Reinforcement Learning Improves Traversal of Hierarchical Knowledge in LLMs

Renfei Zhang
Simon Fraser University
rza104@sfu.ca

Manasa Kaniselvan
FAIR at Meta
ETH Zurich
mkaniselvan@iis.ee.ethz.ch

Niloofar Mireshghallah
FAIR at Meta
nmireshg@andrew.cmu.edu

Abstract

Reinforcement learning (RL) is often credited with improving language model reasoning and generalization at the expense of degrading memorized knowledge. We challenge this narrative by observing that **RL-enhanced models consistently outperform their base and supervised fine-tuned (SFT) counterparts on pure knowledge recall tasks**, particularly those requiring traversal of hierarchical, structured knowledge (e.g., medical codes). We hypothesize these gains stem not from newly acquired data, but from improved procedural skills in navigating and searching existing knowledge hierarchies within the model parameters. To support this hypothesis, we show that structured prompting—which explicitly guides SFTed models through hierarchical traversal—recovers most of the performance gap (reducing 24pp to 7pp on MedConceptsQA for DeepSeek-V3/R1). We further find that while prompting improves final-answer accuracy, RL-enhanced models retain superior ability to recall correct procedural paths on deep-retrieval tasks. Finally our layer-wise internal activation analysis reveals that while factual representations (e.g., activations for the statement "code 57.95 refers to urinary infection") maintain high cosine similarity between SFT and RL models, query representations (e.g., "what is code 57.95") diverge noticeably, indicating that **RL primarily transforms how models traverse knowledge rather than the knowledge representation itself**.

1 Introduction

Large Language Models (LLMs) acquire vast parametric knowledge during pretraining, encoding facts, concepts, and their relationships across billions of parameters. Post-training techniques—including supervised fine-tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and specialized reasoning-focused RL—are then applied to transform these base models into instruction-following agents capable of complex reasoning [49, 43, 6]. While these methods improve performance on reasoning benchmarks and user preference metrics, a growing body of evidence reveals a concerning trade-off known as the "alignment tax" [27, 5, 39]: models sacrifice factual memorization capabilities to optimize for other objectives, leading to reduced performance on knowledge-intensive benchmarks [50, 13]. However, existing work has primarily focused on direct factual recall tasks over unstructured knowledge, leaving a critical gap: *do these degradation patterns hold for all forms of parametric knowledge retrieval tasks?*

To address this question, we investigate tasks where retrieval demands navigating hierarchical structures encoded within the model's parameters. Consider medical code lookup (Figure 1): to identify that ICD-9-CM code 57.95 refers to "Replacement of indwelling urinary catheter," a model can attempt direct recall—often failing due to the vast code space—or systematically traverse the taxonomy (Chapter 11 → codes 57.0-57.99 → specific procedure). Surprisingly, reasoning-enhanced models outperform their base counterparts by 24 percentage points on MedConceptsQA, directly

challenging the conventional wisdom that RL sacrifices memorization for reasoning [14, 11]. We hypothesize these models succeed through systematic hierarchical navigation rather than direct recall, proposing that **reinforcement learning enhances navigation of existing parametric knowledge rather than adding new factual content**.

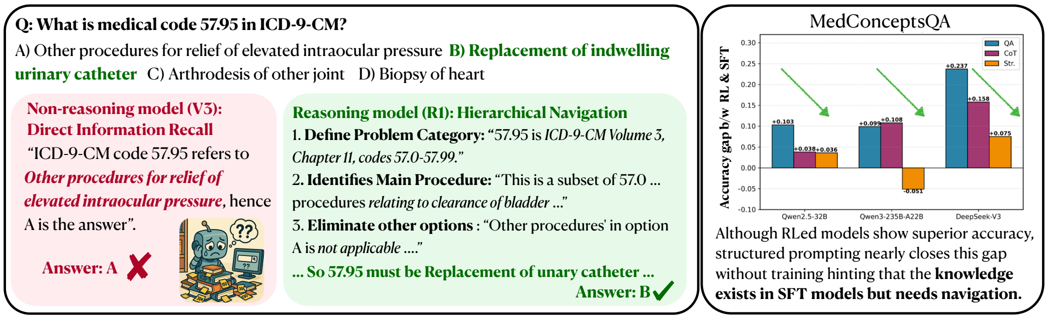


Figure 1: (Left) Overview of our main observation: When querying structured medical codes, non-reasoning models (DeepSeek-V3) rely on direct memorization attempts, often selecting incorrect answers (here choosing A). In contrast, reasoning-enhanced RL models (DeepSeek-R1) employ systematic hierarchical navigation—first categorizing the problem domain, then identifying relevant procedures, and finally interpreting ambiguous terminology—to successfully retrieve the correct answer (B). (Right) Reasoning models consistently outperform their instruction-tuned counterparts when prompted with conventional QA templates. This gap decreases when we optimize the prompt and is minimized with our hand-crafted structured prompt, hinting that the necessary knowledge exists in the instruct models.

To disentangle knowledge acquisition from navigation, we design three complementary experiments. First, inspired by work showing prompt optimization can match RL gains [4, 21, 53], we develop structured prompting that explicitly guides base models through hierarchical traversal. If knowledge exists in base models, prompting should surface it. **Structured prompting reduces the 24pp gap between DeepSeek-V3 and DeepSeek-R1 to 7pp, suggesting information is present but inaccessible without proper navigation** (Figure 1, right-hand side). Second, to validate that improved traversal drives these gains, we introduce a complexity-stratified patent classification dataset and Path Matching Score metric measuring traversal accuracy. We show that as recall depth increases (from fewer than 3 hops to more than 5), **reasoning models demonstrate superior path recall accuracy**, with the performance gap widening from 5pp to 9pp, demonstrating that reasoning models excel at complex hierarchical navigation (Table 5).

Third, to provide internal validation, we conduct layer-wise representational analysis inspired by work examining how post-training modifies internal model structure [31, 38, 18, 2]. We extract layer-wise representations for matched query-answer pairs, comparing interrogative queries (e.g., "What is the medical code 57.95?") versus declarative statements (e.g., "Code 57.95 refers to urinary catheter replacement"). We find a striking pattern (Figure 3): declarative statements maintain high cosine similarity (0.85-0.92) between base and RL models throughout most layers, while interrogative queries diverge substantially (similarity dropping to 0.65-0.73 in middle layers). This asymmetry reveals that **RL and instruction tuning primarily transforms how models process *questions* while leaving factual knowledge representations intact**, consistent with our hypothesis that RL enhances navigation mechanisms rather than knowledge content.

We further conduct ablation studies comparing distilled R1 models to R1 and base models [22, 9], finding that distilled models capture only surface-level improvements without acquiring robust navigation capabilities—achieving intermediate performance on complex retrieval tasks. Structured prompting provides minimal gains for distilled models, and layer-wise analysis reveals greater representational changes than instruction-tuned variants, yet without improved deep-retrieval navigation.

Our findings carry important implications: RL-enhanced models succeed not through expanded knowledge but through improved cognitive scaffolding—the ability to systematically traverse structures already encoded during pretraining, which is inline with recent work showing that RL surfaces intelligence [18, 45]. While our experiments focus on two datasets (MedConceptsQA and IPC) and specific model families (Qwen2.5, DeepSeek, Mistral), the patterns suggest more efficient training paradigms separating knowledge acquisition (pretraining) from organization (post-training). We encourage future work to investigate these phenomena across broader domains and develop RL mechanisms that explicitly optimize for hierarchical navigation.

