. This structured mapping ensures that each abstract Toulmin component corresponds to a concrete, clinically-meaningful reasoning step, providing a transparent foundation for our staged curriculum and evaluation framework. Because MedCaseReasoning requires **free-form** diagnostic conclusions rather than selecting from a fixed option list, exact-string-match accuracy is ill-defined (e.g., synonyms, abbreviations, clinically equivalent variants, or multi-condition diagnoses). We therefore operationalize **Accuracy** using T-Eval's rubric-guided **claim correctness** score $S_Y \in \{1, \ldots, 5\}$, where 5 denotes an exact match and 4 denotes a clinically equivalent variant. We normalize S_Y to $[0, 1]$ and report the dataset-level mean scaled to $[0, 100]$ as Accuracy (Eq. 13), making it consistent with our component-wise scoring protocol.

B Related Work

Medical Large Models and Diagnostic Reasoning. Research on medical large language models has primarily advanced along two directions: 1) developing domain-specific architectures and pre-training strategies, such as ClinicalBERT and BioBERT, which are continually pre-trained on clinical texts (Huang et al., 2019; Lee et al., 2020); and 2) enhancing the performance of general-purpose large models on medical question-answering tasks through instruction fine-tuning, as seen with Med-PaLM and ChatDoctor (Singhal et al., 2023; Li et al., 2023). These studies typically rely on standardized benchmarks like MedQA and PubMedQA for evaluation, employing a paradigm centered on answer-matching (Jin et al., 2019; Pal et al., 2022). This approach, however, struggles to detect instances of "correct answers from flawed reasoning." Although recent work has attempted to elicit reasoning steps via Chain-of-Thought (CoT) prompting, the resulting free-form reasoning chains lack structural constraints, and their logical rigor and medical reliability cannot be effectively evaluated or guaranteed (Wei et al., 2022; Wang et al., 2022; Zhou et al., 2022; Yao et al., 2023).

The Toulmin Model of Argumentation. The Toulmin model serves as a classic and robust theoretical framework for analyzing informal arguments by deconstructing them into six core structured components: Claim, Data (Grounds), Warrant, Backing, Qualifier, and Rebuttal (Erduran et al., 2004). This model has seen widespread application in fields such as computational argumentation and educational assessment, where it has been used, for instance, to automatically analyze the argumentation quality of student essays (Lawrence and Reed, 2020). In the AI domain, it has also been explored as an explainability tool for parsing the decision logic of intelligent systems (Vassiliades et al., 2021). However, while Toulmin-style argumentation has been discussed in prior medical AI literature (Fejer et al., 2022), existing work has predominantly focused on using the model for post-hoc analysis of arguments rather than as an a priori structural constraint to be systematically integrated into a model's training process to directly shape its reasoning capabilities. In the more specific setting of long-form clinical diagnostic reasoning, using Toulmin as an explicit training objective together with a scalable evaluation framework remains relatively underexplored.

Curriculum Learning and Goal-Conditioned Learning. Curriculum Learning (CL) enhances model training by presenting data or tasks in a sequence of increasing difficulty, a strategy validated for its ability to improve generalization and convergence in tasks like machine translation and visual reasoning (Soviany et al., 2022; Graves et al., 2017). Goal-Conditioned Learning (GCL), on the other hand, guides a model to produce target-oriented outputs by incorporating a goal signal into its input. This paradigm has proven highly effective in reinforcement learning and robotics and is gradually being introduced to sequence generation tasks (Yuan et al., 2022; Ghosh et al., 2019; Riedmiller et al., 2018; Dathathri et al., 2019). To date, research has largely explored these two paradigms in isolation: CL has centered on data scheduling, while GCL has focused on goal encoding. How to deeply synthesize the staged progression of CL with the explicit guidance of GCL to construct a systematic training pathway for complex reasoning capabilities remains a compelling open research question.

C Algorithm Details

C.1 Notation and Data Objects

Table 4 summarizes the notation used in CGCL and T-Eval to support reproducibility and ease cross-referencing with Algorithms 1–2.

C.2 CGCL Trajectory Construction

Algorithm 1 describes how we construct the stage-wise optimal trajectories using the frozen strategy model and T-Eval.

