Skill-Level Distillation: Self-Evolving Non-Parametric Reasoning Transfer for Medical Question Answering

Anonymous ACL submission

Abstract

Knowledge distillation (KD) has emerged as a mainstream paradigm for transferring capabilities from Large Language Models (LLMs) to smaller models. However, existing methods predominantly rely on imitating teacher outputs or instance-level reasoning trajectories. This binds capability transfer to specific samples and parameter updates, failing to cultivate reusable, interpretable, and continually updatable reasoning skills. In this paper, we propose Skill-Level Distillation (SLD), a novel non-parametric reasoning transfer framework. Rather than mimicking raw teacher outputs, SLD transfers de-instantiated, structured reasoning skills. We decompose medical reasoning into two reusable cognitive operators: Clinical Information Extraction and Integration (CIEI) and Discriminative Decision Reasoning (DDR). Through a teacher-guided reflection-correction loop, SLD distills student reasoning failures into reusable skills, which are accumulated in an explicit skill memory. During inference, a parameter-frozen 8B student model dynamically retrieves these skills to guide its reasoning. Empirical evaluations on the MedQA and PubMedQA benchmarks demonstrate that, without any student-parameter updates, SLD significantly narrows the performance gap between the frozen 8B student and GPT-4o, reducing the residual accuracy gaps to merely 2.44 and 0.80 percentage points, respectively. This framework provides a novel and dynamically evolvable technical pathway for the deployment of medical LLMs.

1 Introduction

Large Language Models (LLMs) have recently demonstrated strong reasoning capabilities in medical knowledge Q&A, clinical inference, and exam-like medical tasks. Prior studies such as Multi-MedQA, Med-PaLM, and Med-PaLM 2 illustrate that large models can achieve near-expert performance on multiple medical benchmarks, while

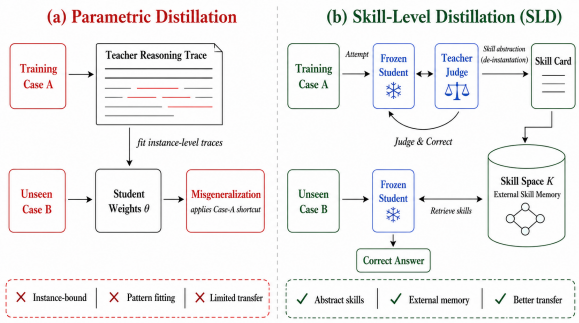

Figure 1: The paradigm shift from traditional supervised fine-tuning-based KD to Skill-Level Distillation (SLD). Left: traditional KD transfers sample-bound teacher behavior into student parameters. Right: SLD freezes student parameters and uses teacher-guided correction to construct reusable skills that can be retrieved for new queries.

models such as GPT-4 exhibit strong zero-shot clinical reasoning on USMLE-style and other challenging medical tasks (Singhal et al., 2023, 2025; Nori et al., 2023). However, such capability does not directly transfer to real clinical environments. For example, loading the FP16 weights of the GPT-3 175B model alone requires roughly 350GB of GPU memory, and inference often demands high-end accelerators or complex offloading strategies (Brown et al., 2020; Sheng et al., 2023). For hospitals, clinics, and bedside systems, this requirement conflicts with low-latency, high-throughput, privacy-controllable, cost-acceptable, and maintainable deployment constraints (Dennstädt et al., 2025). This creates a central deployment tension: clinical reasoning benefits from the complex capabilities of large models, while real-world deployment often favors smaller and more controllable models.

Knowledge Distillation (KD) thus emerges as a natural pathway for compressing LLM capability. Classic KD transfers knowledge by aligning student model outputs with teacher model distributions (Hinton et al., 2015); later, Transformer-era

distillation extends this objective to hidden states, attention maps, and intermediate representations (Jiao et al., 2020). With growing interest in LLM reasoning, researchers have also explored rationale distillation, Chain-of-Thought (CoT) distillation, and multi-step supervision rather than only distilling final answers (Magister et al., 2022; Fu et al., 2023; Hsieh et al., 2023). These methods improve the reasoning expressiveness of smaller models, but they still largely assume that the student should acquire teacher behavior by fitting outputs, rationales, or trajectories at the instance level.

This assumption constrains KD in complex reasoning tasks. Even when the distillation target is extended from logits to natural-language rationales or full CoT traces, the student often learns example-level reasoning patterns rather than cross-question reusable reasoning operations. Moreover, CoT text is not always faithful to the model's actual decision process; a superficially coherent explanation may be a post-hoc rationalization (Turpin et al., 2023). Recent work also shows that CoT distillation can inherit spurious question–answer correlations from training data, leading to out-of-distribution failures (Dai et al., 2024). In short, traditional KD compresses teacher statistics into student parameters, but this does not necessarily yield stable, interpretable, and reusable solution strategies.