2 Experimental Methodology

Our investigation into how reinforcement learning enhances hierarchical knowledge traversal is guided by three research questions:

Research Questions

1. **RQ1: Does explicit prompting close the performance gap?** If instruction-tuned models contain the required knowledge, can structured prompts that explicitly instruct hierarchical traversal match the performance of RL-enhanced models?
2. **RQ2: Do reasoning models navigate deeper hierarchies better?** On tasks requiring multi-step hierarchical traversal, do reasoning models demonstrate superior path accuracy beyond what prompting achieves?
3. **RQ3: How do internal representations differ?** Do reasoning models transform how they encode queries, factual knowledge, or both?

We address these questions through three complementary experiments. Section 2.1 demonstrates that structured prompting can induce hierarchical reasoning in instruction-tuned models, reducing the performance gap by up to 68%. Section 2.2 introduces retrieval tasks of varying complexity with a path matching metric, revealing that reasoning models excel particularly on deep-retrieval tasks requiring extensive hierarchical navigation. Section 2.3 presents layer-wise activation analysis showing that while factual representations remain largely unchanged, query processing diverges substantially between SFT and RL models, supporting our hypothesis that RL primarily enhances navigation mechanisms rather than knowledge content.

2.1 Hierarchical Navigation Through Structured Prompting

We investigate tasks requiring pure information recall without multi-step computation or logical deduction, to determine whether the performance gap between base and reasoning models can be mitigated through prompting strategies alone. Remarkably, structured prompting reduces the performance gap for 671B base models such as DeepSeek-V3 from 23.7 pp to 7.5 pp (a 68% gap reduction), demonstrating the effectiveness of our method.

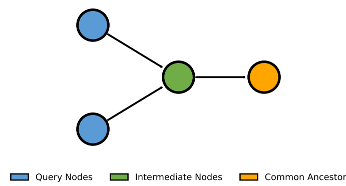
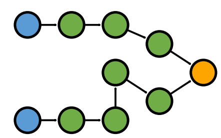
Datasets

- **MedConceptsQA:** A multiple-choice question answering dataset focused on biomedical and clinical concepts. The questions are designed to test factual recall of medical terminology, concept definitions, and their relationships, without reasoning over patient cases or performing calculations.
- **International Patent Classification (IPC):** A dataset consists of queries mapped to patent classification codes. The task requires identifying the correct category for a given technical description, relying on recalling standardized knowledge of patent domains rather than multi-step reasoning.

Prompting

- **Direct Question-Answering (QA) Prompting**: This baseline requires the model to provide only a single-letter answer to each multiple-choice question without any explanation.
- **Standard Chain-of-Thought (CoT) Prompting**: This template requests both a final answer and a supporting explanation, aiming to capture the model's intrinsic reasoning without imposing any procedural constraints.
- **Structured Prompting**: We introduce hierarchical instructions that enforce systematic reasoning. This strategy involves a two-stage process: (1) recall the hierarchical structural

Table 1: Stratification of the "Nearest Common Ancestor" task by retrieval complexity, defined by the number of unique ancestor nodes (traversals) recalled to find the common node.

Task Complexity	Traversals	Figure Example	Example
Memory-Light	< 3		**Question:** Nearest common ancestor of H04B 1/7075 and H04B 1/7083 is: A) H04B 1/707 B) H04B 1/7073 C) H04B 1/69 D) H04B **Hierarchical Paths:** • H04B 1/7075 → H04B 1/7073 • H04B 1/7083 → H04B 1/7073 **Answer:** B
Memory-Heavy	5+		**Question:** Nearest common ancestor of A01B 3/421 and A01B 15/06 is: A) A01B 3/00 B) A01B 15/00 C) A01B D) A01 **Hierarchical Paths:** • A01B 3/421 → A01B 3/42 → A01B 3/40 → A01B 3/36 → A01B 3/00 → A01B • A01B 15/06 → A01B 15/04 → A01B 15/02 → A01B 15/00 → A01B **Answer:** C

breakdown of the relevant medical code or concept, and (2) systematically evaluate each option with justification before elimination. This approach tests our hypothesis that enforcing structured knowledge recall and stepwise elimination can reduce performance gaps (see Appendix B.1 for complete prompt templates).

Models We evaluate a diverse set of large language models, focusing on comparisons between base, instruction-tuned, reasoning, and distilled models. The first group includes instruction-tuned models, such as the Qwen2.5 family (7B, 14B, 32B, and 72B parameters) [40] and Mistral-Small-3.1-24B-Instruct [20], each paired with their respective base models. The second group consists of reasoning models, including QwQ-32B (reasoning-enhanced Qwen2.5-32B), DeepSeek-R1 (from DeepSeek-V3), Magistral (from Mistral-Small-3.1-24B), and the reasoning model of Qwen3-235B-A22B [28, 16, 46]. The third group includes models distilled from DeepSeek-R1: Qwen2.5-Math-7B, Qwen2.5-32B, and Llama3.3-70B, each compared against their pre-distillation ones [15]. We sample from all models using a temperature of 0.8 and top-p of 0.7 across three independent runs. Performance is reported as both mean accuracy (± standard deviation) and majority-voted accuracy, where majority voting selects the most frequent answer among the three runs for each question.

2.2 Hierarchical Navigation Across Retrieval Complexity

While we previously conclude that reasoning models use hierarchical navigation that can be externalized through structured prompting, a fundamental question remains: *do reasoning models merely execute these strategies more consistently, or are there tasks that they execute fundamentally better?* To address this, we need to analyze not just whether models retrieve correct answers but how they traverse knowledge hierarchies to reach those answers. Therefore, we extend the original IPC dataset to stratify it by retrieval complexity and introduce a new metric to measure path traversal quality. Subsequent results reveal that reasoning models show superior hierarchical traversal-an ability that emerges on complex tasks requiring deeper knowledge navigation.

IPC Multi-Level Retrieval Dataset As shown in Table 1, this expanded dataset tests basic structural knowledge, including identifying common ancestors of a given pair of nodes. The questions are categorized by retrieval complexity, defined as the total number of ancestor nodes that must be recalled along both hierarchical paths (excluding the initial query nodes) to reach the nearest common ancestor. This stratification allows us to isolate the effect of retrieval depth on model performance.

- **Memory-Light (ML)** tasks require recalling < 3 ancestor nodes total across both paths to reach the common ancestor.
- **Memory-Heavy (MH)** tasks demand recalling ≥ 5 ancestor nodes across both paths.

Table 2: Performance comparison of Instruct vs. Reasoning models on MedConceptsQA and IPC datasets. The first column indicates the dataset. Models are evaluated across three prompt templates (QA, CoT, Structured). Metrics shown are majority voting accuracy (Maj. Vote Acc.) and mean accuracy (Mean Acc.). Mean accuracy is reported as Mean Acc. $_{\text{(Std.)}}$, with the standard deviation in subscripted parentheses. For each model pair, a Δ row shows the gap from the reasoning model for both Maj. Acc. (red) and Mean Acc. (green). Bold values indicate the best performance within each model pair. Δ values are highlighted, with darker shades indicating larger gaps.