Algorithm 1 CGCL trajectory construction for a single case P (sequential, context-conditioned selection).

Require: Clinical case P; strategy model M_{strategy}; evaluator $\textsc{TEval}(\cdot)$; candidate budgets $\{K_D, K_R, K_W, K_B, K_Q, K_Y\}$; fusion operator $\textsc{Fuse}(\cdot)$.
Ensure: Stage trajectories $C^{(1)}, C^{(2)}, C^{(3)}$.
Notes: C_{ctx} denotes the current partial trajectory (context).
$\oplus$ merges a candidate component into C_{ctx}.
$\textsc{TEval}(P, C_{\text{ctx}} \oplus \cdot, \text{dim} = Z)$ returns a Likert score $S_Z \in \{1, \ldots, 5\}$.

function $\textsc{SelectBest}(Z, C_{\text{ctx}}, K_Z)$ ▷ $Z \in \{D, R, W, B, Q, Y\}$
  $\{C_j^Z\}_{j=1}^{K_Z} \leftarrow M_{\text{strategy}}(P, C_{\text{ctx}}, \text{prompt}_Z)$
  for $j = 1$ to K_Z **do**
    $S_j^Z \leftarrow \textsc{TEval}(P, C_{\text{ctx}} \oplus C_j^Z, \text{dim} = Z)$
  end for
  return $C^{Z*} \leftarrow C^Z_{\arg\max_j S_j^Z}$
end function
Stage 1: construct (D, R)
$C^{D*} \leftarrow \textsc{SelectBest}(D, \emptyset, K_D)$
$C^{R*} \leftarrow \textsc{SelectBest}(R, C^{D*}, K_R)$
$C^{(1)} \leftarrow \textsc{Fuse}(P; C^{D*}, C^{R*})$
Stage 2: augment with (W, B)
$C^{W*} \leftarrow \textsc{SelectBest}(W, C^{(1)}, K_W)$
$C^{B*} \leftarrow \textsc{SelectBest}(B, C^{(1)} \oplus C^{W*}, K_B)$
$C^{(2)} \leftarrow \textsc{Fuse}(P; C^{(1)}, C^{W*}, C^{B*})$
Stage 3: augment with (Q, Y)
$C^{Q*} \leftarrow \textsc{SelectBest}(Q, C^{(2)}, K_Q)$
$C^{Y*} \leftarrow \textsc{SelectBest}(Y, C^{(2)} \oplus C^{Q*}, K_Y)$
$C^{(3)} \leftarrow \textsc{Fuse}(P; C^{(2)}, C^{Q*}, C^{Y*})$
return $C^{(1)}, C^{(2)}, C^{(3)}$

Fusion operator. FUSE produces a coherent structured output with a fixed schema (Appendix E.4). The FUSE operator synthesizes selected components into a coherent, structured output adhering to a strict schema. To ensure integrity, we employ a deterministic post-processing pipeline consisting of three steps: (i) De-duplication to remove redundant clinical evidence; (ii) Consistency Verification between the evidence (D) and final diagnosis (Y), where a detected conflict triggers a one-time regeneration of the qualifier (Q) and diagnosis (Y); and (iii) Format Validation to guarantee machine parsability.

C.3 T-Eval Scoring and Multi-Judge Aggregation

Algorithm 2 details how T-Eval uses multiple independent LLM judges and robust aggregation to score each dimension.

Algorithm 2 T-Eval scoring with multi-judge robust aggregation.

Require: Case P; candidate (partial) trajectory C; target dimension $\delta \in \{D, R, W, B, Q, Y\}$; judges $\{J_m\}_{m=1}^{3}$; rubric $\mathcal{R}_\delta$.
Ensure: Aggregated Likert score $S_\delta(C) \in \{1, 2, 3, 4, 5\}$.
for $m = 1$ to 3 **do**
  Query judge J_m with $(P, C, \mathcal{R}_\delta)$ and obtain $s_m \in \{1, \ldots, 5\}$.
end for
Sort $\{s_1, s_2, s_3\}$ to obtain $s_{(1)} \le s_{(2)} \le s_{(3)}$.
if $s_{(3)} - s_{(1)} \ge 3$ **then** ▷ high disagreement
  $\hat{s} \leftarrow s_{(2)}$ ▷ use median
else
  $\hat{s} \leftarrow \text{round}\left(\frac{s_1+s_2+s_3}{3}\right)$
end if
return $S_\delta(C) \leftarrow \min(5, \max(1, \hat{s}))$

Rubrics. Each rubric $\mathcal{R}_\delta$ contains anchor descriptions for scores 1–5. All rubrics and judge prompts are provided in Appendix E.5.