A complementary line of work offers a different perspective. ReAct interleaves reasoning and external action to improve task-solving, Toolformer shows that models can learn to use tools, and Reflexion demonstrates that linguistic feedback and external memory can improve reasoning without updating core parameters (Yao et al., 2023; Schick et al., 2023; Shinn et al., 2023). These studies suggest that reasoning capability can be carried not only by model weights, but also by structured intermediate representations, tools, and memory traces. This perspective is particularly relevant to medicine, where reasoning patterns and clinical recommendations may need to be updated after deployment.

Based on these observations, we propose Skill-Level Distillation (SLD), a reasoning augmentation framework that freezes student model parameters and supports the continued evolution of reasoning capabilities. As contrasted in Figure 1, the core idea is that clinically transferable reasoning should not be bound to specific input–output instances, but should operate through intermediate skills reusable across examples. Instead of directly imitating the teacher's final answer or full CoT for a single example, SLD treats training instances as *diagnostic probes*: the student first attempts to solve the problem, thereby exposing where evidence extraction, information integration, or candidate discrimination fails. The teacher then diagnoses these failures and converts its corrective feedback into structured abstract skills. In this way, the transfer target becomes not "how the teacher answered this instance", but "how the student should revise reasoning in similar tasks."

In medical Q&A, we operationalize SLD into two core skill spaces: *Clinical Information Extraction and Integration* (CIEI) and *Discriminative Decision Reasoning* (DDR). CIEI supports the extraction, normalization, and integration of clinically relevant evidence from questions, while DDR supports candidate comparison, error elimination, and discriminative answer selection. Through a student-trial–teacher-diagnosis loop, SLD transforms accumulated instance-level failures into reusable skill modules that are dynamically invoked by the parameter-frozen student model. Crucially, SLD does not store teacher answers or raw CoT traces; it abstracts reusable correction patterns from student failures, allowing new skills to be added, corrected, or merged without retraining the base student model.

Experiments on MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019) show that an SLD-enhanced frozen 8B student achieves 82.40% and 77.20% accuracy, respectively. Compared with the frozen Qwen3-8B backbone, SLD yields substantial gains of 15.55 and 14.40 accuracy points on MedQA and PubMedQA, respectively. When placed alongside larger direct-inference open-model references, SLD also achieves competitive performance, suggesting that teacher-guided skill memory can substantially improve a small frozen student without parameter updates. Our main contributions are:

- We propose Skill-Level Distillation (SLD), a non-parametric reasoning transfer framework that freezes student parameters and transfers reusable reasoning skills through an external memory.
- We design an error-driven skill construction mechanism. An initially empty skill memory grows through a student-trial–teacher-diagnosis loop without retraining the base model.

