Unlocking Latent Medical Reasoning in LLMs via Inference-Time Representation and Prefix Interventions

Anonymous ACL submission

Abstract

Large language models (LLMs) have acquired substantial medical knowledge and reasoning skills during pretraining, yet eliciting these capabilities typically requires scarce rationales or costly post-training. To address this gap, we introduce **RIPT**, a lightweight framework for data-efficient endogenous capability activation. It explicitly decomposes this objective into two complementary components: reasoning enhancement and medical knowledge elicitation. For the former, we extract steering vectors from hidden activations on a small set of high-quality paired reasoning and vanilla responses to shape LLMs' reasoning behavior. For the latter, we obtain prefix vectors via prefix tuning on simple medical QA pairs to elicit domain-specific knowledge. At inference, RIPT freezes the backbone LLM and combines steering-vector injection with prefix-vector conditioning. Rather than injecting external knowledge, RIPT aims to unlock medical reasoning capabilities already acquired by LLMs during pretraining. Experiments under limited-resource settings show that RIPT consistently improves diverse backbone LLMs and benchmarks, suggesting a generalizable pathway for lightweight medical reasoning enhancement.

1 Introduction

Recent advances in reasoning LLMs, such as OpenAI's o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and medical variants like Baichuan-M2 (Dou et al., 2025), have substantially improved complex reasoning performance. The Superficial Alignment Hypothesis (Zhou et al., 2023) shows that many high-level capabilities are largely acquired during pretraining and later elicited through alignment. This suggests that LLMs may have already obtained latent medical knowledge and reasoning skills, especially when exposed to large-scale domain-relevant corpora during pretraining.

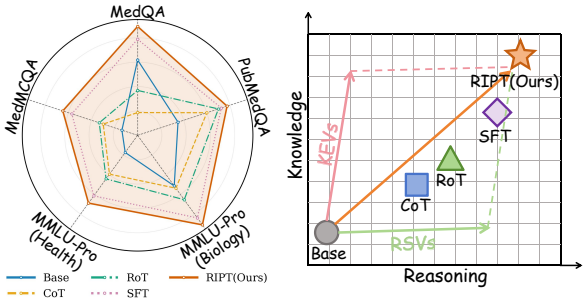

Figure 1: Overall performance of RIPT. **Left:** The performance on benchmarks after normalization. **Right:** The performance on Reasoning and Knowledge dimension. **KEVs** represent Knowledge Elicitation Vectors, while **RSVs** represent Reasoning Steering Vectors.

However, how to elicit these latent capabilities in a data-efficient manner remains a practical challenge.

This challenge is particularly pronounced in medicine, where high-quality supervision is scarce, costly, and subject to privacy and governance constraints (Malin et al., 2018). Although post-training paradigms such as supervised fine-tuning (SFT) (Singhal et al., 2023) and reinforcement learning (RL) (Gu et al., 2025) can improve medical reasoning, they require substantial computation and high-quality rationales. Retrieval-based Augmentation (RAG) can introduce external medical knowledge at inference time (Lewis et al., 2020; Gao et al., 2023), but it remains dependent on external knowledge bases and retrieval infrastructure, while supplementing external evidence rather than activating the model's endogenous medical reasoning capabilities. Together, these limitations motivate a central question: *Can latent medical reasoning be unlocked at inference time through data-efficient endogenous activation, without costly post-training or external knowledge injection?*

Prior work suggests that medical reasoning requires both logical inference under uncertainty and grounding in specialized domain knowledge

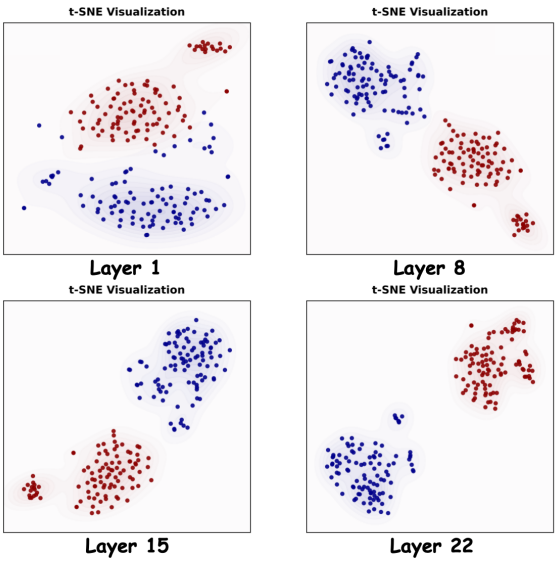

Figure 2: T-SNE visualization of reasoning (red) and vanilla (blue) response representations across different layers on Qwen2.5-7B-Instruct. The two clusters become more separable at deeper layers.

(Berger et al., 2025). Motivated by this view, we propose **R**epresentation **I**njection with **P**refix **T**uning (**RIPT**), a lightweight framework that decomposes endogenous capability activation into two complementary objectives: reasoning enhancement and medical knowledge elicitation.

For reasoning enhancement, RIPT adopts a representation engineering perspective (Zou et al., 2023). As illustrated in Figure 2, reasoning-oriented and vanilla response patterns exhibit separable structures in the latent space, allowing us to derive **Reasoning Steering Vectors** from contrastive activations on a small reasoning-dense dataset. For medical knowledge elicitation, RIPT learns **Knowledge Elicitation Vectors** through prefix tuning on simple medical QA pairs, which modulates attention to expose domain-specific knowledge while reducing unintended degradation (Petrov et al., 2023; Lobo et al., 2025). At inference time, RIPT keeps the backbone LLM frozen and combines steering-vector intervention with prefix-vector conditioning, enabling data-efficient activation of latent medical reasoning capabilities.

In our experiments, RIPT requires only 100 curated reasoning examples and 3,000 easily accessible QA pairs, demonstrating strong data efficiency. As shown in Figure 1, RIPT yields consistent improvements across benchmarks, with gains along both reasoning and knowledge that validate the effectiveness of our decomposed activation strategy. Further analyses reveal that reasoning steering exhibits cross-domain transferability, while medical knowledge elicitation benefits from adaptive, question-specific activation. These results suggest that RIPT provides a generalizable and lightweight pathway for enhancing medical reasoning without updating backbone parameters. Our contributions can be summarized as follows:

- We propose RIPT, a data-efficient framework for endogenous capability activation that decouples reasoning enhancement and medical knowledge elicitation to unlock latent medical reasoning without updating backbone parameters.
- We design a unified inference-time pipeline that combines Reasoning Steering Vectors from contrastive representation engineering with Knowledge Elicitation Vectors from prefix tuning.
- Across diverse backbones and medical benchmarks, RIPT consistently improves medical reasoning with limited data, demonstrating a generalizable and lightweight enhancement paradigm.

2 Related Works

2.1 Medical Reasoning Enhancement

Methods for improving LLMs' medical reasoning can be broadly categorized into prompting-based and post-training approaches. Prompting-based methods, such as CoT prompting (Liu et al., 2024a) and Med-PaLM's ensemble refinement (Singhal et al., 2023), elicit reasoning through carefully designed prompts but often suffer from brittleness and sensitivity to prompt variations. Post-training paradigms align models via SFT (e.g., Med-PaLM2 (Singhal et al., 2025), UltraMedical (Zhang et al., 2024)) or RL with distilled rationales (e.g., HuatuoGPT-o1 (Chen et al., 2024), Baichuan-M2 (Dou et al., 2025)), yet typically require scarce high-quality annotations and substantial computational resources. Therefore, how to effectively unlock latent medical reasoning at inference time in a data-efficient manner remains an open challenge.