C.4 Clinician Validation Protocol (Evaluation Method)

This section specifies the clinician-validation protocol used to ground T-Eval in expert judgment. We describe only the protocol here; results are reported in the main paper (Section 3.3) and additional breakdowns are provided in Appendix D.7.

Sampling. We randomly sample $N_{\text{clin}} = 50$ cases from the official test split (stratified by specialty) and collect one reasoning trace per evaluated method under the same decoding settings as the main experiments.

Raters and blinding. We recruit $K_{\text{clin}} = 3$ board-certified clinicians to independently rate anonymized reasoning traces. Raters are blinded to model identity, training method, and T-Eval scores; we randomize presentation order to mitigate fatigue and ordering effects.

Rating task. Clinicians rate all six components (D, R, W, B, Q, Y) using the same 1–5 Likert rubrics as T-Eval.

Aggregation and agreement. We average ratings across raters to form a clinician score per component. We compute Clinician TrustScore by averag-

Table 5: Training hyperparameters for Qwen2.5-3B-Instruct. We use long-context training ($L_{\max} = 4096$), hence micro-batch is set to 1 and we vary gradient accumulation to match compute budgets.

Method	LR	Global batch	Steps/Epochs	Max len	Warmup	Notes
SFT-GT	2e-4	16	3 epochs	4096	0.03	LoRA $r = 64, \alpha = 32$
SFT-CoT	2e-4	16	3 epochs	4096	0.03	LoRA $r = 64, \alpha = 32$
SFT-CGCL(Stage3)	1e-4	16	3 epochs	4096	0.03	single-stage baseline
CGCL Stage 1	2e-4	16	1 epoch	4096	0.03	train on $\mathcal{D}^{(1)}$
CGCL Stage 2	2e-4	16	1 epoch	4096	0.03	train on $\mathcal{D}^{(2)}$
CGCL Stage 3	1e-4	16	1 epoch	4096	0.03	train on $\mathcal{D}^{(3)}$
DPO	5e-6	8	1 epoch	4096	0.10	$\beta = 0.1$; pair construction
GRPO	3e-6	4	$\sim$230 steps	4096	0.10	$G = 8$, $\epsilon = 0.2$, KL penalty

ing normalized scores over (D, R, W, B, Q), and compute Clinician Accuracy by normalizing and averaging the claim score S_Y as in Eq. 13. We quantify (i) inter-rater reliability via ICC (two-way random effects; ICC$(2, k)$) and (ii) alignment between clinician and T-Eval TrustScore using Spearman's ρ.

Ethics. The study uses only de-identified, publicly available benchmark cases and collects no patient-identifying information. Clinician raters participated voluntarily under standard institutional guidelines.

D Detailed Experimental Setup

D.1 Models and Training Recipes

Base models. We evaluate multiple instruction-tuned LLM backbones spanning model sizes, including: LLaMA3.2-3B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and LLaMA3.1-8B-Instruct. All models are used in their original released checkpoints.

Training framework. All training-based methods are implemented using open-source training stacks (LLaMA-Factory) (Zheng et al., 2024b) with identical tokenization, data packing, and logging across methods. Unless otherwise stated, we adopt parameter-efficient tuning for stability under long-context training(§D.3).