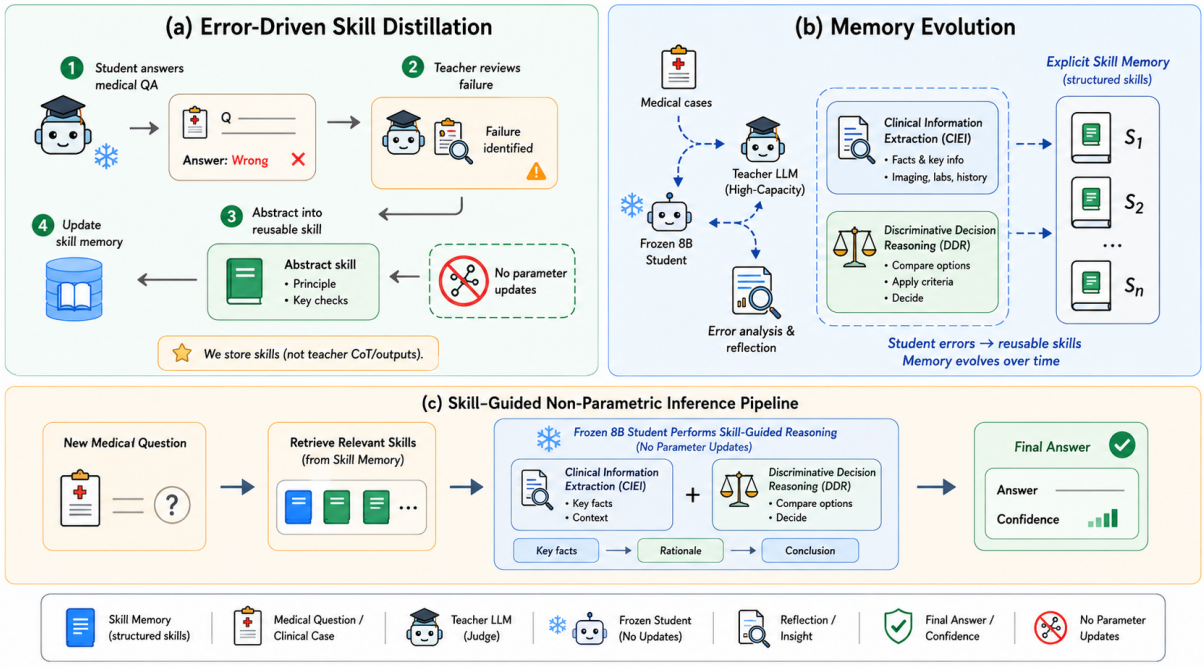

Figure 2: The overall framework of Skill-Level Distillation (SLD). (a) Error-Driven Skill Distillation: the teacher model diagnoses reasoning failures exposed by the frozen student on training instances. (b) Memory Evolution: corrective feedback is de-instantiated from specific samples into reusable cognitive operators, which are accumulated or merged into evolving CIEI and DDR skill spaces. (c) Skill-Guided Non-Parametric Inference: for new queries, the frozen student retrieves relevant skills and uses them as reasoning guidance.

- We decompose medical reasoning into CIEI and DDR, making evidence extraction, information integration, candidate comparison, and decision-making more modular.

- We empirically validate SLD on the MedQA and PubMedQA benchmarks. We show that without updating student parameters, our framework substantially bridges the capability gap between a frozen 8B student and GPT-4o, bringing the accuracy difference down to 2.44 and 0.80 percentage points, respectively.

2 Related Work

Recent research on medical LLMs demonstrates that foundation models with massive parameter scales exhibit strong performance on medical question answering and clinical reasoning tasks. MultiMedQA established a comprehensive evaluation framework covering professional exams, research queries, and consumer health inquiries, while Med-PaLM 2 further advanced benchmark results and improved physician preference for long-form answers (Singhal et al., 2023, 2025). Meanwhile, general-purpose frontier models such as GPT-4 have demonstrated strong zero-shot reasoning capabilities on challenging medical problems (Nori et al., 2023). To elicit these reasoning capabilities, researchers have explored structured prompting and reasoning methods for medical QA (Lucas et al., 2024). However, existing work focuses primarily on improving the upper bound of large models, with less attention to how clinically valuable reasoning skills can be transferred to resource-constrained lightweight systems.

To address the deployment challenges of LLMs, KD has emerged as a mainstream research direction. Classical KD transfers knowledge by training students to fit teacher output distributions, while Transformer-era distillation extends the objective to hidden states, attention maps, and intermediate representations (Hinton et al., 2015; Jiao et al., 2020). Recent studies further explore rationale distillation, CoT distillation, and step-by-step supervision (Magister et al., 2022; Fu et al., 2023; Hsieh et al., 2023; Dai et al., 2024). Nevertheless, these methods still largely compress complex reasoning into the student parameters and rely on sample-level trajectories. This parametric route suffers from limitations: CoT explanations are not always faithful to model decisions, and CoT distillation can remain vulnerable to OOD generalization chal-

lenges (Turpin et al., 2023; Dai et al., 2024).

Research on agents and non-parametric reasoning provides a complementary perspective: reasoning capabilities can be externalized rather than necessarily internalized within model weights. ReAct interleaves reasoning with external actions, Toolformer allows models to decide when to use tools, and Reflexion uses linguistic feedback and episodic memory to improve decisions without updating underlying parameters (Yao et al., 2023; Schick et al., 2023; Shinn et al., 2023). In medicine, knowledge and clinical recommendations also evolve over time, creating a need for updateable reasoning mechanisms (Pielenz et al., 2022; Dennstädt et al., 2025). These observations motivate SLD: instead of distilling complete final answers or sample-bound CoT trajectories into small-model weights, we use student-exposed capability gaps to distill teacher corrections into explicit, invokable, and dynamically updatable skills.