Dataset	Model	Model Type	Maj. Vote Acc.			Mean Acc.$_{\text{(Std.)}}$		
			QA	CoT	Structured	QA	CoT	Structured
MedConceptsQA	Qwen2.5-32B	Instruct	0.379	0.475	0.469	0.371$_{(.012)}$	0.449$_{(.010)}$	0.454$_{(.007)}$
		Reasoning	**0.482**	**0.513**	**0.505**	**0.470$_{(.012)}$**	**0.487$_{(.009)}$**	**0.481$_{(.005)}$**
		Δ	**+0.103**	**+0.038**	**+0.036**	**+0.099**	**+0.038**	**+0.027**
	Qwen3-235B-A22B	Instruct	0.542	0.548	**0.631**	0.503$_{(.004)}$	0.528$_{(.005)}$	**0.589$_{(.007)}$**
		Reasoning	**0.641**	**0.656**	0.580	**0.599$_{(.003)}$**	**0.617$_{(.003)}$**	0.554$_{(.008)}$
		Δ	**+0.099**	**+0.108**	**-0.051**	**+0.096**	**+0.089**	**-0.035**
	DeepSeek-V3	Instruct	0.541	0.632	0.717	0.551$_{(.014)}$	0.636$_{(.049)}$	0.701$_{(.026)}$
		Reasoning	**0.778**	**0.790**	**0.792**	**0.830$_{(.006)}$**	**0.774$_{(.013)}$**	**0.775$_{(.026)}$**
		Δ	**+0.237**	**+0.158**	**+0.075**	**+0.279**	**+0.138**	**+0.074**
IPC Codes	Qwen2.5-32B	Instruct	0.759	0.754	0.774	0.759$_{(.007)}$	0.754$_{(.000)}$	0.774$_{(.007)}$
		Reasoning	**0.777**	**0.875**	**0.790**	**0.713$_{(.015)}$**	**0.754$_{(.070)}$**	**0.769$_{(.033)}$**
		Δ	**+0.018**	**+0.121**	**+0.016**	**-0.046**	**+0.000**	**-0.005**
	Qwen3-235B-A22B	Instruct	0.800	**0.846**	0.846	0.800$_{(.013)}$	**0.846$_{(.013)}$**	0.846$_{(.013)}$
		Reasoning	**0.908**	0.877	**0.893**	**0.846$_{(.013)}$**	0.836$_{(.026)}$	**0.851$_{(.015)}$**
		Δ	**+0.108**	**+0.031**	**+0.047**	**+0.046**	**-0.010**	**+0.005**
	DeepSeek-V3	Instruct	0.831	**0.923**	0.877	0.846$_{(.000)}$	**0.882$_{(.007)}$**	0.872$_{(.007)}$
		Reasoning	**0.923**	0.892	**0.923**	**0.913$_{(.019)}$**	0.867$_{(.026)}$	**0.903$_{(.007)}$**
		Δ	**+0.092**	**-0.031**	**+0.046**	**+0.067**	**-0.015**	**+0.031**

Path Matching Score To evaluate the quality of predicted hierarchical paths for IPC codes, we propose the path matching score, which combines two metrics:

- **F1-Score:** Measures precision and recall of hierarchical ancestor identification, defined as $F_1 = \frac{2 \times P \times R}{P + R}$, where P and R denote precision and recall over the set of hierarchical ancestors [8].
- **Common Subsequence Score (CSS):** Evaluates structural integrity of sequential paths via the ratio of the Longest Common Subsequence (LCS) [33] between the predicted and true paths to the length of the true path: $\text{CSS} = \frac{|\text{LCS}(\text{predicted}, \text{ground truth})|}{|\text{ground truth ancestors}|}$.

The path matching score combines both components via harmonic mean: Path Matching $= \frac{2 \times F_1 \times \text{CSS}}{F_1 + \text{CSS}}$. This metric captures both structural accuracy and hierarchical coherence in patent classification navigation.

Models To analyze the impact of retrieval complexity, we conduct a case study using the DeepSeek-V3 and R1 pair on our expanded IPC dataset. While a broader evaluation would be ideal, we select the DeepSeek pair due to their instruction-following capabilities suitable for a reliable analysis.

2.3 Hierarchical Navigation in Internal Representations

To investigate whether base and specialized models[1] possess equivalent knowledge for hierarchical reasoning, we analyze their internal activations on MedConceptsQA using contrastive question-answer pairs. We conduct two complementary analyses: an *inter-model* comparison to show how enhancement modifies representations relative to the base model, and an *intra-model* comparison to trace how individual models transform questions into answers across layers. Our findings show that enhancement refines query processing while preserving factual knowledge.

Probe Construction. We construct probes from the MedConceptsQA dataset, which spans five medical vocabularies: ATC, ICD9CM, ICD10CM, ICD9PROC, and ICD10PROC. To ensure balanced representation, we randomly sample 100 question-answer pairs from each vocabulary. Each probe

[1] Here "base models" refer to the foundation model from which "specialized models" (instruction-tuned/reasoning/distilled) variants are derived. We adopt this terminology throughout the section to clearly distinguish the two categories.

Table 3: Performance comparison of Base vs. Instruct models on MedConceptsQA and IPC datasets. The first column indicates the dataset. Models are evaluated across three prompt templates (QA, CoT, Structured). Metrics shown are majority voting accuracy (Maj. Vote Acc.) and mean accuracy (Mean Acc.). Mean accuracy is reported as Mean $\text{Acc.}_{(\text{Std.})}$, with the standard deviation in subscripted parentheses. For each model pair, an Δ row shows the gap from the instruct model for both Maj. Acc. (red) and Mean Acc. (green). Bold values indicate the best performance within each model pair. Δ values are highlighted, with darker shades indicating larger gaps. **This gap shrinks as we optimize the prompt, showing that the knowledge exists in the instruct model, it just needs to surface.**

Dataset	Model	Model Type	Maj. Vote Acc.			Mean $\text{Acc.}_{(\text{Std.})}$		
			QA	CoT	Structured	QA	CoT	Structured
MedConceptsQA	Qwen2.5-7B	Base	0.148	0.277	0.286	$0.159_{(.007)}$	$0.239_{(.036)}$	$0.270_{(.012)}$
		Instruct	**0.295**	**0.329**	**0.313**	$\mathbf{0.289_{(.006)}}$	$\mathbf{0.316_{(.008)}}$	$\mathbf{0.307_{(.015)}}$
		Δ	**+0.147**	**+0.052**	**+0.027**	**+0.130**	**+0.077**	**+0.037**
	Qwen2.5-14B	Base	0.335	0.332	0.386	$0.316_{(.015)}$	$0.293_{(.025)}$	$0.372_{(.007)}$
		Instruct	**0.395**	**0.420**	**0.420**	$\mathbf{0.385_{(.006)}}$	$\mathbf{0.415_{(.007)}}$	$\mathbf{0.409_{(.012)}}$
		Δ	**+0.060**	**+0.088**	**+0.034**	**+0.069**	**+0.122**	**+0.037**
	Qwen2.5-32B	Base	0.221	0.332	0.404	$0.219_{(.012)}$	$0.260_{(.071)}$	$0.372_{(.007)}$
		Instruct	**0.379**	**0.475**	**0.469**	$\mathbf{0.371_{(.012)}}$	$\mathbf{0.449_{(.010)}}$	$\mathbf{0.454_{(.007)}}$
		Δ	**+0.158**	**+0.143**	**+0.065**	**+0.152**	**+0.189**	**+0.082**
	Qwen2.5-72B	Base	0.443	0.351	0.468	$0.389_{(.005)}$	$0.305_{(.028)}$	$0.418_{(.008)}$
		Instruct	**0.546**	**0.520**	**0.546**	$\mathbf{0.519_{(.007)}}$	$\mathbf{0.512_{(.005)}}$	$\mathbf{0.537_{(.008)}}$
		Δ	**+0.103**	**+0.169**	**+0.078**	**+0.130**	**+0.207**	**+0.119**
IPC Codes	Qwen2.5-7B	Base	0.463	0.436	**0.588**	$0.349_{(.040)}$	$0.364_{(.038)}$	$\mathbf{0.585_{(.038)}}$
		Instruct	**0.615**	**0.554**	0.574	$\mathbf{0.615_{(.025)}}$	$\mathbf{0.554_{(.013)}}$	$0.574_{(.015)}$
		Δ	**+0.152**	**+0.118**	**-0.014**	**+0.266**	**+0.190**	**-0.011**
	Qwen2.5-14B	Base	0.526	0.608	0.609	$0.421_{(.038)}$	$0.492_{(.033)}$	$0.600_{(.013)}$
		Instruct	**0.708**	**0.691**	**0.718**	$\mathbf{0.708_{(.025)}}$	$\mathbf{0.687_{(.029)}}$	$\mathbf{0.718_{(.007)}}$
		Δ	**+0.182**	**+0.083**	**+0.109**	**+0.287**	**+0.195**	**+0.118**
	Qwen2.5-32B	Base	0.644	0.641	**0.777**	$0.482_{(.059)}$	$0.503_{(.038)}$	$0.769_{(.013)}$
		Instruct	**0.759**	**0.754**	0.774	$\mathbf{0.759_{(.007)}}$	$\mathbf{0.754_{(.000)}}$	$\mathbf{0.774_{(.007)}}$
		Δ	**+0.115**	**+0.113**	**-0.003**	**+0.277**	**+0.251**	**+0.005**