SFT baselines. SFT-GT performs supervised fine-tuning on (P, Y) pairs, i.e., predicting only the final diagnosis without explicit intermediate reasoning. **SFT-CoT** fine-tunes on the expert-written chain-of-thought rationales released with MedCaseReasoning. The model is trained to reproduce the provided rationale and final diagnosis under the official formatting, without enforcing the Toulmin-structured schema used by CGCL. This may underperform structured prompting because the released CoT rationales are not explicitly aligned with the Toulmin schema and can be noisy. Fine-tuning on such traces may encourage superficial pattern imitation rather than improved structured reasoning. **SFT-CGCL(Stage3)** trains directly on Stage 3 trajectories without curriculum staging.

CGCL training. CGCL trains sequentially on $\mathcal{D}^{(1)}$, then $\mathcal{D}^{(2)}$, then $\mathcal{D}^{(3)}$. At each stage k, we initialize $\theta^{(k)}$ from $\theta^{(k-1)}$ and perform maximum-likelihood training on $(P, C^{(k)})$. Stage-specific training lengths and learning rates are reported in Table 5.

D.2 Baselines: RL-based Methods

Initialization. To ensure that the policy can reliably generate parsable structured trajectories, both DPO and GRPO start from the corresponding **SFT-CGCL(Stage3)** checkpoint of the same backbone.

DPO. We implement Direct Preference Optimization (DPO) by constructing preference pairs (C^+, C^-) for the same case P. For each case, we sample multiple complete trajectories from the current policy and categorize them by diagnosis correctness: C^+ is any trajectory whose final claim Y matches the gold diagnosis after output normalization, and C^- is any trajectory that does not. We construct preference pairs using a gold-label exact-match indicator for training stability. If both sets are non-empty, we uniformly sample one element from each set to form a pair; otherwise the case is skipped for that epoch. We optimize DPO with $\beta = 0.1$ and the hyperparameters in Table5.

GRPO. We implement Group Relative Policy Optimization (GRPO) by sampling a group of $G = 8$ trajectories per case and computing a trajectory-level reward. The primary reward is diagnosis correctness: $r = 1$ if the normalized final diagnosis matches gold, and $r = 0$ otherwise. The primary reward is a gold-label exact-match indicator ($r = 1$

Table 6: Estimated training cost for Qwen2.5-3B-Instruct. Peak memory denotes the maximum memory allocated per GPU during training. DPO requires fewer GPU-hours than CGCL but substantially higher peak memory. GRPO has the highest cost in both peak memory and GPU-hours and additionally requires online rollouts. CGCL uses single-GPU offline training with moderate memory requirements while maintaining competitive reasoning quality.

Method	GPUs	Available GPU memory	Peak memory (GB)	GPU-hours	Online rollouts
SFT-GT	1	80GB	18	1.1	No
SFT-CoT	1	80GB	20	1.2	No
DPO	1	80GB	68	2.6	No
GRPO	2	80GB	76	8.3	Yes
CGCL (Stage 1–3)	1	80GB	21	3.4	No

Table 7: Clinician-rated component scores (1–5 Likert) on the same validation subset as Table 2. TrustScore is computed by normalizing each dimension score $\tilde{S}_Z = (S_Z - 1)/4$ and averaging over $Z \in \{D, R, W, B, Q\}$, then scaling to $[0, 100]$. To simplify calculations, values are rounded to one decimal place.

Method	S_D	S_R	S_W	S_B	S_Q	TrustScore ↑
Trust-Think & Answer	3.4	3.0	3.1	3.0	2.8	51.5
SFT-CoT	3.6	3.1	3.2	3.1	2.9	54.5
GRPO	3.8	3.3	3.5	3.4	3.2	61.0
CGCL (Ours)	**4.1**	**3.7**	**3.9**	**3.8**	**3.7**	**71.0**

Table 8: Clinician agreement and alignment with T-Eval on the clinician validation subset.

Metric	Value	Notes
ICC(2,k)	0.63	Inter-rater reliability (overall)
Spearman ρ	0.74	Clinician TrustScore vs T-Eval TrustScore

if the normalized final diagnosis matches the gold label, else 0), which serves as an RL training signal and should not be confused with the reported ACCURACY derived from S_Y (Eq. 13). To discourage invalid structured outputs, unparsable trajectories receive a reward of 0 and an additional small formatting penalty of -0.1. We use a standard clipped policy update ($\epsilon = 0.2$) with a KL regularization term to stabilize training, following the hyperparameters in Table 5. Unlike CGCL, GRPO requires online rollouts during training.