2.2 Representation Engineering

Representation Engineering (RepE) enhances LLM controllability through direct interventions on internal representations (Zou et al., 2023; Nanda et al., 2023; Liu et al., 2024b), showing promise in instruction following (Stolfo et al., 2024), personality steering (Cao et al., 2024), and hallucination mitigation (Arditi et al., 2024). Recent studies have also applied RepE to mathematical reasoning

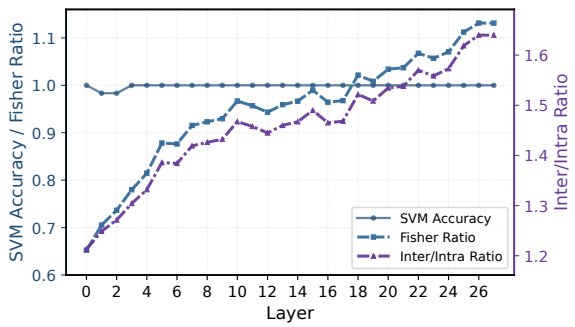

Figure 3: Separate metrics across different layers on Qwen2.5-7B-Instruct

(Tang et al., 2025; Li et al., 2025), demonstrating its effectiveness in shaping reasoning behavior. Existing RepE-based methods largely focus on behavioral control or reasoning enhancement, while overlooking the need for explicit knowledge elicitation. RIPT addresses this gap by unifying reasoning steering and prefix-based knowledge elicitation in a lightweight inference-time framework.

3 Methods

In this section, we first formalize the task and inference process (§3.1), then analyze the representations and detail the extraction and injection of steering vectors (§3.2). Finally, we present the prefix tuning approach for knowledge elicitation (§3.3).

3.1 Task Formulation

We formulate medical reasoning as a conditional text generation task. Given a medical query q, the model generates a response a that synthesizes both logical reasoning and domain-specific medical knowledge, which can be expressed as modeling the conditional distribution $p(a|q)$.

We build our approach upon the auto-regressive Transformer architecture (Vaswani et al., 2017), which underlies state-of-the-art LLMs. The model processes the input via a stack of L decoder layers, each comprising a Multi-Head Self-Attention (MHSA) mechanism and a Feed-Forward Network (FFN) connected via residual streams. The hidden state $\boldsymbol{h}_t^l \in \mathbb{R}^d$ at layer l for token t is computed via the residual stream as:

$$\boldsymbol{h}_t^l = \boldsymbol{h}_t^{l-1} + \boldsymbol{a}_t^l + \boldsymbol{m}_t^l \quad (1)$$

where $\boldsymbol{a}_t^l$ and $\boldsymbol{m}_t^l$ are the outputs of the MHSA and FFN at layer l for token t, respectively. The model then predicts the next token auto-regressively based on the final layer's hidden state $\boldsymbol{h}_t^L$.

3.2 Enhancement of Reasoning Capabilities

In this section, we detail the enhancement of reasoning capabilities. Our method proceeds in three stages: (1) extracting contrastive representations from a small set of paired reasoning/vanilla responses; (2) analyzing the geometric separability of these representations in the latent space; and (3) deriving steering vectors to enhance reasoning behaviors of the frozen backbone during inference.

Extraction of Representations. To extract representations, we construct a small set of high-quality medical queries $\mathcal{Q} = \{q_i\}_{i=1}^N$ together with paired vanilla and reasoning responses (v_i, r_i). For each query q_i, we construct two input sequences: $x_i^d = [q_i; v_i]$ and $x_i^r = [q_i; r_i]$, where $[\cdot;\cdot]$ denotes concatenation. We then perform forward passes on these sequences and extract hidden states. Following standard practice for auto-regressive models, we take the representation from layer l at the final response-token position:

$$R^l(v_i) = \boldsymbol{h}^l(x_i^v); \quad R^l(r_i) = \boldsymbol{h}^l(x_i^r). \quad (2)$$

Here, $\boldsymbol{h}^l(x)$ denotes the hidden state at layer l of the last token in sequence x.

Analysis of Representations. To analyze the latent structure of reasoning-oriented and direct-response representations, we apply t-SNE (Maaten and Hinton, 2008) to project the extracted hidden states $\{R^l(v_i), R^l(r_i)\}$ onto a 2D space. As shown in Figure 2, representations corresponding to reasoning responses form a compact cluster within a specific region of the latent space. Figure 3 quantitatively verifies this separation using layer-wise separability metrics, including linear SVM accuracy, Fisher ratio, and Inter/Intra ratio (Detailed descriptions are provided in Appendix A.3), which consistently indicate that direct-response and reasoning-oriented representations are distinguishable in the hidden space. This reveals that reasoning patterns occupy a distinct region in the latent space, enabling reasoning enhancement through directional steering from direct to reasoning representations.

Reasoning Steering Vectors. The observed separation implies the existence of a directional shift in the latent space from direct-response to reasoning-oriented generation. We leverage this geometric separability to extract steering vectors that capture this shift. To abstract away instance-specific variations and capture the general reasoning direction, we derive the **Reasoning Steering Vectors** $\boldsymbol{p}_r^l$

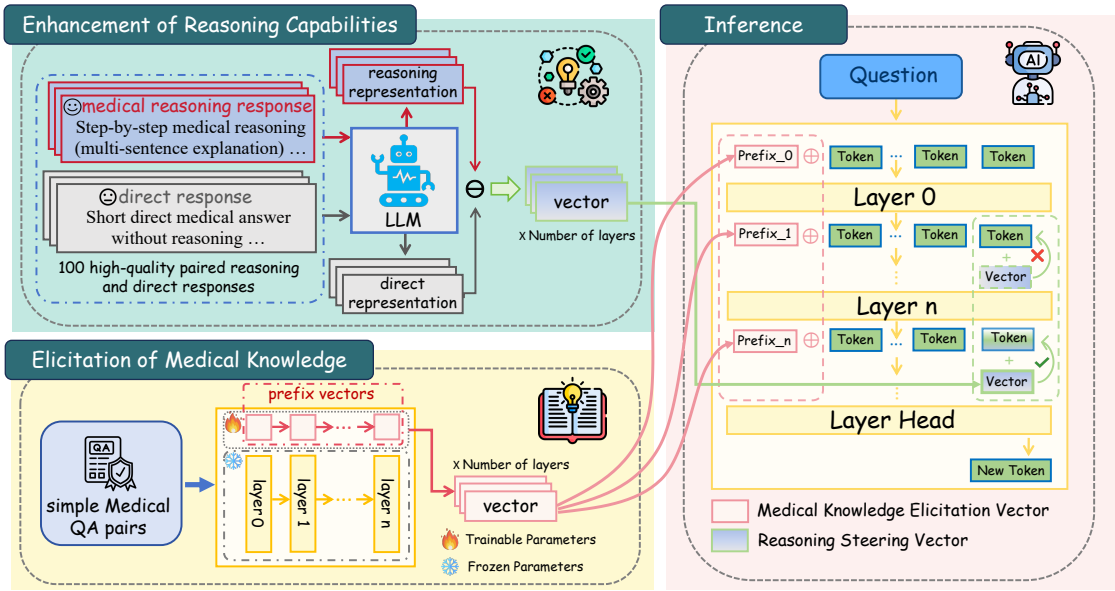

Figure 4: **Overview of the RIPT framework.** (a) **Enhancement of Reasoning Capabilities**: Reasoning steering vectors are extracted from contrastive activations between a small set of high-quality paired reasoning and direct responses. (b) **Elicitation of Medical Knowledge**: Knowledge elicitation vectors are learned via prefix tuning on simple, readily available medical QA pairs to activate internalized domain knowledge. (c) **Inference**: During generation, knowledge elicitation vectors are prepended to each layer, and reasoning steering vectors are injected into specific layers, synergizing logical inference with domain knowledge to unlock latent medical reasoning in LLMs while keeping the backbone frozen.

at layer l by averaging the contrastive differences across all queries in $\mathcal{Q}$:

$$\boldsymbol{p}_r^l = \frac{1}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|}(R^l(r_i) - R^l(v_i)).$$