3 Method

This section introduces the proposed Skill-Level Distillation (SLD) framework. Unlike conventional KD, which updates student parameters to imitate teacher outputs, SLD keeps the student parameters frozen and transforms training into an error-driven skill construction procedure. The student first attempts to solve each training example. If the answer is incorrect or the reasoning is insufficient, the teacher diagnoses the failure and provides corrective feedback, which is distilled into reusable reasoning skills. As training progresses, the CIEI and DDR skill spaces evolve from empty to functional skill memories. At inference time, the frozen student performs skill-guided reasoning with the fixed skill memory.

3.1 Overall Framework and Problem Definition

Let the training dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a medical question and y_i is the corresponding answer. Denote the student model as S_θ and the teacher model as T. Traditional KD optimizes student parameters to fit teacher outputs, reasoning traces, or intermediate representations, effectively compressing teacher knowledge into student parameters (Hinton et al., 2015; Jiao et al., 2020; Hsieh et al., 2023).

SLD follows a different transfer paradigm. The training data is used not to update student weights, but to expose deficiencies in the student's reasoning; the teacher acts as a critic and corrector, transforming student failures into reusable skills. Throughout skill distillation, student parameters remain frozen:

$$\theta^{(t+1)} = \theta^{(t)}. \tag{1}$$

The actual evolution occurs in the external skill space. Denote the skill space at step t as $\mathcal{K}^{(t)}$. The update process can be abstracted as:

$$\mathcal{K}^{(t+1)} = \mathrm{Update}\left(\mathcal{K}^{(t)}, \Delta k_i\right), \tag{2}$$

where Δk_i is the skill update distilled from teacher feedback on the student's performance for sample x_i. Thus, the core of SLD is not optimizing student parameters, but converting student errors into reusable skill knowledge.

As illustrated in Figure 2, SLD consists of two stages: skill distillation and skill-guided inference. In the distillation stage, the CIEI and DDR skill spaces are initially empty. For each training sample, the student first produces an answer using the current skills. If the answer is incorrect or the reasoning is insufficient, the teacher evaluates the student's output and reasoning, then provides corrective feedback. This feedback is de-instantiated into reusable skills and written into the corresponding skill space.

In the inference stage, the student parameters remain frozen and the skill memory is fixed. For a new medical question, the student dynamically retrieves relevant skills from the CIEI and DDR spaces, performs clinical evidence extraction and integration, and conducts candidate comparison and discrimination.

This design differentiates SLD from three common approaches. First, it differs from parameter-level KD because student weights remain frozen. Second, it differs from ordinary retrieval-augmented generation (RAG) or example retrieval, as the skill space stores reasoning operations distilled from teacher corrections rather than raw samples, historical answers, or external knowledge snippets (Lewis et al., 2020). Third, it differs from CoT distillation because SLD focuses on cross-sample reusable error-correction strategies rather than instance-level full reasoning texts.

3.2 Skill Spaces: CIEI and DDR

In SLD, CIEI and DDR are not hand-crafted fixed rules but functional skill spaces designed for medical reasoning. They start empty and evolve as

the student model exposes errors on training examples. The student first attempts to answer; if the answer is incorrect or the reasoning is insufficient, the teacher diagnoses the error and distills corrective feedback into skills that are reusable across similar questions.

CIEI (*Clinical Information Extraction and Integration*) primarily contains skills for processing clinical evidence. Many medical QA errors are not due to a complete lack of medical knowledge, but because the model fails to identify key clues or fails to integrate symptoms, signs, history, and test results. CIEI helps the student determine "what information is important" and organize dispersed medical clues into structured evidence. For example, if the student overlooks a key lab result, confuses chief complaints with distractors, or fails to link multiple symptoms to the same pathological mechanism, the teacher distills such corrections into CIEI skills.

DDR (*Discriminative Decision Reasoning*) primarily contains skills for candidate comparison, error elimination, and discriminative decision reasoning. Medical questions often contain multiple superficially plausible answers. DDR helps the student judge "why this option is more reasonable than others" and avoid traps in distractors. For example, if the student chooses a common but incomplete diagnosis or fails to exclude options that conflict with key evidence, the teacher converts this correction into DDR skills.

CIEI and DDR correspond to two critical stages of medical reasoning: the former focuses on discovering, extracting, and integrating evidence, while the latter focuses on comparing, eliminating, and selecting among candidates. Because the skills are distilled from student errors and teacher corrections rather than manually enumerated, both CIEI and DDR evolve as training progresses. This design avoids solidifying reasoning ability into static prompts. CIEI and DDR are not fixed instructions or hand-crafted medical rules; they are error-driven, evolving skill spaces that guide the student to improve reasoning on similar problems.