D.3 Hyperparameters and Implementation Details

Hyperparameters. Table 5 reports training hyperparameters for `Qwen2.5-3B-Instruct`. In particular, to support $L_{\max} = 4096$, we fix micro-batch $= 1$ and adjust gradient accumulation to realize the listed effective global batch sizes. This matches the design intent of comparing methods under realistic memory constraints, rather than forcing a single batch size across all objectives.

Parameter-efficient tuning. Unless otherwise stated, we use LoRA for all supervised objectives with the configuration reported in Table 5. For RL baselines, we use LoRA for the policy and keep the reference model frozen.

D.4 Evaluation Protocol and Output

Decoding for evaluation. For the main test evaluation, we use greedy decoding (temperature $= 0$) to reduce stochastic variance across methods. When sampling is required (e.g., for DPO pair construction or GRPO rollouts), we use temperature $= 0.7$, top-$p = 0.95$, and $max\_new\_tokens = 512$.

T-Eval implementation. We use Qwen2.5-32B-Instruct as the LLM judge for T-Eval, as it provides a practical balance of open-weight reproducibility, judging quality, and computational cost, while also showing good alignment with clinician ratings in our validation study. For each dimension $\delta \in \{D, R, W, B, Q, Y\}$, we run three independent judge instances and aggregate scores with median-based outlier handling (Algorithm 2). The three-judge setup primarily serves as a fault-tolerant mechanism for large-scale automated evaluation: if one instance produces malformed or unparsable output, the remaining instances provide a stable fallback for score aggregation. Judge decoding uses temperature $= 0$, top-$p = 1.0$, and `max_new_tokens` $= 64$.

Comparison	Metric	Δ (CGCL $-$ baseline)	95% CI / p
CGCL vs. GRPO	Accuracy	+0.05	[-0.41, 0.52], p=0.81
CGCL vs. GRPO	TrustScore	+0.05	[-0.38, 0.47], p=0.79
CGCL vs. SFT-CGCL(Stage3)	Accuracy	+2.37	[0.96, 3.74], p=0.002
CGCL vs. SFT-CGCL(Stage3)	TrustScore	+2.63	[1.54, 3.71], p<0.001

Table 9: Paired bootstrap significance tests on MedCaseReasoning for Qwen2.5-3B-Instruct.

D.5 Statistical Testing and Confidence Intervals

We compute case-level T-Eval TrustScore and diagnosis accuracy on the same $N = 897$ official test cases, and perform paired bootstrap resampling with $B = 10000$ resamples. Table 9 reports the paired difference, 95% confidence interval, and p-value for the two key comparisons discussed in the main text on Qwen2.5-3B-Instruct. The CGCL–GRPO gaps are very small for both Accuracy and TrustScore and are not statistically significant, consistent with our claim that the two methods are comparable on this backbone. In contrast, CGCL significantly outperforms SFT-CGCL(Stage3) on both metrics, supporting the benefit of progressive curriculum learning over single-stage training on the same final structured trajectories.

D.6 Compute and Training Cost

Measurement. Peak GPU memory is the maximum memory allocated per GPU during training. GPU-hours are computed as the product of the number of GPUs and wall-clock training time. Table 6 summarizes representative profiling results for Qwen2.5-3B-Instruct under a fixed software stack and sequence length, with method-specific batch and gradient-accumulation settings chosen to match comparable memory budgets.

D.7 Clinician Validation: Additional Details

This section provides additional details complementing the clinician validation.

Sampling and setup. We randomly sample $N_{\text{clin}} = 50$ cases from the official test split. For each case, we collect one model-generated reasoning trace per evaluated method under the same decoding settings used in the main experiments. All traces are anonymized to remove method-identifying markers.

Raters and blinding. Board-certified clinicians independently rate each trace while blinded to model identity, training method, and T-Eval scores. We randomize the presentation order across cases and methods.