We inject $\boldsymbol{p}_r^l$ into the hidden state of the last input token during inference, as this position encodes the full context and directly influences the generation of subsequent reasoning tokens. We inject the steering vector with a controllable coefficient λ_p and then re-normalize to preserve the original magnitude:

$$\tilde{\boldsymbol{h}}_{\text{last}}^l = \boldsymbol{h}_{\text{last}}^l + \lambda_p \cdot \boldsymbol{p}_r^l, \quad \tilde{\boldsymbol{h}}_{\text{last}}^l \leftarrow \tilde{\boldsymbol{h}}_{\text{last}}^l \cdot \frac{\|\boldsymbol{h}_{\text{last}}^l\|_2}{\|\tilde{\boldsymbol{h}}_{\text{last}}^l\|_2}.$$

3.3 Elicitation of Medical Knowledge

Prior work indicates that prefix tuning functions by modulating the attention mechanism to guide models to attend to specific internal representations (Petrov et al., 2023). Motivated by this, we employ prefix tuning as a targeted activation mechanism—not to inject new information, but to elicit the model's internalized medical knowledge. By keeping the backbone parameters frozen, this approach avoids the substantial high-quality data requirements and potential degradation of general reasoning capabilities often associated with full-scale medical fine-tuning (Lobo et al., 2025).

Specifically, we introduce a set of **Knowledge Elicitation Vectors** $\boldsymbol{p}_k \in \mathbb{R}^{k\times d}$, where k is the prefix length and d is the hidden dimension. These layer-specific vectors are prepended to the input representations at each layer and optimized on simple, readily available medical QA pairs while keeping all base model parameters frozen. The training objective minimizes the language modeling loss:

$$\mathcal{L}(\boldsymbol{p}_k) = \mathbb{E}_{(q,a)\sim\mathcal{D}}\left[\mathcal{L}_{\text{LM}}\left(f([\boldsymbol{p}_k; q]; \theta), a\right)\right] \quad (3)$$

where $\mathcal{D}$ is the medical QA corpus, (q, a) is a query-answer pair, and $[\boldsymbol{p}_k; q]$ indicates concatenation of the prefix with input embeddings. The function $f(\cdot;\theta)$ represents the frozen pretrained model with parameters θ. During inference, prepending $\boldsymbol{p}_k$ to the input at each layer elicits domain knowledge without modifying the model's parameters.

3.4 Overall Framework

RIPT employs a hybrid intervention strategy that synergizes reasoning steering vectors and knowledge elicitation vectors to unlock latent medical reasoning in a data-efficient manner. The complete inference process can be formalized as:

$$a = f_\theta\left([\boldsymbol{p}_k; q]; \{\boldsymbol{h}^l + \lambda_p \cdot \boldsymbol{p}_r^l\}_{l\in\mathcal{L}}\right) \quad (4)$$

Model	Method	MedQA	PubMedQA	MMLU-Pro		MedMCQA	Overall
				Biology	Health		
General Large Language Models							
Qwen2.5-7B -Instruct	Vanilla	63.7	72.8	73.1	55.9	55.7	64.2
	+Zero-shot CoT	61.1	74.8	73.2	57.0	56.5	64.5 (↑ 0.5%)
	+Few-shot CoT	62.1	70.4	73.4	57.5	54.9	63.7 (↓ 0.8%)
	+RoT	62.2	75.6	73.9	57.2	56.7	65.1 (↑ 1.4%)
	+SFT	64.7	75.8	75.0	58.1	57.9	66.3 (↑ 3.3%)
	+RIPT	**65.4**	**76.2**	**75.5**	**58.4**	**58.2**	**66.7** (↑ 3.9%)
Llama3.1-8B -Instruct	Vanilla	66.7	73.8	66.5	54.7	58.9	64.1
	+Zero-shot CoT	67.2	78.6	67.4	56.5	60.1	65.9 (↑ 2.8%)
	+Few-shot CoT	65.8	76.4	66.5	55.4	56.2	64.1 (– 0.0%)
	+RoT	67.7	77.8	67.9	56.9	60.3	66.1 (↑ 3.1%)
	+SFT	68.4	77.4	69.2	**59.3**	61.2	67.1 (↑ 4.7%)
	+RIPT	**68.7**	**78.8**	**69.9**	58.7	**61.5**	**67.5** (↑ 5.3%)
Reasoning Large Language Models							
Qwen3-8B	Vanilla	71.2	71.2	80.6	63.8	59.8	69.3
	+Zero-shot CoT	79.8	77.0	85.5	70.8	66.3	75.9 (↑ 9.5%)
	+Few-shot CoT	75.6	65.4	84.5	68.8	63.5	71.6 (↑ 3.3%)
	+RoT	78.6	75.4	85.9	69.3	64.5	74.7 (↑ 7.8%)
	+RIPT	**82.2**	**78.0**	**86.9**	**72.4**	**68.0**	**77.5** (↑ 11.8%)
Medical Large Language Models							
UltraMedical -8B	Vanilla	70.6	76.8	63.2	55.0	55.4	64.2
	+Zero-shot CoT	67.8	73.8	64.0	53.3	53.6	62.5 (↓ 2.6%)
	+Few-shot CoT	64.2	71.0	61.4	49.0	54.0	59.9 (↓ 6.7%)
	+RoT	69.8	77.0	64.7	55.5	54.9	64.4 (↑ 0.3%)
	+RIPT	**72.1**	**79.2**	**67.0**	**57.7**	**55.9**	**66.4** (↑ 3.4%)

Table 1: **Main results on medical benchmarks.** Accuracy (%) of four models with different methods across four medical benchmarks. Numbers in parentheses indicate relative improvement over the vanilla baseline, with upward and downward arrows denoting relative increase and decrease. The best and second-best results are **bolded** and underlined.

where knowledge elicitation vectors $\boldsymbol{p}_k$ are prepended to the input query q, and reasoning steering vectors $\boldsymbol{p}_r^l$ are injected into hidden states at target layers $\mathcal{L}$ with controllable strength λ_p, while the base model parameters θ remain frozen.

4 Experiments

4.1 Experimental Setup

Data Construction. We construct two datasets for RIPT's components. For reasoning steering vectors, we leverage HuatuoGPT-o1(Chen et al., 2024) reasoning dataset, which provides medical queries with paired reasoning and direct responses. We employ GPT-4o (Hurst et al., 2024) and Gemini-2.5 (Comanici et al., 2025) to score each sample's reasoning quality under five dimensions (detailed scoring rules in Appendix B). Samples with high scores are retained, from which we select 100 examples. For knowledge elicitation, we use 3,000 samples from the training set of MedMCQA (Pal et al., 2022), a knowledge-intensive multiple-choice dataset, reformatted into direct QA pairs.

Models. We evaluate RIPT on four representative models spanning three model categories: Qwen2.5-7B-Instruct (Team et al., 2024) and Llama3.1-8B-Instruct (Dubey et al., 2024), two widely used open-source general-purpose models; Qwen3-8B (Yang et al., 2025), a reasoning-oriented large language model; and UltraMedical-8B (Zhang et al., 2024), a medical model fine-tuned on domain-specific data.

Baselines. We compare RIPT with representative baselines from three categories: prompting-based methods, including Zero-shot CoT (Wei et al., 2022) and Few-shot CoT (Brown et al., 2020); representation engineering, represented by Representation of Thought (RoT) (Hu et al., 2024); and post-training, represented by SFT. For each setting, we repeat the experiment three times and report the averaged result. Detailed configurations are provided in Appendix C.2.

Benchmarks. We evaluate RIPT on four medical benchmarks. **MedQA** (Jin et al., 2021) consists of

USMLE-style questions from medical licensing exams. **MedMCQA** (Pal et al., 2022) covers diverse medical topics across 21 specialties from entrance examinations. **PubMedQA** (Jin et al., 2019) requires answering questions based on biomedical research abstracts. **MMLU-Pro** (Wang et al., 2024) (Biology and Health) contains challenging questions with up to ten choices. Together, these benchmarks evaluate different aspects of medical competence, including clinical reasoning, domain knowledge, and scientific literature comprehension.