Table 4: Performance of distilled models compared to the **DeepSeek-R1 (reasoning model)**. Each cell for a distilled model shows its absolute score, followed in parentheses by the Δ gap (reasoning - distilled). Δ values for Maj. Vote Acc. are shaded red, and Δ values for Mean Acc. are shaded green. Darker shades indicate a larger performance gap. All Δ values are positive, showing the gap to the stronger R1 model.

Dataset	Model	Maj. Vote Acc. (Δ vs. R1)			Mean $\text{Acc.}_{(\text{Std.})}$ (Δ vs. R1)		
		QA	CoT	Structured	QA	CoT	Structured
MedConceptsQA	**DeepSeek-R1 (Reasoning)**	**0.778**	**0.790**	**0.792**	$\mathbf{0.830_{(.006)}}$	$\mathbf{0.774_{(.013)}}$	$\mathbf{0.775_{(.026)}}$
	Qwen2.5-Math-7B (Dist.)	0.296 (**+0.482**)	0.256 (**+0.534**)	0.282 (**+0.510**)	$0.292_{(.010)}$ (**+0.538**)	$0.250_{(.017)}$ (**+0.524**)	$0.289_{(.017)}$ (**+0.486**)
	Qwen2.5-32B (Dist.)	0.375 (**+0.403**)	0.380 (**+0.410**)	0.447 (**+0.345**)	$0.351_{(.009)}$ (**+0.479**)	$0.369_{(.005)}$ (**+0.405**)	$0.420_{(.002)}$ (**+0.355**)
	Llama3.3-70B (Dist.)	0.537 (**+0.241**)	0.633 (**+0.157**)	0.610 (**+0.182**)	$0.495_{(.002)}$ (**+0.335**)	$0.609_{(.011)}$ (**+0.165**)	$0.596_{(.012)}$ (**+0.179**)
IPC Codes	**DeepSeek-R1 (Reasoning)**	**0.923**	**0.892**	**0.923**	$\mathbf{0.913_{(.019)}}$	$\mathbf{0.867_{(.026)}}$	$\mathbf{0.903_{(.007)}}$
	Qwen2.5-32B (Dist.)	0.778 (**+0.145**)	0.730 (**+0.162**)	0.788 (**+0.135**)	$0.754_{(.038)}$ (**+0.159**)	$0.667_{(.019)}$ (**+0.200**)	$0.780_{(.019)}$ (**+0.123**)
	Llama3.3-70B (Dist.)	0.785 (**+0.138**)	0.831 (**+0.061**)	0.815 (**+0.108**)	$0.785_{(.015)}$ (**+0.128**)	$0.785_{(.041)}$ (**+0.082**)	$0.790_{(.018)}$ (**+0.113**)

consists of a factual question and its corresponding ground-truth answer, formatted as declarative statements. For example, a probe for medical code `0QD20Z` from `ICD10PROC` takes the following form:

> *Question:* What is the description of the medical code `0QD20Z` in `ICD10PROC`?
>
> *Answer:* The description of the medical code `0QD20Z` in `ICD10PROC` is extraction of right pelvic bone, open approach.

We process questions and answers independently through each model to extract their respective layer-wise representations, enabling both inter-model and intra-model comparative analyses.

Representation Extraction. For a model with L layers and hidden dimension d, we extract the hidden state at the final token position for each layer $\ell \in \{1, \ldots, L\}$ as the layer's representation vector $\mathbf{h}_\ell \in \mathbb{R}^d$. This representation attends to all preceding tokens, thereby capturing the full input context at that layer.

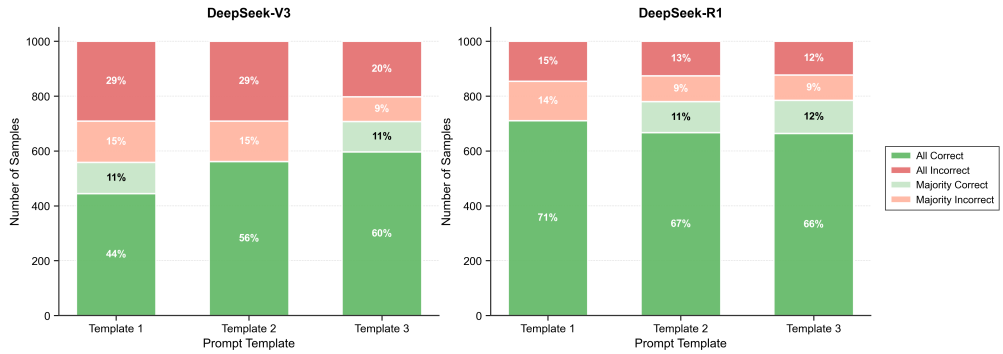


Figure 2: Comparative performance analysis of DeepSeek-V3 and DeepSeek-R1 across prompt strategies: direct question-answering (Template 1), chain-of-thought (Template 2), and structured prompting (Template 3) on MedConceptsQA dataset. Four categories are defined based on the number of correct votes across three independent runs: "All Incorrect" (0/3 correct), "Majority Incorrect" (1/3 correct), "Majority Correct" (2/3 correct), and "All Correct" (3/3 correct).