3.3 Error-Driven Skill Distillation and Memory Evolution

The core of SLD is error-driven skill distillation. Unlike conventional training, the dataset does not provide gradient signals; each sample instead exposes gaps in the student's reasoning. If the student answers correctly, the skill space remains unchanged; otherwise, the teacher provides corrective feedback.

Formally, given a training sample (x_i, y_i), the student first generates an answer and reasoning process using current skills:

$$(\hat{y}_i, r_i) = S_\theta(x_i; \mathcal{K}^{(t)}), \tag{3}$$

where $\hat{y}_i$ is the student's predicted answer and r_i is the reasoning trace. The teacher evaluates:

$$j_i = T_{\text{judge}}(x_i, y_i, \hat{y}_i, r_i), \tag{4}$$

where j_i indicates the quality of the student's answer and reasoning. If the answer is incorrect or the reasoning is insufficient, the teacher produces corrective feedback, which is then de-instantiated into reusable skills:

$$\Delta k_i = T_{\text{skill}}(x_i, y_i, \hat{y}_i, r_i, j_i). \tag{5}$$

Δk_i abstracts reusable reasoning operations from the student's failure and teacher correction, such as evidence extraction strategies, integration methods, candidate comparison principles, or ways to avoid common medical pitfalls.

The skill space is updated via:

$$\mathcal{K}^{(t+1)} = \begin{cases} \text{InsertOrMerge}\left(\mathcal{K}^{(t)}, \Delta k_i\right), \\ \quad \text{if } j_i \text{ indicates failure}, \\ \mathcal{K}^{(t)}, \quad \text{otherwise}. \end{cases} \tag{6}$$

New skills are added if they correspond to novel failure patterns; similar skills are merged or refined. This prevents uncontrolled growth and ensures that the skill memory evolves compactly and reusably.

At inference, the student first retrieves relevant skills for a new question x:

$$\mathcal{K}_x = \text{Retrieve}(x, \mathcal{K}^{(*)}), \tag{7}$$

where $\mathcal{K}^{(*)}$ is the skill space after distillation. It then generates an answer and reasoning trace:

$$(\tilde{y}, \tilde{r}) = S_\theta(x; \mathcal{K}_x). \tag{8}$$

In summary, SLD transforms distillation from "teacher outputs $\rightarrow$ student parameters" into a closed loop of "student errors $\rightarrow$ teacher corrections $\rightarrow$ skill evolution." Student errors expose capability gaps, teacher diagnostics supervise de-instantiated skill construction, and the explicit skill space enables frozen students to perform complex reasoning. This mechanism transfers sophisticated medical reasoning in a non-parametric manner with respect to the student weights, while maintaining interpretability and update flexibility.

4 Experiments

4.1 Datasets and Metrics

We evaluate SLD on MedQA and PubMedQA (Jin et al., 2021, 2019). We use the English USMLE subset of MedQA. PubMedQA is treated as a three-way classification task over yes/no/maybe. We report accuracy as the primary metric for both datasets.

Consistent with this benchmark-based evaluation setting, we use MedQA and PubMedQA only as research benchmarks for medical and biomedical question answering evaluation. We do not redistribute the original datasets, model weights, or derived benchmark data. The model artifacts used in this work, including Qwen-family models and GPT-4o, are used only for research evaluation and skill construction under their respective access conditions. Since our experiments are conducted on existing public benchmark datasets, we do not collect new personal data.

4.2 Baselines and Comparison Scope

We include GPT-4o as a strong proprietary reference model(OpenAI, 2024). We also compare with direct-inference open-model baselines, including Qwen3-8B, Qwen3-14B(Qwen Team, 2025), and Qwen2.5-72B(Yang et al., 2025). In addition, we include Query-Efficient DKS as an 8B teacher-assisted transfer reference(Li et al., 2026).

4.3 Implementation Details

Unless otherwise stated, the student model is Qwen3-8B (Qwen Team, 2025) with frozen parameters, and the teacher model is GPT-4o (OpenAI, 2024). Skills are stored in an external vector database and retrieved separately from the CIEI and DDR memories. The student parameters are not updated during either skill construction or inference.

Figure 3 summarizes the skill-memory construction process. Out of 11,450 training samples, the frozen student answers 7,364 samples correctly on the first attempt, while 4,086 samples trigger teacher correction. The failed samples are further grouped by medical system or question type and routed to three construction outcomes: CIEI skill writing, DDR skill writing, or no accepted skill write. Among the failed cases, 2,262 lead to CIEI skill writes, 1,145 lead to DDR skill writes, and 679 remain unrescued. This distribution shows that, under our construction procedure, evidence extraction and integration corrections are invoked more frequently than candidate-discrimination corrections. The rightmost nodes further distinguish whether each successful write inserts a new skill or merges/updates an existing one, indicating that the memory evolves through both expansion and consolidation.