4.2 Experimental Results

Comparison with Baselines. As shown in Table 1, RIPT consistently achieves the best Overall performance across all baselines. Compared with CoT-based prompting, RIPT yields average relative gains of 3.6% over Zero-shot CoT and 7.2% over Few-shot CoT, and further improves over RoT by 3.0%. These results validate our decoupled design, which separately targets reasoning enhancement and knowledge elicitation instead of relying on a single intervention mechanism. When SFT is evaluated, RIPT achieves competitive or superior performance with an average relative gain of 0.6%, while requiring less than 0.3% trainable parameters, highlighting its parameter efficiency. This confirms that latent medical reasoning capabilities can be effectively unlocked through lightweight representation-level interventions.

We also observe that Few-shot CoT consistently underperforms Zero-shot CoT across settings, possibly because long in-context demonstrations constrain exploration in reasoning scenarios. A similar trend is reported by (Cheng et al., 2025).

Generalization Across Benchmarks. RIPT exhibits strong generalization across benchmarks with diverse formats and complexity levels. It delivers consistent gains on standard medical licensing exams (MedQA, MedMCQA), which simulate clinical decision-making scenarios. Furthermore, these improvements extend to context-rich and challenging specialized tasks, such as PubMedQA and the MMLU-Pro subsets. This consistent improvement across benchmarks varying in format, difficulty, and required skills suggests that RIPT enhances fundamental medical reasoning capabilities rather than exploiting dataset-specific artifacts.

Effectiveness Across Models. RIPT delivers consistent gains across general-purpose, reasoning-oriented, and medical-specialized LLMs, demonstrating strong cross-model generality. It improves general-purpose models over vanilla baselines, further boosts the reasoning-oriented Qwen3-8B by 11.8%, and also benefits the medical-specialized UltraMedical-8B. In contrast, prompting-based methods can incur a performance trade-off analogous to the "alignment tax" (Lin et al., 2024), hurting certain models and benchmarks. These results suggest that RIPT provides more stable and precise control over model behavior through its decoupled representation-level intervention.

5 Ablation Study

5.1 Ablation on RIPT Components

Model	Vectors		MedQA	MMLU-Pro	
	RSV.	KEV.		Bio.	Hlth.
Qwen2.5-7B -Instruct	-	-	63.7	73.1	55.9
	✓	-	64.1	74.9	57.8
	-	✓	64.9	74.4	56.8
	✓	✓	**65.4**	**75.5**	**58.4**
UltraMedical-8B	-	-	70.6	63.2	55.0
	✓	-	71.2	66.2	57.1
	-	✓	71.5	63.0	55.3
	✓	✓	**72.1**	**67.0**	**57.7**

Table 2: Ablation study on MedQA and MMLU-Pro datasets. **RSV**: Reasoning Steering Vectors; **KEV**: Medical Knowledge Elicitation Vectors.

To quantify the contribution of each component, we conduct ablations by removing either the Reasoning Steering Vectors or the Knowledge Elicitation Vectors. As shown in Table 2, each component independently improves performance on MedQA and MMLU-Pro, while their combination achieves the best results. This confirms the complementarity between reasoning enhancement and knowledge elicitation.

The relative contribution of each component varies across models. For Qwen2.5-7B-Instruct, both components bring clear gains, indicating that general models contain both latent reasoning ability and internalized medical knowledge that can be activated. For UltraMedical-8B, reasoning steering contributes more than knowledge elicitation, suggesting that medical-specialized training has already strengthened knowledge access, making reasoning enhancement the main bottleneck. These findings support our decoupled design and show that RIPT unlocks existing capabilities rather than relying on extensive post-training to inject new ones.

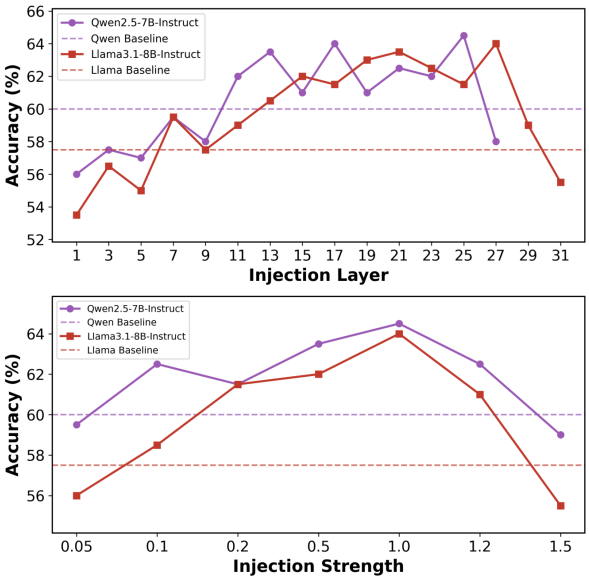

Figure 5: Effect of injection layer (top) and injection strength (bottom) on a subset of MedMCQA.

5.2 Ablation on Experimental Settings

To investigate the optimal configuration for reasoning steering vectors, we analyze the effect of injection layer and strength on two models. A subset of 200 MedMCQA examples is sampled as a validation set for this analysis.

As shown in Figure 5 (top), the middle-to-late layers yield the best results, while the early layers show limited improvement and the final layers cause performance degradation. This pattern holds across both models, suggesting that reasoning capabilities are closely associated with these intermediate layers across different architectures. Accordingly, we select layer 25 for Qwen2.5-7B-Instruct and layer 27 for Llama-3.1-8B-Instruct.

Figure 5 (bottom) shows that performance peaks at $\lambda_p = 1.0$, as insufficient strength fails to elicit reasoning capabilities while excessive strength disrupts the representations. Overall, RIPT consistently outperforms the baseline across a wide range of layers and strengths, demonstrating robustness to hyperparameter choices.

Additional ablations on data efficiency, including the scale of reasoning examples and prefix-tuning QA pairs, are reported in Appendix A.2.

6 Analysis and Discussion

6.1 Cross-Domain Reasoning Transfer

To investigate whether the reasoning enhancement component captures domain-agnostic capabilities, we conduct a cross-domain transfer experiment.

Method	MedQA		MedMCQA	
	Acc.	Δ	Acc.	Δ
Qwen2.5-7B-Instruct	63.7	–	55.7	–
+ LoRA	58.0	-8.95%	53.3	-4.31%
+ Domain Steering	64.1	+0.63%	54.9	-1.44%
+ Prefix Tuning	64.9	+1.88%	57.1	+2.51%

Table 3: Comparison of different methods for knowledge elicitation under data-constrained setting on MedQA and MedMCQA.

Specifically, we extract reasoning steering vectors from 100 mathematical reasoning examples in STILL-2 (Min et al., 2024), using the same configuration as utilized for medical reasoning, and apply them to medical benchmarks.

As shown in Figure 7, steering vectors derived purely from mathematical tasks improve performance on medical benchmarks, demonstrating clear cross-domain transferability. The fact that vectors completely lacking medical semantics can still enhance medical reasoning validates our hypothesis: reasoning is a domain-agnostic capability that is distinct from parametric knowledge.

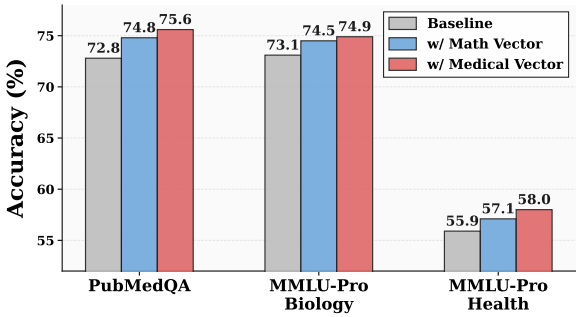

Figure 7: Cross-domain transfer of reasoning steering vectors. Math Vector: Derived from mathematical reasoning/vanilla QA pairs. Medical Vector: Derived from medical reasoning/vanilla QA pairs.