Table 5: Comparison of structured prompting performance by task complexity for DeepSeek-R1 and DeepSeek-V3 models. Memory-Light tasks (1-2 hierarchical recalls); Memory-Heavy tasks (5+ hierarchical recalls). Bold values indicate the best performance for each metric within each complexity category. **As we move to retrieval heavier tasks with structure, the gap between path matching score of R1 and V3 increases.**

Task Complexity	Model	Accuracy (%)	Path Matching Score
Memory-Light	DeepSeek-R1	**44.8**	**0.681**
	DeepSeek-V3	37.9	0.627
Memory-Heavy	DeepSeek-R1	**67.7**	**0.597**
	DeepSeek-V3	**67.7**	0.503

Representation Analysis. We quantify representational differences across and within models using *inter-model* and *intra-model* analyses:

- **Inter-Model (Q-Q / A-A) Analysis.** By comparing the question–question (Q–Q) and answer–answer (A–A) representations between the base and specialized models, we assess how they differ at understanding query and retrieving factual knowledge.
- **Intra-Model (Q-A) Comparison.** This analysis investigates the internal transformation of information within a single model. By comparing a model's question and answer representations layer by layer, we trace how internal activations evolve from encoding a problem to producing a solution.

Comparison Metric For each layer $\ell \in \{1, \ldots, L\}$, we use cosine similarity, a measure of directional alignment, to define representation similarity:

$$d_{\text{cos}}^{(a,b)}(\ell) = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{h}_\ell^{(a)}(i)^\top \mathbf{h}_\ell^{(b)}(i)}{\|\mathbf{h}_\ell^{(a)}(i)\|_2 \, \|\mathbf{h}_\ell^{(b)}(i)\|_2}, \tag{1}$$

Here, $\mathbf{h}_\ell^{(s)}(i)$ denotes the layer-ℓ hidden representation for probe i from a source s. The set of sources $\mathcal{S} = \{\text{Q}^{\text{base}}, \text{A}^{\text{base}}, \text{Q}^{\text{specialized}}, \text{A}^{\text{specialized}}\}$ includes representations for both the question (Q) and answer (A) components from the base and specialized models. Pair (a, b) represents either inter-model (e.g., Q^{base} vs $\text{Q}^{\text{specialized}}$, A^{base} vs $\text{A}^{\text{specialized}}$) or intra-model comparisons (e.g., Q^{base} vs A^{base}, $\text{Q}^{\text{specialized}}$ vs $\text{A}^{\text{specialized}}$). Results are reported per vocabulary using $N = 100$ probes.

Models We compare Qwen2.5-32B (base) against three specialized variants: Qwen2.5-32B-Instruct (instruction-tuned), DeepSeek-R1-Distill-Qwen-32B (distilled), and QwQ-32B (reasoning). We

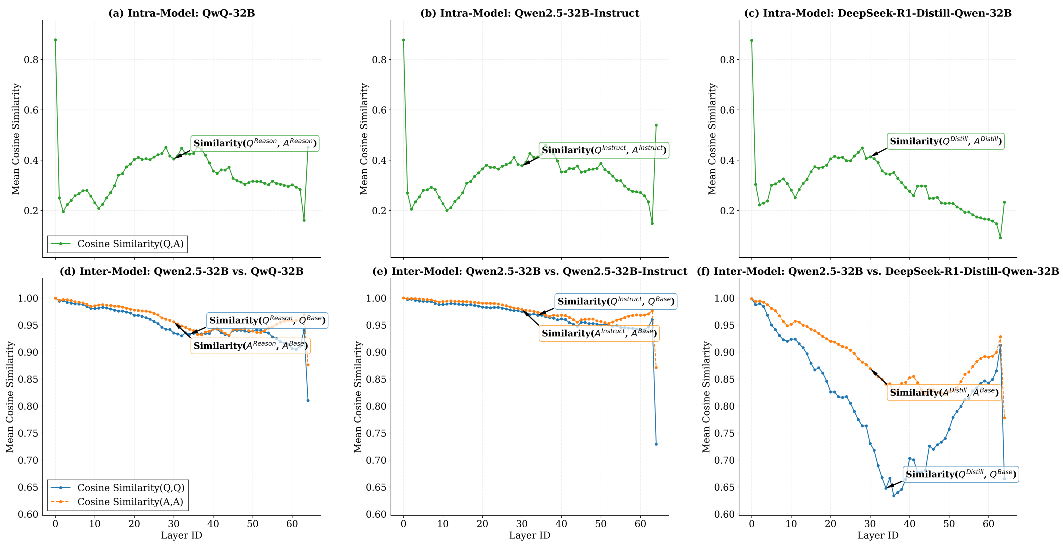


Figure 3: Layerwise Representation Similarity for ICD9PROC Vocabulary from MedConceptsQA. Plots compare last-token hidden state representations across layers (x-axis) using cosine similarity. Top Row (Intra-Model): Question vs. Answer representation similarity within QwQ-32B, Qwen2.5-32B-Instruct, and DeepSeek-R1-Distill-Qwen-32B. Bottom Row (Inter-Model): Similarity between the base model (Qwen2.5-32B) and each respective advanced model, comparing Question representations (Q_{Reason} vs. Q_{Base}) and Answer representations (A_{Reason} vs. A_{Base}) separately. **The representations of questions diverge more, specially in the last layer, compared to the answers. This hints at the knowledge being encoded similarly in base and reasoning models, but navigated differently.**

select this 32B parameter family because it spans multiple enhancement methods while remaining computationally tractable for single-GPU inference. A supplementary analysis comparing variants of the Mistral-Small-24B family (base, instruct, and reasoning) is included in Appendix C.

3 Experimental Results

3.1 Hierarchical Navigation Through Structured Prompting

Hierarchical navigation and stepwise elimination strategies systematically narrow the accuracy gap between base models and their reasoning-enhanced, or instruction-tuned versions across both MedConceptsQA and IPC code datasets. For example, on MedConceptsQA, structured prompting allows the Qwen3-235B Instruct model (Table 2) to outperform its reasoning counterpart, reversing a +0.108 majority vote accuracy gap (CoT) to a -0.051 advantage. Similarly, on the IPC dataset (Table 3), this prompting reduces the gap between the Qwen2.5-32B base and instruct models from +0.115 (QA) to -0.003. However, this effect is less pronounced for distilled models (Table 4), where the performance gap relative to the reasoning model remains substantial, even with structured prompts (e.g., Llama3.3-70B on MedConceptsQA, +0.182 gap).

To understand the mechanisms underlying structured prompting's effectiveness, we examine response consistency patterns. Figure 2 presents results for DeepSeek-V3 and R1 across three independent runs under majority voting on MedConceptsQA. When transitioning from the baseline to the structured prompt, DeepSeek-V3 shows significant sample migration: questions initially categorized as "All Incorrect" and "Majority Incorrect" shift toward "Majority Correct" and "All Correct". In contrast, R1 exhibits static distribution across these categories, suggesting it already operates near its ceiling. This redistribution in V3 indicates that explicit structural guidance improves the consistency of the model's internal reasoning and that its underlying knowledge is sufficient. Therefore, the primary role of specialized post-training is not to introduce entirely novel knowledge, but rather to enhance the procedural consistency and strategic reasoning of existing knowledge structures.

3.2 Hierarchical Navigation Across Retrieval Complexity

Stratifying performance by retrieval complexity highlights a distinction between the base and reasoning models. Despite similar overall accuracy, R1 consistently achieves a higher path matching score, particularly on complex tasks such as common ancestor identification, suggesting it can correctly navigate the hierarchy step-by-step (Table 5). This is a deeper form of understanding that goes beyond simple memorization. Ultimately, R1 understands the process of navigating a knowledge hierarchy better than the base model (V3), even when their final-answer accuracy is similar.

3.3 Hierarchical Navigation in Internal Representations

Intra-Model Representational Similarity. Within each model, representations for questions and answers are initially highly similar, but this similarity decreases in later layers, suggesting that the representations accumulate increasingly distinct features.