Group	Model	MedQA	PubMedQA
Proprietary	GPT-4o	84.84	78.00
Direct	Qwen2.5-72B	76.75	73.74
Direct	Qwen3-14B	74.71	63.80
Direct	Qwen3-8B	66.85	62.80
Teacher-assisted	Query-Efficient DKS (8B)	62.00	74.90
Teacher-assisted	**SLD (8B)**	**82.40**	**77.20**

Table 1: Main results on MedQA and PubMedQA.

4.4 Main Results

Table 1 reports the main results. On MedQA, SLD improves the 8B backbone from 66.85% to 82.40% (+15.55), leaving a 2.44-point gap to GPT-4o. On PubMedQA, SLD improves the same backbone from 62.80% to 77.20% (+14.40), narrowing the gap to GPT-4o to 0.80 points. SLD also numerically outperforms the direct-inference Qwen2.5-72B reference by 5.65 points on MedQA and 3.46 points on PubMedQA.

The improvement over Qwen3-8B indicates that reusable external skills can provide substantial reasoning guidance to a frozen student model. The small remaining gap to GPT-4o further suggests that a significant portion of the teacher's medical reasoning behavior can be amortized into an explicit skill memory.

Figure 4 provides a qualitative example of how a distilled skill is reused across instances. In the training case, the frozen student selects an incorrect option because it fails to integrate shortness of breath, pulmonary findings, and the left atrial mass into the correct cardiac-reasoning frame. After teacher correction, SLD stores this failure pattern as a CIEI skill rather than as the original answer. At test time, a different question retrieves the same skill: the case contains worsening cough, dyspnea, elevated jugular venous pressure, decreased breath sounds at the lung bases, lower-extremity edema, and an enlarged cardiac silhouette. Guided by the retrieved skill, the student identifies the need to evaluate the underlying cardiac structure and selects echocardiography. Notably, the training case asks for an inflammatory mediator associated with a cardiac mass, whereas the test case asks for the