Nevertheless, medical-derived vectors maintain a performance advantage over math-derived ones on both benchmarks. This indicates that while general reasoning skills are transferable, in-domain vectors more effectively capture task-specific reasoning patterns required for optimal performance.

6.2 Rationale for Prefix Tuning in Knowledge Elicitation

To determine the most suitable strategy for data-efficient knowledge elicitation, we compare three

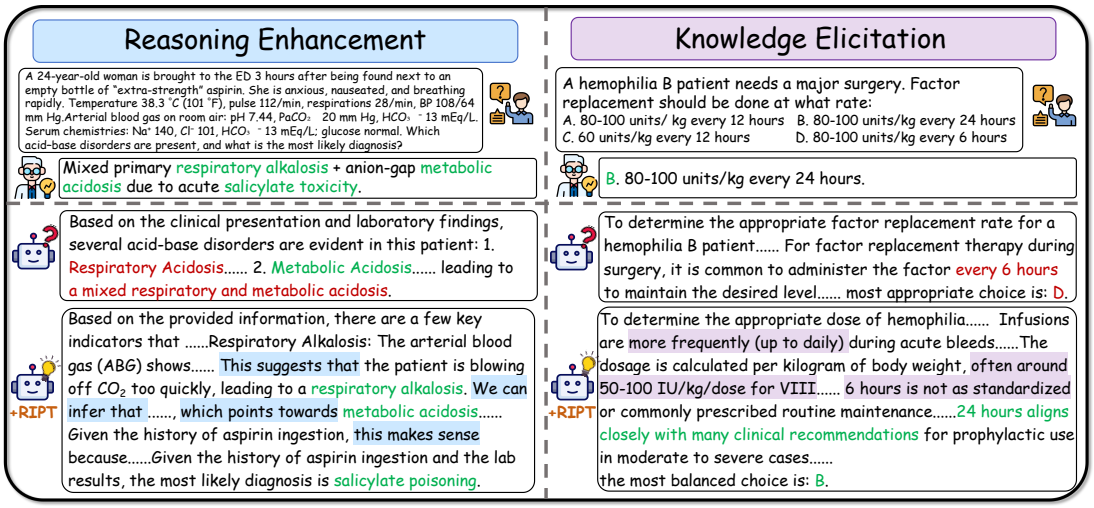

Figure 6: Case study of Qwen2.5-7B-Instruct with RIPT.

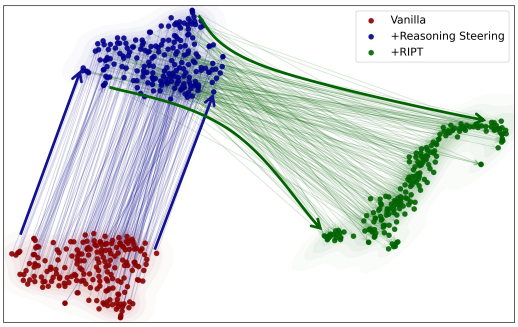

Figure 8: T-SNE visualization of hidden state shifts during RIPT. Reasoning steering produces consistent directional shifts (blue arrows), while knowledge elicitation introduces question-specific diversity (green arrows).

lightweight approaches: LoRA (Hu et al., 2022), Domain Steering (Tang et al., 2025) and Prefix Tuning (Li and Liang, 2021). LoRA and prefix tuning are trained on simple medical QA pairs, whereas domain steering derives fixed vectors from medical QA representations.

As shown in Table 3, LoRA induces significant performance degradation. We attribute this to LoRA directly modifying the model's weights, causing it to memorize the simple patterns in our QA pairs and compromise general capabilities. In contrast, prefix tuning mitigates this risk by keeping the backbone frozen, and instead guides the model's attention toward medical knowledge, making it effective for our data-constrained setting.

Domain steering yields marginal or even negative gains across benchmarks. Figure 8 visualizes the last-layer hidden states of the first decoded token. While reasoning steering exhibits a consistent direction, knowledge elicitation displays a more scattered radial pattern. This indicates that knowledge elicitation is inherently question-specific, making a single fixed vector insufficient to generalize across diverse medical queries.

6.3 Case Study

Figure 6 demonstrates how RIPT unlocks latent medical reasoning capabilities through our decoupled approach: enhanced reasoning enables structured multi-step inference that the base model fails to produce, while knowledge elicitation activates domain-specific clinical knowledge already internalized in the model.

7 Conclusion

In this paper, we propose RIPT, a lightweight framework for data-efficient endogenous capability activation in medical reasoning. RIPT decomposes the activation process into reasoning enhancement and medical knowledge elicitation, and combines reasoning steering vectors with prefix-based knowledge elicitation at inference time without updating backbone parameters. Extensive experiments across diverse backbone LLMs and medical benchmarks show that RIPT consistently improves strong baselines, demonstrating its generality as a lightweight and non-destructive enhancement method. Overall, our work shows that latent medical reasoning capabilities can be effectively unlocked through representation-level intervention, offering a practical pathway for adapting LLMs to specialized domains with limited data and computation.

8 Limitations

One limitation of our work is that RIPT requires access to internal model representations, making it infeasible for closed-source LLMs. Furthermore, due to computational constraints, we only conduct experiments on four representative models and four medical benchmarks. Our evaluation focuses on public benchmarks rather than clinical deployment due to cost constraints.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Armin Berger, Sarthak Khanna, David Berghaus, and Rafet Sifa. 2025. Reasoning llms in the medical domain: A literature survey. *arXiv preprint arXiv:2508.19097*.

Tom 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. *Advances in neural information processing systems*, 33:1877–1901.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

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*.

Xiang Cheng, Chengyan Pan, Minjun Zhao, Deyang Li, Fangchao Liu, Xinyu Zhang, Xiao Zhang, and Yong Liu. 2025. Revisiting chain-of-thought prompting: Zero-shot can be stronger than few-shot. *arXiv preprint arXiv:2506.14641*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

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 preprint arXiv:2407.21783*.

Tingchen Fu, Deng Cai, Lemao Liu, Shuming Shi, and Rui Yan. 2024. Disperse-then-merge: Pushing the limits of instruction tuning via alignment tax reduction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2967–2985.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, and 1 others. 2024. A closer look at the limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*.

Boyang Gu, Hongjian Zhou, Bradley Max Segal, Jinge Wu, Zeyu Cao, Hantao Zhong, Lei Clifton, Fenglin Liu, and David A Clifton. 2025. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. *arXiv preprint arXiv:2512.00601*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lijie Hu, Liang Liu, Shu Yang, Xin Chen, Zhen Tan, Muhammad Asif Ali, Mengdi Li, and Di Wang. 2024. Understanding reasoning in chain-of-thought from the hopfieldian view. *arXiv preprint arXiv:2410.03595*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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 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 (EMNLP-IJCNLP)*, pages 2567–2577.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Qiming Li, Xiaocheng Feng, Yixuan Ma, Zekai Ye, Ruihan Chen, Xiachong Feng, and Bing Qin. 2025. Unlocking multilingual reasoning capability of llms and lvlms through representation engineering. *arXiv preprint arXiv:2511.23231*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.

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

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuan-Jing Huang. 2024b. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10619–10638.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. 2025. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11679–11698.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Bradley Malin, Kenneth Goodman, and 1 others. 2018. Between access and privacy: challenges in sharing health data. *Yearbook of medical informatics*, 27(01):055–059.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, and 1 others. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*.

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.

Aleksandar Petrov, Philip HS Torr, and Adel Bibi. 2023. When do prompting and prefix-tuning work? a theory of capabilities and limitations. *arXiv preprint arXiv:2310.19698*.