Inter-Model Representational Similarity. Instruction-tuned and reasoning models show strong directional alignment with the base model for both question and answer representations, whereas the distilled model shows much greater divergence (Figure 3(d-f)). Notably, question representations diverge more than answer representations across all specialized models, suggesting that performance gains arise primarily from refining question understanding rather than reorganizing factual knowledge.

4 Conclusion

This work challenges the view that reinforcement learning enhances reasoning at the expense of memorization. We demonstrate that RL-enhanced models outperform base counterparts by 24pp on hierarchical knowledge tasks, not through acquiring new knowledge, but by improving navigation of existing structures. Structured prompting reduces this gap to 7pp on simple tasks, yet reasoning models maintain superior path traversal on complex deep-retrieval tasks (5pp to 9pp gap widening). Layer-wise analysis reveals that RL transforms query processing (similarity drops to 0.65-0.73) while preserving factual representations (0.85-0.92), confirming that improvements stem from enhanced navigation mechanisms rather than knowledge content changes.

Several open questions warrant investigation. First, do similar navigation mechanisms underlie RL improvements on other structured reasoning tasks such as mathematical proof generation, code debugging, or multi-hop question answering? Second, can we develop RL objectives that explicitly optimize for hierarchical navigation rather than relying on implicit emergence? Third, how do these findings extend to knowledge domains with different structural properties—flat versus deeply nested hierarchies, dense versus sparse connectivity? Finally, can we design hybrid approaches that combine the efficiency of structured prompting with the robustness of RL-trained navigation for practical deployment? Addressing these questions will deepen our understanding of how language models organize and access parametric knowledge, ultimately enabling more capable and efficient reasoning systems.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shawn Guo, Chris Hallacy, Jesse

Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023. URL `https://arxiv.org/abs/2303.08774`.

[2] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

[3] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning, 2024. URL `https://arxiv.org/abs/2507.19457`. Note: Citation key was xu2024gepa but first author is Agrawal.

[4] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.

[5] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL `https://arxiv.org/abs/2112.00861`.

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal

Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022. URL `https://arxiv.org/abs/2212.08073`.

[7] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv preprint arXiv:2309.12288*, 2024.

[8] Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

[9] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.

[10] Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

[11] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[12] Matteo Gabburo, Nicolaas Paul Jedema, Siddhant Garg, Leonardo FR Ribeiro, and Alessandro Moschitti. Measuring retrieval complexity in question answering systems. *arXiv preprint arXiv:2406.03592*, 2024.

[13] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*, 2024.

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

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

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

[18] Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*, 2025.

[19] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[20] Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. Mistral–a journey towards reproducible language model training, 2021.

[21] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[22] Minwu Kim, Anubhav Shrestha, Safal Shrestha, Aadim Nepal, and Keith Ross. Reinforcement learning vs. distillation: Understanding accuracy and capability in llm reasoning. *arXiv preprint arXiv:2505.14216*, 2025.

[23] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

[24] Junliang Li, Yucheng Wang, Yan Chen, Yu Ran, Ruiqing Zhang, Jing Liu, Hua Wu, and Haifeng Wang. Knowledge-level consistency reinforcement learning: Dual-fact alignment for long-form factuality. *arXiv preprint arXiv:2509.23765*, 2025.

[25] Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, et al. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 2025.

[26] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL `https://arxiv.org/abs/2305.20050`.

[27] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

[28] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[29] Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. Hierarchical label-wise attention transformer model for explainable icd coding. *Journal of biomedical informatics*, 133:104161, 2022.

[30] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

[31] Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tür, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=0NdS4xCngO`.

[32] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, 2024.

[33] Mike Paterson and Vlado Dančík. Longest common subsequences. In *International symposium on mathematical foundations of computer science*, pages 127–142. Springer, 1994.

[34] Hoang Phan, Xianjun Yang, Kevin Yao, Jingyu Zhang, Shengjie Bi, Xiaocheng Tang, Madian Khabsa, Lijuan Liu, and Deren Lei. Beyond reasoning gains: Mitigating general capabilities forgetting in large reasoning models. *arXiv preprint arXiv:2510.21978*, 2025.

[35] Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. *arXiv preprint arXiv:2411.12580*, 2024.

[36] Cansu Sen, Bingyang Ye, Javed Aslam, and Amir Tahmasebi. From extreme multi-label to multi-class: A hierarchical approach for automated icd-10 coding using phrase-level attention. *arXiv preprint arXiv:2102.09136*, 2021.

[37] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. Curran Associates Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html`. arXiv:2303.11366.

[38] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

[39] Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Park, Jillian Fisher, Niloofar Mireshghallah, Liwei Jiang, and Yejin Choi. Spectrum tuning: Post-training for distributional coverage and in-context steerability. *arXiv preprint arXiv:2510.06084*, 2025.

[40] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2:3, 2024.

[41] Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

[42] Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. Emergent hierarchical reasoning in llms through reinforcement learning. *arXiv preprint arXiv:2509.03646*, 2025.

[43] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025.

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html`.

[45] Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

[46] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[47] Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *ArXiv*, abs/2502.06772, 2025. URL `https://api.semanticscholar.org/CorpusID:276250066`.

[48] Yufan Ye, Ting Zhang, Wenbin Jiang, and Hua Huang. Process-supervised reinforcement learning for code generation. *arXiv preprint arXiv:2502.01715*, 2025.

[49] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[50] Jiaqing Yuan, Lin Pan, Chung-Wei Hang, Jiang Guo, Jiarong Jiang, Bonan Min, Patrick Ng, and Zhiguo Wang. Towards a holistic evaluation of llms on factual knowledge recall. *arXiv preprint arXiv:2404.16164*, 2024.

[51] Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-STaR: Language models can teach themselves to think before speaking, 2024. URL `https://arxiv.org/abs/2403.09629`.

[52] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=WZH7099tgfM`.

[53] Noah Ziems, Dilara Soylu, Lakshya A Agrawal, Isaac Miller, Liheng Lai, Chen Qian, Kaiqiang Song, Meng Jiang, Dan Klein, Matei Zaharia, et al. Multi-module grpo: Composing policy gradients and prompt optimization for language model programs. *arXiv preprint arXiv:2508.04660*, 2025.

A Related Work

A.1 The Alignment Tax and Factual Degradation

The trade-off between alignment and factual accuracy has been extensively explored. Lin et al. [27] introduced the concept of "alignment tax", showing systematic degradation on factual benchmarks as RLHF reward strength increases. Achiam et al. [1] similarly reported that RLHF "does not improve exam performance (without active effort, it actually degrades it)" and can reduce calibration. Mechanistic analyses in Ghosh et al. [14] reveal that instruction tuning primarily adjusts style rather than new knowledge, with responses generated from pre-trained knowledge consistently outperforming those from models learning new knowledge through instruction tuning. Both Li et al. [24] and Kirk et al. [23] show that base models' parametric knowledge originates from pre-training while aligned models learn how to express it—training directly from base models mitigates knowledge forgetting and alignment tax incurred by SFT-based distillation. Recent work by Phan et al. [34] reveals that optimizing for narrow verifiable rewards in reasoning-focused RL leads to regression in general capabilities, with models exhibiting increased hallucinations despite improved reasoning.

While these studies document factual degradation from alignment, our work reveals a contrasting phenomenon: RL-enhanced models *outperform* their base counterparts on structured knowledge recall tasks. This apparent contradiction suggests that alignment tax may not uniformly affect all forms of parametric knowledge retrieval—particularly when retrieval demands systematic navigation through hierarchical structures rather than direct factual recall.