Rating task and aggregation. Clinicians rate all six components (D, R, W, B, Q, Y) using the same 1–5 Likert rubrics as T-Eval. We average ratings across clinicians to obtain a clinician score per component. We compute Clinician TrustScore by averaging normalized scores over $\{D, R, W, B, Q\}$ (Eq. 12), and compute Clinician Accuracy by normalizing and averaging the claim score S_Y as in Eq. 13. Component-level clinician ratings are summarized in Table 7.

Agreement and alignment. We compute Spearman's ρ between Clinician TrustScore and T-Eval TrustScore, and quantify inter-rater reliability with ICC$(2, k)$. The resulting agreement and alignment statistics are summarized in Table 8.

D.8 Out-of-Distribution Evaluation on MedFound

To assess whether the gains of CGCL generalize beyond MedCaseReasoning, we additionally evaluate representative methods on MEDFOUND, an external benchmark of real clinical cases, in a strictly zero-shot out-of-distribution (OOD) setting. No method is trained, tuned, or otherwise adapted on MEDFOUND. For training-based baselines (SFT/RL), we directly evaluate the same checkpoints trained on MedCaseReasoning. Prompt-based baselines are evaluated under the same zero-shot protocol as in the main paper. We report three metrics: **DxCode Accuracy**, which measures ICD-code prediction accuracy against the gold labels provided by MedFound; **T-Eval Accuracy**, which applies the same claim-score-based evaluation protocol used in the main paper; and **T-Eval TrustScore**, which measures the quality of the generated diagnostic reasoning under the identical Toulmin-aligned rubric and judge setup. This design ensures strict comparability with the main MedCaseReasoning results. As shown in Tables 10 and 11, the overall ranking trend remains consistent under OOD evaluation. CGCL outperforms the tested prompting and standard SFT baselines, and remains competitive with RL-based methods. In particular, CGCL achieves the best or near-best

Category	Variant	DxCode Acc ↑	T-Eval Acc ↑	TrustScore ↑
Prompt	Vanilla	19.85	20.50	58.24
Prompt	Think & Answer	23.12	24.20	60.15
Prompt	Trust-Think & Answer	26.24	27.51	62.45
SFT	SFT-GT	24.38	25.40	62.10
SFT	SFT-CoT	25.10	26.20	63.88
SFT	SFT-CGCL(Stage3)	27.35	28.62	67.84
RL	DPO	28.15	29.50	68.90
RL	GRPO	29.01	30.35	70.76
Ours	CGCL	**29.28**	**30.64**	**70.89**

Table 10: Zero-shot OOD evaluation on MEDFOUND using Qwen2.5-3B-Instruct.

Category	Variant	DxCode Acc ↑	T-Eval Acc ↑	TrustScore ↑
Prompt	Trust-Think & Answer	27.90	28.95	63.12
SFT	SFT-CGCL(Stage3)	28.45	29.68	68.21
RL	GRPO	28.17	29.46	70.58
Ours	CGCL	**29.83**	**31.24**	**71.36**

Table 11: Zero-shot OOD evaluation on MEDFOUND using LLaMA3.2-3B-Instruct.

TrustScore across both backbones, suggesting that Toulmin-structured curriculum learning improves reasoning quality in a way that transfers beyond the training dataset.

D.9 Epistemic Calibration and Overconfidence Analysis

To examine whether CGCL improves epistemic calibration rather than merely output formatting, we compare it with SFT-CGCL(Stage3), a strong single-stage baseline trained on the same final schema and able to generate confidence qualifiers. Following our rebuttal protocol, we perform a stochastic evaluation on 50 cases with 5 runs per case and group predictions into three confidence bins (*low*, *medium*, *high*) according to the generated qualifier. For each bin, Table 12 reports the diagnosis accuracy within the bin, together with the proportion and count of samples assigned to that bin. We further define **Overconfidence Error** as the percentage of total samples that are incorrect but assigned *high confidence*. As shown in Table 12, the baseline exhibits weak confidence discrimination, with accuracy remaining relatively flat across confidence levels. In contrast, CGCL shows substantially improved calibration: accuracy increases monotonically from low to medium to high confidence, and the high-confidence bin reaches 82.2% accuracy. Moreover, CGCL reduces Overconfidence Error from 17.4% to 3.1%, indicating that its qualifier is not merely stylistic, but provides a meaningful signal of epistemic confidence.