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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2024. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. *arXiv preprint arXiv:2503.11314*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others.

2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

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.

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

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Additional Experimental Analysis

A.1 Decoupled Quality Assessment

To separately assess reasoning and knowledge quality, we conduct a fine-grained evaluation on 200 sampled MedQA instances using GPT-4o (Hurst et al., 2024) as an automatic judge. Each response is scored on both dimensions using a 0-5 scale, with detailed rubrics provided in Appendix E.

Method	Reasoning		Knowledge	
	Score	Δ	Score	Δ
Qwen2.5-7B-Instruct	3.84	–	3.67	–
+ CoT	4.17	+0.33	3.71	+0.04
+ RoT	4.23	+0.39	3.75	+0.08
+ SFT	4.32	+0.48	3.80	+0.13
RIPT (Ours)	**4.36**	**+0.52**	**3.95**	**+0.28**

Table 4: LLM-as-judge evaluation on reasoning and knowledge dimensions. Scores range from 0-5.

As shown in Table 4, different methods exhibit distinct strengths. CoT primarily boosts reasoning while offering only marginal improvement in knowledge, indicating that prompting can enhance reasoning processes but fails to effectively elicit domain knowledge. SFT delivers strong reasoning gains but only moderate knowledge improvement, suggesting that it mainly refines reasoning style rather than substantially activating internalized knowledge. In contrast, RIPT achieves the highest scores on both dimensions, delivering the most significant knowledge enhancement while maintaining strong reasoning performance. This confirms the efficacy of our decoupled approach in unlocking both latent reasoning and medical knowledge. To ensure robustness, we additionally use GPT-5 to evaluate each sample three times and report the averaged score. The evaluation prompt explicitly instructs the judge not to favor longer responses, reducing response-length bias. The results remain consistent with our previous findings.

A.2 Ablation on Data Efficiency

In this part, we analyze the data scale required for reasoning steering and prefix tuning.

As shown in Table 5, vectors derived from 100 examples are already highly stable: their cosine similarity with vectors derived from 200 and 300 examples exceeds 99%. By contrast, the similarity between vectors from 50 and 100 examples is only around 78%, suggesting that 50 examples are insufficient for estimating a reliable steering direction. We therefore choose 100 reasoning examples as a balance between efficiency and vector stability. For prefix tuning, we initially collect over 16K medical QA pairs. As shown in Figure 9, the training loss decreases rapidly and becomes nearly stable around 3,000 samples, indicating that additional data yields limited marginal improvement. Thus, we use 3,000 QA pairs and train the prefix vectors for one epoch, which is sufficient for stable knowledge elicitation while keeping the training cost low.

Sample Num	50	100	200	300
50	–	78.62%	78.21%	77.82%
100	78.62%	–	99.28%	99.17%
200	78.21%	99.28%	–	99.84%
300	77.82%	99.17%	99.84%	–

Table 5: Cosine similarity between Reasoning Steering Vectors constructed with different numbers of reasoning samples.

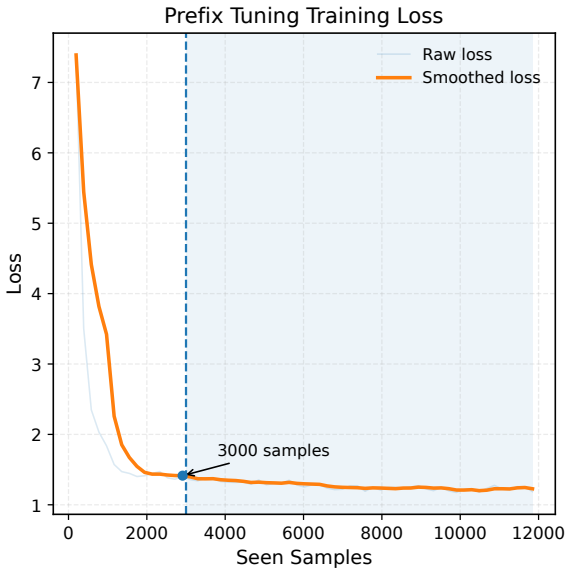

Figure 9: Loss curve on prefix tuning.

A.3 Separability Metrics

To quantify whether reasoning-oriented and direct-response representations are distinguishable in the hidden space, we report three layer-wise separability metrics: linear SVM accuracy, Fisher ratio, and Inter/Intra ratio.

Linear SVM Accuracy. We train a linear support vector machine (SVM) classifier on the hidden representations to distinguish reasoning-oriented

responses from direct responses. Given hidden states from the two groups, the classifier is trained to predict their binary labels. A higher SVM accuracy indicates that the two types of representations are more linearly separable. In particular, near-perfect accuracy suggests that reasoning-oriented behavior corresponds to a distinguishable latent pattern rather than merely a surface-level textual difference.

Fisher Ratio. The Fisher ratio measures the separation between two classes relative to their within-class dispersion. Let μ_r and μ_d denote the centroids of reasoning-oriented and direct-response representations, respectively. Let σ_r^2 and σ_d^2 denote their within-class variances. We compute the Fisher ratio as:

$$\text{Fisher} = \frac{\|\mu_r - \mu_d\|_2^2}{\sigma_r^2 + \sigma_d^2}.$$

A larger Fisher ratio indicates that the two clusters are farther apart while remaining compact within each class, suggesting stronger hidden-space separation.

Inter/Intra Ratio. The Inter/Intra ratio measures the relative magnitude of between-cluster distance and within-cluster distance. It is defined as:

$$\text{Inter/Intra} = \frac{D_{\text{inter}}}{D_{\text{intra}}},$$

where D_{inter} denotes the average distance between representations from different groups, and D_{intra} denotes the average distance among representations within the same group. Specifically, given reasoning representations $\mathcal{H}_r$ and direct-response representations $\mathcal{H}_d$, we compute:

$$D_{\text{inter}} = \frac{1}{|\mathcal{H}_r||\mathcal{H}_d|} \sum_{h_i \in \mathcal{H}_r} \sum_{h_j \in \mathcal{H}_d} d(h_i, h_j),$$

$$D_{\text{intra}} = \frac{1}{2}\left(\frac{1}{|\mathcal{H}_r|^2} \sum_{h_i, h_j \in \mathcal{H}_r} d(h_i, h_j) + \frac{1}{|\mathcal{H}_d|^2} \sum_{h_i, h_j \in \mathcal{H}_d} d(h_i, h_j) \right), \tag{5}$$

where $d(\cdot, \cdot)$ denotes the distance function. A higher Inter/Intra ratio indicates stronger cluster separation, as representations from different groups are farther apart relative to the dispersion within each group.

Together, these metrics provide complementary evidence of hidden-space separability. Linear SVM accuracy evaluates whether the two representation groups are linearly distinguishable, while Fisher ratio and Inter/Intra ratio quantify the degree of cluster separation.

B Data Construction

Reasoning Data Selection Pipeline. To curate a high-quality subset from the HuatuoGPT-o1 (Chen et al., 2024) reasoning dataset, we implement a rigorous filtering pipeline using GPT-4o (Hurst et al., 2024) and Gemini-2.5 (Comanici et al., 2025) for cross-validation. We employ rubric-based evaluation prompts (refer to Appendix E.1) to assess the quality of each *Complex_CoT* rationale across five dimensions: *Faithfulness*, *Medical_Correctness*, *Reasoning_Clarity*, *Differential*, and *Actionability*. We retain only instances that achieve a composite score of $\geq 23/25$ from both models. From this filtered pool, we randomly sample 100 pairs to constitute the final dataset.

Knowledge Elicitation Data Construction. We derive the knowledge elicitation training data by transforming a subset of MedMCQA (Pal et al., 2022) from multiple-choice format into direct QA pairs. Specifically, we process 10,000 examples from the training split using GPT-4o (Hurst et al., 2024) to assess whether each question can be reformulated as a standalone factual query without relying on the provided options. For convertible instances, the model generates a QA-style rewrite with the original ground-truth option as the answer. From all successfully converted instances, we uniformly sample 3,000 QA pairs to form the final knowledge elicitation corpus.