A.2 Reasoning Enhancement Through RL

RL is commonly viewed as a means of amplifying reasoning ability. Process supervision and reward-driven methods [26, 48] demonstrate clear improvements on reasoning tasks, with process-supervised models solving substantially more problems than outcome-supervised variants. However, recent work hints at a more nuanced picture. Zelikman et al. [51] introduce Quiet-STaR, showing that training models to generate internal rationales improves downstream reasoning by teaching systematic exploration of solution spaces—essentially navigation skills that achieve zero-shot improvements from 5.9% to 10.9% on GSM8K. Shinn et al. [37] demonstrate that reinforcement learning primarily helps models learn from feedback to refine their search through problem spaces, rather than acquiring new problem-solving rules. Most strikingly, Guo et al. [17] show that DeepSeek-R1 develops self-reflection, verification, and dynamic strategy adaptation through RL alone, without human-labeled reasoning trajectories, increasing pass@1 scores on AIME 2024 from 15.6% to 71.0%. Recent work on search-augmented reasoning [19, 10] demonstrates models learning to autonomously generate search queries and self-correct, with behaviors like pausing when detecting knowledge gaps emerging naturally.

These findings align with our hypothesis that RL enhances navigation of existing knowledge structures. However, while prior work focuses on mathematical and algorithmic reasoning, we examine whether these navigation improvements extend to retrieval from structured factual hierarchies, providing complementary evidence that RL's benefits stem from improved access patterns rather than new knowledge acquisition.

A.3 Hierarchical Reasoning and Structured Navigation

Hierarchical reasoning frameworks further support our knowledge navigation hypothesis. Wang et al. [41] present the Hierarchical Reasoning Model (HRM), a brain-inspired recurrent architecture that achieves near-perfect performance on complex tasks with only 27 million parameters trained on 1000 samples, without pre-training or chain-of-thought data. HRM's architecture features interdependent modules for high-level abstract planning and low-level detailed computation, achieving 40.3% on ARC-AGI—precisely the type of structured traversal we hypothesize enables medical code lookup. Yang et al. [47] show that hierarchical reinforcement learning on template sequences rather than long chain-of-thought data achieves 91.2% on MATH, outperforming models trained on detailed reasoning traces. Wang et al. [42] reveal RL training induces emergent separation between high-level strategic planning and low-level procedural execution, with two-phase learning of procedural consolidation followed by strategic exploration.

In the medical domain, structured approaches demonstrate substantial gains. Liao et al. [25] report that EHR-R1 achieves over 30 percentage points improvement on MIMIC-Bench (F1 of 0.6744 vs 0.3155 for GPT-4o) through graph-driven structured medical reasoning that converts raw EHR records into thinking graphs encoding temporal relations and causal hypotheses. Work on ICD code classification [29, 36] shows that leveraging hierarchical structure through label-wise attention and multi-class reformulation improves classification, particularly at higher hierarchy levels.

While these works demonstrate that hierarchical architectures and structured representations improve reasoning, they typically attribute gains to enhanced reasoning capabilities. Our work provides an alternative interpretation: these improvements may stem from better *navigation* of knowledge hierarchies already encoded during pretraining, rather than acquiring new reasoning abilities. We test this by showing that structured prompting—which explicitly guides traversal without modifying model parameters—recovers most performance gaps between base and RL models.

A.4 Prompting as an Alternative to RL

The possibility of achieving RL-like benefits through prompting has gained increasing attention. [3] demonstrate that Genetic-Evolution Prompt Alignment (GEPA) can outperform Group Relative Policy Optimization by up to 20% while using 35× fewer computational resources. They argue that "the interpretable nature of language provides a richer learning medium than sparse scalar rewards." [44] show that chain-of-thought prompting can match fine-tuned performance on reasoning tasks, while [52] demonstrate that optimized prompts can exceed supervised fine-tuning. The "Invisible Leash" phenomenon [45] reveals that much of RLHF's apparent benefit comes from teaching models to follow implicit formatting patterns—effects reproducible through prompting.

A.5 Knowledge Storage versus Knowledge Access

The distinction between knowledge acquisition and knowledge retrieval is crucial to our thesis. [32] show that models fine-tuned on new knowledge often "hallucinate" by incorrectly combining existing knowledge rather than storing new information. [35] provide key insights with their finding that models rely on procedural knowledge extracted from documents involving similar reasoning processes rather than memorizing new facts. This aligns with our hypothesis that RL enhances navigation strategies rather than expanding knowledge. [7] further support this through their "Reversal Curse" findings—models trained on "A is B" cannot infer "B is A," suggesting that training affects access patterns rather than creating bidirectional knowledge representations.

A.6 Retrieval Complexity in Knowledge-Intensive Tasks

Recent work has begun to to examine the relationship between retrieval complexity and model performance in knowledge-intensive tasks. [12] show that retrieval complexity extend beyond simple multi-hop reasoning—including temporal (15%), comparative (10%), and aggregate (16%) questions—suggesting that different types of knowledge organization require distinct retrieval strategies. [30] demonstrate that in long-form generation, factual accuracy in biographies drops as entity rarity increases, suggesting that retrieval difficulty directly impacts knowledge accessibility.

B Technical Appendices and Supplementary Material

B.1 Zero-Shot Prompt Templates

We present three prompt templates used in MedConceptsQA and IPC, which are designed to elicit specific responses from language models. These templates request:

- Direct answers, both with and without explanations.
- Structural recall of codes and a stepwise elimination of incorrect options.

Prompt Template 1: MCQ with Final Answer Only

Answer only A,B,C,D according to the answer to this multiple choice question.
[... Insert Question Text Here ...]
Answer (only the letter of your choice (A, B, C, or D)):

Prompt Template 2: MCQ with Explanation

You are a medical research assistant. Read the following multiple-choice question carefully. Your task is to:

1. Answer each question with one of A/B/C/D, which corresponds to the four options.
2. For my convenience, please give me a list of ANSWERs for the given instances in the format 'Answer: ...', with additional explanation for each answer in the format 'Explanation: ...'.

Respond in the following format:

```
Answer: <A/B/C/D>
Explanation: <your explanation here>
```

[... Insert Question Text Here ...]

Answer:
Explanation:

Prompt Template 3: MCQ with Stepwise Reasoning

You are a medical classification expert. For each option, first **recall the general category and structure breakdown of the medical code**, then explain **why it might be wrong**. Finally pick the correct one.

[... Insert Question Text Here ...]
Steps to follow:

1. Recall the general category and structural break down of the code.
2. Evaluate each option (A–D) briefly.
3. Choose the best option and justify.

Answer format:

```
Step 1: ...
Step 2A: ...
Step 2B: ...
Step 2C: ...
Step 2D: ...
Final Answer: [A/B/C/D] because ...
```

C Layer-wise Representation Analysis

C.1 Question-Answer Pairwise Probing

This section provides supplementary results for the layer-wise representation divergence analysis presented in Figure 3, extending the comparison across additional MedConceptsQA vocabularies for two model families.

C.1.1 Qwen2.5 Series

Figure 5 presents the analysis for the Qwen2.5-32B base model compared against its instruction-tuned (-Instruct), distilled (DeepSeek-R1-Distill-), and reasoning-enhanced (QwQ-32B) variants across the ATC, ICD10PROC, ICD9CM, and ICD10CM vocabularies.