E Prompt Details

E.1 Conventions

This appendix provides the key prompt templates used in our pipeline. We keep the prompt set compact and include only templates that (i) define the I/O contract (schema), (ii) specify the generic candidate-generation interface shared across stages, (iii) implement trajectory consolidation (fusion), and (iv) define the T-Eval judging interface. All prompts are written to be machine-parseable and to support deterministic filtering and evaluation.

Placeholders. Prompts use placeholders `CASE`, `STAGE_CONTEXT`, and `OUTPUT_FORMAT`. Unless explicitly stated otherwise, models must output only valid JSON (no additional natural language).

Component semantics. We represent a reasoning trajectory with six components {D, R, W, B, Q, Y}. D extracts objective evidence (facts only); R lists a ranked differential diagnosis (3–5 items) with brief reasons; W provides supporting evidence and clinical/pathophysiological logic for the top-ranked diagnosis; B provides ruling-out reasoning for alternatives (as a string); Q provides confidence, uncertainty, and missing information; and Y is the final diagnosis. **Field naming.** To improve readability, we use `reason` in `R` in place of `why_not`. For non-top diagnoses (rank ≥ 2), `reason` must include the key counterpoint(s) that make the diagnosis less likely, preserving the original semantics. **Revision encoding.** Because `Q` contains only {`confidence`, `uncertainty`, `missing_info`}, any evidence-based revision is

Method	Overall Acc. (↑)	TrustScore (↑)	Confidence Bins [Acc. (Prop., n)] Low	Med	High	Overconf. Error (↓)
Baseline (SFT)	32.3	66.8	28.5% (32.8%, n=82)	33.4% (40.4%, n=101)	35.2% (26.8%, n=67)	17.4
CGCL (Ours)	**37.3**	**72.9**	14.3% (44.4%, n=111)	43.8% (38.4%, n=96)	**82.2%** (17.2%, n=43)	**3.1**

Table 12: Epistemic calibration analysis on Qwen2.5-3B-Instruct.

encoded inside `Q.uncertainty` by prefixing a marker and a brief rationale: `[Evidence-Based Revision] Initial hypothesis: ... Pivot evidence: ... Therefore revise to: ....`

E.2 Unified Output Schema (I/O Contract)

The schema below defines the unified JSON format used throughout the paper. When a prompt requests only a subset of fields, the model is required to output only those fields and omit all others.

Unified JSON schema (required output format)

Return a single JSON object. Output ONLY JSON (no extra text).
Schema:

```
{
  "D": ["...","..."],
  "R": [{"dx":"...","rank":1,"reason":"..."}],
  "W": "...",
  "B": "...",
  "Q": {
    "confidence": "low|medium|high",
    "uncertainty": ["string"],
    "missing_info": ["string"]
  },
  "Y": "FINAL_DIAGNOSIS"
}
```

Rules:

1. Do NOT fabricate evidence or citations.
2. If a prompt requests only specific field(s), output ONLY those field(s) and omit all others.

E.3 Stage-wise Candidate Generation

The prompt template below defines the generic interface for stage-wise candidate generation across different curriculum stages: Stage 1 populates D and R; Stage 2 populates W and B conditioned on Stage-1 context; Stage 3 populates Q and Y conditioned on Stage-2 context. We enforce strict JSON-only outputs and disallow unsupported additions to enable reliable parsing and filtering.

Stage instantiations. We instantiate as follows. **Stage 1 (D, R).** `OUTPUT_FORMAT` requests only `D` and `R`. D lists objective facts only. R lists 3–5 diagnoses ranked by plausibility given D; for rank ≥ 2, `reason` must include the key counterpoint(s). **Stage 2 (W, B).** `OUTPUT_FORMAT` requests only `W` and `B`. W provides supporting evidence and pathophysiological/clinical logic for the top diagnosis (R rank=1). B is a *string* that rules out alternatives in R (rank ≥ 2), referencing missing/contradictory evidence, without adding new facts. **Stage 3 (Q, Y).** `OUTPUT_FORMAT` requests only `Q` and `Y`. Y is the final diagnosis. Q provides calibrated confidence, uncertainty, and missing information. If the final diagnosis differs from the initial top candidate in R, the model must encode an evidence-based revision inside `Q.uncertainty` using the required marker and rationale format.