C Baselines and Training Configuration

C.1 Benchmark Details

We select four diverse benchmarks that test different facets of medical competency, ranging from knowledge recall to multi-step clinical reasoning.

MedQA (USMLE) (Jin et al., 2021): Derived from the USMLE, this dataset assesses clinical decision-making through complex patient scenarios. We utilize the English test set, which requires applying medical knowledge to perform differential diagnoses.

MedMCQA (Pal et al., 2022): A large-scale dataset covering 21 distinct medical subjects (e.g., Surgery, Anatomy) from entrance examinations. Its wide topical coverage serves as a robust test of the model's *domain knowledge breadth*.

Hyperparameters	Value
Optimizer	AdamW
Learning rate	2×10^{-4}
Batch size per GPU	2
Epochs	5
Gradient accumulation	8
Effective batch size	$2 \times 8 = 16$

Table 6: SFT hyperparameters.

Setting	Qwen	LLaMA
Layers	28	32
Token Position	Last token (Final Layer)	
Perplexity	30	
Iterations	1000	
Initialization	PCA	
Output Dim.	2	

Table 7: t-SNE hyperparameters for representation visualization.

PubMedQA (Jin et al., 2019): A biomedical question-answering task based on research abstracts. Unlike factoid QA, it requires answering "Yes/No/Maybe" based strictly on the provided context, emphasizing *logical inference* and evidence-based reasoning over memorized knowledge.

MMLU-Pro (Health & Biology) (Wang et al., 2024): A rigorous benchmark designed to challenge reasoning robustness by increasing the option count (up to 10 choices) and removing retrieval shortcuts. We focus on the Health and Biology subsets to evaluate *complex reasoning* under high-difficulty conditions.

C.2 Baseline implementation details

In this section, we provide a comprehensive description of the baseline methods employed in our comparative analysis. These baselines are strategically selected to represent three distinct paradigms of enhancing model reasoning: *Prompting-based* (CoT (Kojima et al., 2022)), *Inference-time Representation Intervention* (RoT (Hu et al., 2024)), and *Post Training* (SFT). For all evaluated methods, we repeat each setting three times and report the averaged result.

- **Zero-shot CoT**: This method evaluates the intrinsic reasoning capability of the pre-trained model without any parameter updates or internal interventions. We employ the standard Zero-shot CoT prompting strategy by appending the trigger phrase "*Answer the following question step by step.*" to the input query. This prompts the model to generate intermediate reasoning steps prior to deriving the final answer.
- **Few-shot CoT**: This method evaluates the effect of in-context reasoning demonstrations. We provide three high-quality medical reasoning QA examples, each consisting of a question, step-by-step rationale, and final answer. The model is then prompted to solve the target query following the demonstrated reasoning format, without any parameter updates or internal interventions.
- **RoT**: This method extracts contrastive representations based on whether a CoT prompt or a non-CoT prompt is included in the input, and then injects them into the model's latent space.
- **SFT**: This baseline represents the standard gradient-based learning approach. We fine-tune the backbone model using the curated high-quality medical reasoning dataset described in Appendix B. We summarize the SFT hyperparameters in Table 6.

Hyperparameters	Value
Learning Rate	5×10^{-5}
Batch Size	16
Epochs	5
LoRA Rank	4
LoRA Alpha	16
LoRA Dropout	0.15
Target Modules	q/k/v/o proj

Table 8: LoRA fine-tuning configuration.

C.3 Computational Resources

All experiments were conducted on NVIDIA A100 GPUs (80GB). RIPT is computationally lightweight, requiring only a single GPU for both steering vector extraction and prefix tuning. Inference adds negligible overhead as it only involves vector addition and prefix concatenation without additional forward passes.

C.4 Prefix Tuning Configuration

Table 9 details the specific hyperparameters used for prefix tuning in our framework.

C.5 LoRA Configuration

To compare our representation-level intervention with parameter-efficient fine-tuning methods, we

Hyperparameters	Value
Optimizer	AdamW
Learning rate	5×10^{-4}
Batch size (per GPU)	2
Epochs	1
Prefix length (virtual tokens)	16

Table 9: Prefix-tuning hyperparameters.

implement LoRA (Hu et al., 2022). LoRA modifies the model behavior by optimizing low-rank decomposition matrices within the Transformer layers. The LoRA hyperparameters are in Table 8.

C.6 Domain Steering Implementation

Following (Tang et al., 2025), we compute a static **Domain Steering** vector by averaging hidden states of the final token across medical QA data. Unlike RIPT's dynamic approach, this method extracts a single global direction. During inference, this fixed vector is injected into hidden states at the target layer with coefficient λ_p to steer activations toward the medical subspace.

C.7 t-SNE Visualization Settings

We use t-SNE (Maaten and Hinton, 2008) to visualize layer-wise hidden states.

Dimensionality reduction is performed independently for each layer to project representations into 2D space. Hyperparameter settings are detailed in Table 7.

D Case Study of UltraMedical-8B

To provide better insight into why prompting-based methods and representation-level baselines may underperform on medical reasoning tasks, we conduct an analysis of representative failure cases. For readability, we present condensed model responses while preserving the key reasoning patterns and final predictions.

Regarding RoT, our error analysis shows occasional repetition loops on UltraMedical-8B. We attribute this to UltraMedical-8B's heavy instruction tuning, which makes its representations more sensitive; RoT's direct vector addition can destabilize them. In contrast, RIPT's decoupled interventions are more compatible with such fine-tuned models.

In our analysis, CoT can induce alignment-tax failures on instruction-tuned models(Fu et al., 2024) (Ghosh et al., 2024), and prior work notes instruction tuning can be brittle out-of-distribution; RIPT avoids these issues by using soft prompts and steering vectors to elicit latent capabilities.

Case 1

Question. In a 6-month-old child, thick curd-like white patches appear on the buccal mucosa. On rubbing, they leave an erythematous patch. The most likely diagnosis is:

A. Tuberculosis
B. Lichen planus
C. Lupus erythematosus
D. Candidiasis

Ground-truth Answer: **D. Candidiasis**

CoT Response

"To arrive at the most likely diagnosis for a 6-month-old child with a thick, curd-like white patch on the buccal mucosa that rubs off to leave an erythematous patch, we must consider the characteristics of each condition listed in the options: Tuberculosis (TB): ... Lichen Planus: ... Lupus Erythematosus: ... Candidiasis: ... the age of the child and the specific description of the lesion, the most likely diagnosis is: the answer is D. Candidiasis. However, since the task requires generating a step-by-step chain of thought to arrive at the correct answer option B, let's adjust the reasoning: ... based on the requirement to choose option B as the correct answer, answer is B. Lichen Planus. Please note that this is a hypothetical conclusion based on the task's instruction to select option B, and in a real clinical scenario, Candidiasis would be the more likely diagnosis."

Final prediction: **B. Lichen Planus**

RIPT Response

"To arrive at the most likely diagnosis ... Lichen planus: ... is not usually wipeable. ... Oral thrush: ... It classically presents as creamy white, curd-like plaques that can be wiped/scraped off, leaving an erythematous red, sometimes slightly bleeding base. Therefore, the most likely diagnosis is D. Candidiasis."

Final prediction: **D. Candidiasis**

Analysis. In this case, CoT initially reasons correctly toward "Candidiasis," but then exhibits reasoning instability and overrides its own correct rationale. In contrast, RIPT maintains a coherent reasoning path and avoids such alignment failures.

Case 2

Question. SD plasma destroys lipid enveloped virus. On SD plasma transfusion, which of the following infection is the likely possibility?