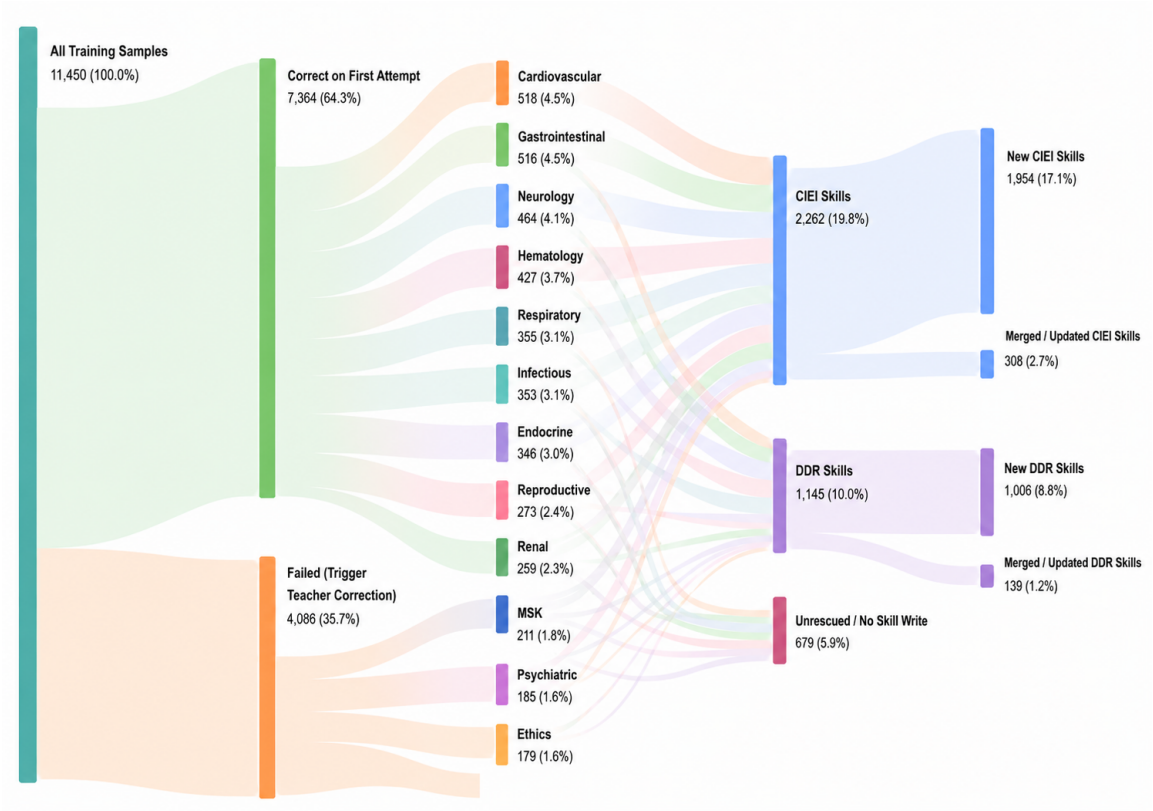

Figure 3: Construction flow of the SLD skill memory. Among 11,450 training samples, the frozen student correctly answers 7,364 samples on the first attempt, while 4,086 samples trigger teacher correction. The failed samples are grouped by medical system or disease type and then distilled into CIEI skills, DDR skills, or unrescued cases. The rightmost nodes distinguish newly inserted skills from merged or updated existing skills.

next diagnostic step in a heart-failure-like presentation. This answer-type shift indicates that the retrieved memory is not a memorized answer or a single medical fact, but a reusable CIEI operator that helps the student organize dispersed cardiopulmonary evidence into a cardiac reasoning frame before mapping it to the required question target.

4.5 Ablation Study

We report ablations on MedQA only. Table 2 evaluates the two components that are most central to SLD: error-conditioned skill induction and memory evolution. Removing error-conditioned skill induction disables the process that converts student failures into reusable CIEI/DDR skills, causing performance to drop from 82.40% to 65.99% (-16.41). This confirms that a useful skill memory requires targeted construction from student-exposed failures. The result should not be interpreted as the teacher directly solving test examples, because the teacher is not queried during inference in the full model.

Removing memory evolution yields 80.91% (-1.49). This setting measures the additional benefit of maintaining an evolving cross-instance memory after skills have been induced. The smaller drop suggests that once high-quality skills are constructed, a static skill set already provides strong guidance, while insertion/merging/update further improves reuse and consolidation. Overall, the ablations support the intended mechanism: teacher judgment supplies skill-construction supervision, de-instantiated skills form the transferred artifact, and memory evolution provides additional cross-instance reuse.

Variant	MedQA
w/o Memory Evolution	80.91
w/o Error-Conditioned Skill Induction	65.99
Full SLD	**82.40**

Table 2: Ablation results on MedQA.

5 Conclusion

We propose Skill-Level Distillation (SLD), a non-parametric framework for transferring medical reasoning capabilities from a large teacher model to