Generic stage-wise candidate generation template

You are a careful clinician. Follow the output format strictly. You will be given a clinical case and (optionally) a stage context. Generate ONLY the requested component(s).
Requested field(s): {TARGET_FIELDS}
Rules:

1. Output MUST be valid JSON and MUST contain ONLY the requested field(s).
2. Do NOT add any other keys.
3. Do NOT invent evidence or diagnoses not supported by the case/context.
4. Keep the output concise, factual, and clinically grounded.
5. Use double quotes for all strings and keys.

Case:
`{CASE}`
Stage context (may be empty):
`{STAGE_CONTEXT}`
Output format:
`{OUTPUT_FORMAT}`

E.4 Fusion Prompt

Fusion consolidates selected best candidates into a single coherent trajectory under the fixed schema. The fusion prompt below defines the consolidation procedure for merging selected candidates into a single coherent trajectory. It also enforces implementation-critical typing constraints (B is a string; Q contains only three fields).

Fusion prompt (trajectory consolidation)

You are an expert clinical writer. Produce a coherent structured argument. Merge the selected components into ONE consistent JSON object with keys: D, R, W, B, Q, Y.
Hard constraints:

1. Do NOT add new evidence beyond the provided D list.
2. Do NOT introduce any diagnosis not already present in the provided R list.
3. The final claim Y must be consistent with D and the reasoning (W, B).
4. Do NOT fabricate citations or new facts not supported by the case.
5. Output ONLY valid JSON. No extra text.

Typing constraints (must hold exactly):

- "B" MUST be a string.
- "Q" MUST contain ONLY {"confidence", "uncertainty", "missing_info"}.

Revision encoding (no extra fields allowed):

If Y differs from the top diagnosis in R (rank = 1), Q.uncertainty MUST begin with:

"[Evidence-Based Revision] Initial hypothesis: ... Pivot evidence: ... Therefore revise to: ..."

Selected components:

D: {D_STAR}
R: {R_STAR}
W: {W_STAR}
B: {B_STAR}
Q: {Q_STAR}
Y: {Y_STAR}

Return exactly one JSON object with keys D, R, W, B, Q, Y.

E.5 T-Eval Judge Prompt

T-Eval judges receive the case P and a candidate trajectory C and output a rubric-guided 1–5 score. To ensure consistent parsing, we enforce a strict two-line output format. The T-Eval judge prompt below provides a representative rubric-guided interface for evaluating diagnostic reasoning quality.

T-Eval judge prompt

You are an expert evaluator of medical reasoning quality. Your task is to rigorously assess the AI model's diagnostic output against the standard diagnosis.
Evaluation Criteria
1. Static Structure Assessment (Toulmin-style) — Score 1.0–5.0

- **data_score**: Are all key facts correctly extracted without errors or omissions?
- **warrant_score**: Is the chain from data to hypothesis clear, sound, and medically valid?
- **backing_score**: Are cited guidelines or medical knowledge accurate and relevant?
- **rebuttal_score**: Are the major alternative diagnoses addressed with specific reasoning for exclusion?
- **qualifier_score**: Does the output appropriately calibrate diagnostic confidence, uncertainty, and missing information?
- **claim_correct**: is a 1.0–5.0 rating of whether the final diagnosis (<answer>) matches the standard diagnosis (A): 5.0 = exact match; 4.0 = near-synonym/variant; 3.0 = partially correct; 2.0 = mostly incorrect; 1.0 = incorrect.

Output Format
Return a strict JSON object only, with no extra text or commentary.

```
{"data_score": 0.0,
"warrant_score": 0.0,
"backing_score": 0.0,
"rebuttal_score": 0.0,
"qualifier_score": 0.0,
"claim_correct": 0.0,
"overall_analysis": "..."}
```

Objects to Evaluate
Standard Diagnosis (A): {A}
Model Output: {model_output}
Begin evaluation. Output valid JSON only.