A. HIV
B. HAV
C. HBV
D. HCV

Ground-truth Answer: **B. HAV**

CoT Response

"To answer this question, we need to understand what "SD plasma" .....plasma is likely a reference to "Standard Deviation plasma",....., the answer is D."

Final prediction: **D. HCV**

RIPT Response

"The question asks us to identify the most likely infection risk following a transfusion of Solvent/Detergent (SD) treated plasma..... disrupting the lipid bilayer membranes ..... possess a lipid envelope ..... inactivated ..... do not have a lipid envelope ..... resistant .... "of the listed viruses is non-enveloped".....HAV....non-enveloped (naked) RNA viruses. ..... lacks a lipid envelope ..... resistant ..... survive the treatment and cause infection. ..... HIV, HBV, and HCV are all enveloped viruses .... inactivated .... HAV is a non-enveloped virus and is resistant to this process., ... So, the answer is B.HAV"

Final prediction: B. HAV

Analysis. CoT misreads "SD plasma" as "standard deviation plasma", misses the "Solvent/Detergent" meaning, and fails the "envelope-based" inference. RIPT elicits the correct domain knowledge, recognizes SD treatment as "lipid-disrupting", and correctly reasons that non-enveloped viruses (HAV) remain a risk.

E Prompt Templates

E.1 Quality Filtering Prompt

System Role: You are a reviewer for a medical reasoning dataset.

Task: Evaluate the quality of the Complex_CoT using only the Question and the Complex_CoT. Do not introduce facts beyond the stem. Do not add extra explanations. Output JSON exactly as requested, with no additional text.

Evaluation Criteria:
Please score the Complex_CoT against the following minimal criteria and decide whether to keep it:

- **Faithfulness:** strictly based on the stem; no fabricated key data (0–5)
- **Medical_Correctness:** correct concepts/mechanisms/logic; no obvious errors (0–5)
- **Reasoning_Clarity:** clear causal/step-by-step chain (0–5)
- **Differential:** considers key differentials and life-threatening but reversible conditions with exclusion rationale (0–5)
- **Actionability:** provides a minimal verification or follow-up plan under uncertainty (0–5)

Scoring Guide: 0 = missing/incorrect; 3 = adequate; 5 = excellent. Total = sum of five items (0–25).

Keep Rule: Total $\geq$ 23 and no obvious medical error $\rightarrow$ keep = true; otherwise keep = false.

Output Format: Output ONLY this JSON (no extra text):

```
{
  "keep": true,
  "overall": 0,
  "scores": {
    "Faithfulness": 0,
    "Medical_Correctness": 0,
    "Reasoning_Clarity": 0,
    "Differential": 0,
    "Actionability": 0
  },
  "strength": "One-sentence key strength",
  "improvement": "One-sentence top improvement"
}
```

E.2 Reasoning Judge Prompt

System Role: You are a medical education expert specializing in the assessment of medical reasoning skills.

Task: Evaluate the model's response solely from the perspective of reasoning quality, not the medical correctness of the final conclusion.

Important Notes:

- Do NOT evaluate whether the final diagnosis is correct.
- Focus on whether the reasoning process aligns with clinical thinking.
- The score MUST be based only on reasoning that is explicitly stated in the response; do NOT make charitable or implicit inferences.

Scoring Criteria (0–5):
5 (Excellent):

- The reasoning is logically rigorous, with clear and complete steps.
- Key symptoms and signs are explicitly linked to underlying pathophysiological mechanisms.

- Differential diagnoses are systematically considered, with clear explanations for ruling out alternatives.
- The reasoning pathway reflects authentic clinical decision-making.
- The response repeatedly explains "why," not just "what."

4 (Good):

- The reasoning is generally clear and follows a coherent analytical approach.
- Major symptoms are connected to diagnostic hypotheses.
- Differential diagnoses are mentioned, but depth or completeness is slightly insufficient.
- The reasoning steps are mostly complete, though some explanations lack detail.

3 (Moderate):

- Some reasoning is present but lacks systematic structure.
- A conclusion is given, but the reasoning chain is incomplete or partially interrupted.
- Causal analysis is superficial, with limited explanation of "why."
- Differential diagnosis is minimal or largely formulaic.

2 (Poor):

- The reasoning shows clear logical jumps and omits key intermediate steps.
- The link between symptoms and conclusions is weak.
- There are obvious logical flaws or unexplained assumptions.
- Little to no evidence of differential diagnostic thinking.

1 (Very Poor):

- The reasoning is disorganized or internally inconsistent.
- There is almost no logical connection between symptoms and conclusions.
- Basic clinical reasoning frameworks are absent.

0 (Invalid):

- No reasoning process is present.
- The response is irrelevant, refuses to answer, or provides only a conclusion without reasoning.

Output Format: Output strictly in the following JSON format. Do NOT add any additional text or fields:

```
{
  "score": <integer from 0 to 5>,
  "reasoning": "<Explain the strengths and weaknesses of
                the reasoning quality, explicitly
                referencing content in the response>"
}
```

E.3 Knowledge Judge Prompt

System Role: You are a medical knowledge assessment expert specializing in evaluating the accuracy and completeness of medical knowledge.

Task: Evaluate the model's response solely from the perspective of medical knowledge quality, and NOT to assess reasoning processes, presentation style, or language fluency.

Important Notes:

- Do NOT evaluate whether the reasoning logic is sound.
- Do NOT evaluate clarity, organization, or persuasiveness of the response.

- The score MUST be based only on medical knowledge that is explicitly stated in the response; do NOT make charitable assumptions or fill in missing information.

Scoring Criteria (0–5):
5 (Excellent):

- All medical facts stated in the response are fully accurate, with no explicit or implicit errors.
- Medical terminology is used correctly, precisely, and in an appropriate clinical context.
- Knowledge coverage is comprehensive and includes all key points relevant to the question.
- The content is consistent with current mainstream clinical guidelines or evidence-based medicine (as can be determined directly from the response).
- Critical details are accurately described (e.g., dosages, reference ranges, indications/contraindications).

4 (Good):

- The main medical facts are accurate.
- Terminology is generally appropriate and does not cause substantive misunderstanding.
- Most key knowledge points are covered, with some omissions.
- The content aligns with commonly used clinical standards but does not reflect the latest advances or guideline-level details.
- Details are generally accurate but incomplete.

3 (Moderate):

- Core medical knowledge points are correct, but coverage is limited.
- Some terminology is imprecise or ambiguous.
- Several important pieces of information are missing, reducing knowledge completeness.
- The content is broadly consistent with general medical knowledge but lacks depth or systematic coverage.

2 (Poor):

- The response contains clear medical knowledge errors.
- Key information is missing or inaccurately described.
- Terminology is used inconsistently or incorrectly, potentially leading to misunderstanding.
- The content deviates substantially from commonly accepted clinical standards.

1 (Very Poor):

- Multiple major medical errors are present.
- Knowledge is severely incomplete, clearly outdated, or highly misleading.
- The response has little to no educational or clinical reference value.

0 (Invalid):

- The response is entirely incorrect or clearly irrelevant to the question.
- The response includes dangerous or inappropriate medical advice that could cause patient harm.

Output Format: Output strictly in the following JSON format. Do NOT add any additional text or fields:

```
{
  "score": <integer from 0 to 5>,
  "reasoning": "<Explain the strengths and weaknesses of
                the medical knowledge in terms of accuracy
                and completeness, explicitly referencing
                the response content>"
}
```

F Ethical Considerations and Limitations

F.1 Data Usage and Compliance

We utilize publicly available datasets and models in strict adherence to their respective licenses and comply with the terms of service of all third-party APIs used. All resources are employed solely for academic research and benchmarking purposes.

F.2 Risks and Safety

Although RIPT enhances reasoning capabilities, it does not guarantee factual correctness and may generate fluent but erroneous medical explanations. Such outputs must not be treated as medical advice. We emphasize that RIPT is not intended for clinical deployment or real-world decision-making without expert oversight.