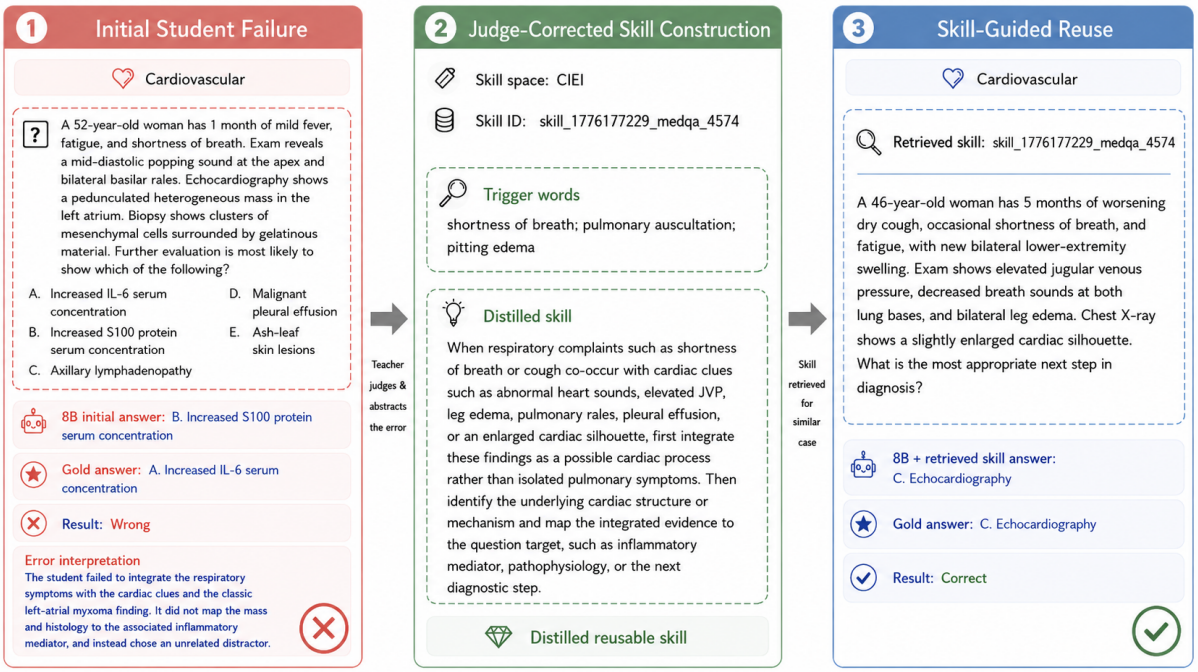

Figure 4: Qualitative example of train-to-test CIEI skill reuse. The frozen 8B student first fails on a training case because it does not integrate respiratory symptoms, pulmonary findings, and cardiac evidence into a unified cardiac-reasoning frame. The teacher correction is abstracted into a CIEI skill that instructs the student to integrate dispersed cardiopulmonary clues before selecting the question-specific target. On a held-out test case with different surface form and a different answer target, the student retrieves the skill and correctly selects echocardiography. This example illustrates that SLD stores reusable reasoning operations rather than memorized answers.

a frozen 8B student model. Unlike traditional methods that distill teacher outputs or instance-level CoT trajectories into model parameters, SLD leverages a student-trial–teacher-diagnosis loop to compile student failures into reusable and continuously evolving skill patterns. These skills are organized around two operators: CIEI for evidence extraction and integration, and DDR for discriminative decision-making. Experiments on MedQA and PubMedQA show that SLD improves Qwen3-8B by 15.55 and 14.40 accuracy points, respectively, and reduces the gap to GPT-4o without requiring student-parameter updates. More broadly, our results suggest that medical reasoning transfer is not exclusively bound to parameter-level imitation; teacher supervision can be amortized into an explicit, reusable, and updatable skill memory, providing a flexible deployment pathway for lightweight medical LLMs.

Limitations

This work has several limitations. First, our evaluation is currently limited to MedQA and PubMedQA. These benchmarks cover exam-style and biomedical question answering, but they do not fully represent the diversity of real clinical reasoning scenarios, such as multi-turn consultation, temporal patient records, or institution-specific clinical workflows. Future work should evaluate SLD on more diverse medical benchmarks and clinical task formats.

Second, the current skill memory bank is constructed from model errors collected on training data. Accordingly, its quality and coverage are closely correlated with the distribution of observed failures. There is still room for improvement regarding rare reasoning paths and underrepresented medical specialties.

Third, although SLD de-instantiates teacher feedback into reusable skills rather than storing raw answers or CoT traces, future work should provide larger-scale leakage audits and cross-dataset transfer analyses to further quantify how much each skill generalizes beyond its source failure pattern.

Ultimately, while our explicit skill memory mechanism effectively enhances the interpretability of the reasoning guidance process, discrepancies in skill retrieval may still occur during practical model operations. Consequently, when applying this framework to clinical decision-making, it

must be deployed under the supervision of qualified medical professionals, complemented by rigorous model calibration and safety assessments.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024. Improve student's reasoning generalizability through cascading decomposed CoTs distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15643, Miami, Florida, USA. Association for Computational Linguistics.

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):143.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Z. Li, X. Yuan, B. Shen, K. Le, H. Wang, X. Zhou, S. Gao, and Y. Dong. 2026. Query-efficient domain knowledge stealing against large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mary M. Lucas, Justin Yang, Jon K. Pomeroy, and Christopher C. Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *Preprint*, arXiv:2212.08410.

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*.

OpenAI. 2024. GPT-4o system card. `https://openai.com/index/gpt-4o-system-card/`.

Charline Pielenz, Marco Schneider, Eva Salveridou-Hof, Marisa Flick, Gabriele Gaigl, Naiiri Khorikian-Ghazari, Duygu Güler, Theresa Halms, Angelika Kapfhammer, Carolin Lorenz, Astrid Röh, Elias Wagner, Peter Falkai, Stefan Leucht, Wolfgang Gaebel, and Alkomiet Hasan. 2022. From conventional to living guidelines–faster updates for better informed guidance? a scoping review. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 174:20–31.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. FlexGen: High-throughput generative inference of large language models with a single GPU.

In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31094–31116. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Karan Singhal, Shekoofeh Azizi, Tao Tu, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

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

An Yang and 1 others. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.