C.1.2 Mistral-Small-24B Series

Figure 4 shows the corresponding analysis for the Mistral-Small model family, comparing the base (-Base-2503), instruction-tuned (-Instruct-2503), and reasoning-enhanced (Magistral-Small-2507) variants across all five MedConceptsQA vocabularies (ATC, ICD9PROC, ICD9CM, ICD10CM, ICD10PROC).

C.2 CoT Prompt Stepwise Probing

To analyze model representations under chain-of-thought (CoT) prompting, we construct a series of hierarchical prompts. For example, for the question "What is the description of the medical code 743.63 in ICD9CM?", the CoT series builds incrementally:

- "hmm let me think. 001-999.99 refers to diseases and injuries"
- "hmm let me think. 001-999.99 refers to diseases and injuries, and 740-759.99 refers to congenital anomalies"
- ...
- "hmm let me think. ... and 743.63 refers to other specified congenital anomalies of eyelid"

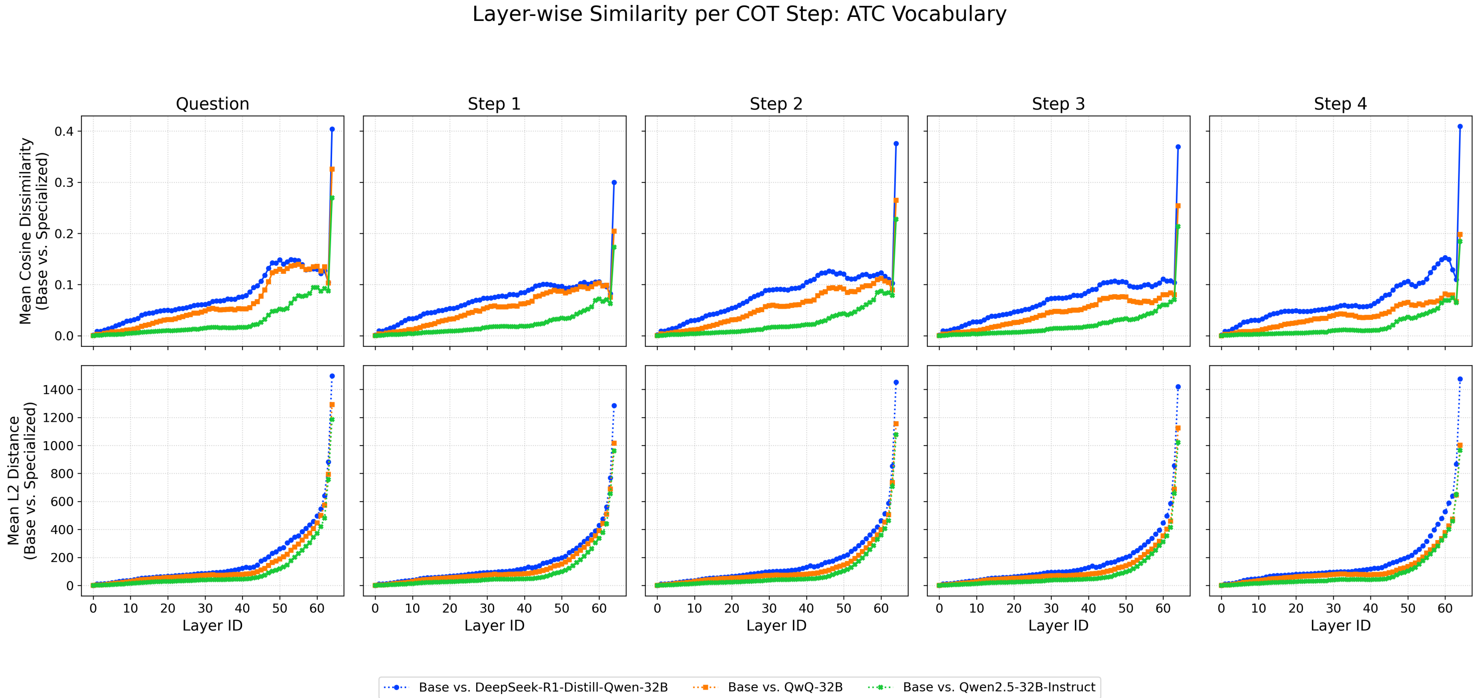
For each prompt in this series, we extract the activations from each layer of the model and group them by their corresponding vocabularies.

Additionally, we use **L2 distance** captures both directional and magnitude differences:

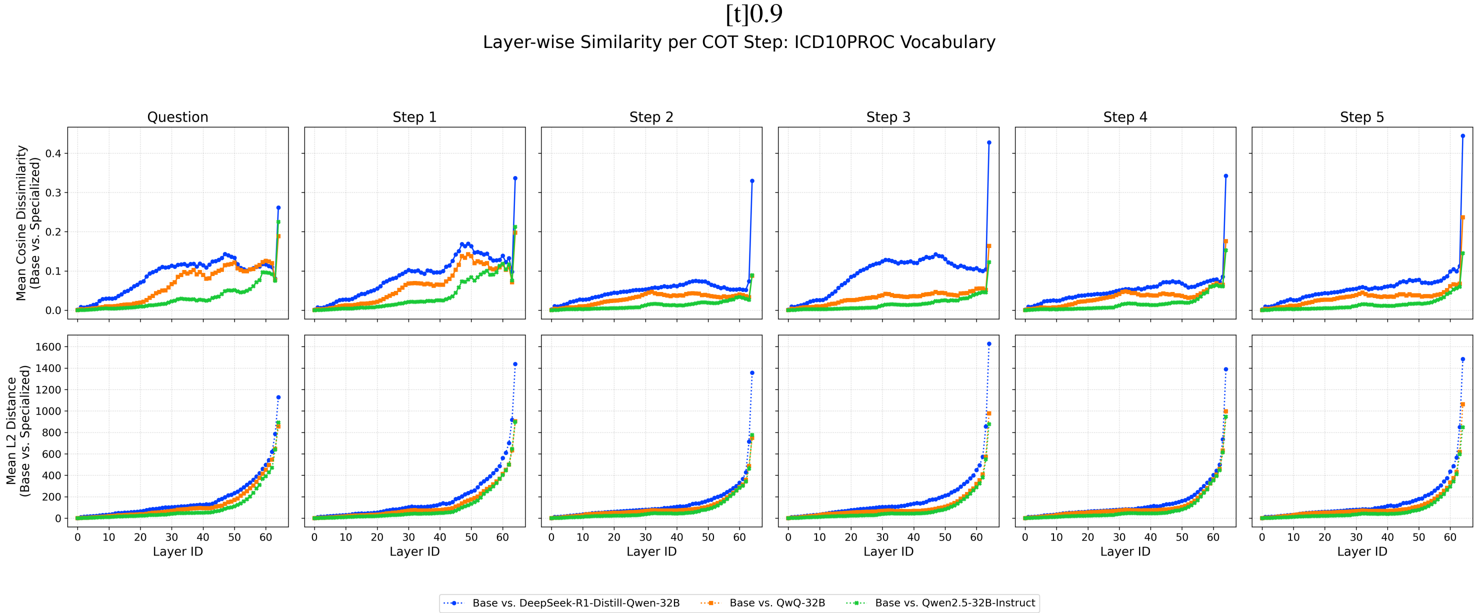
$$d_{\mathrm{L2}}^{(a,b)}(\ell) = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{h}_\ell^{(a)}(i) - \mathbf{h}_\ell^{(b)}(i) \right\|_2. \tag{2}$$

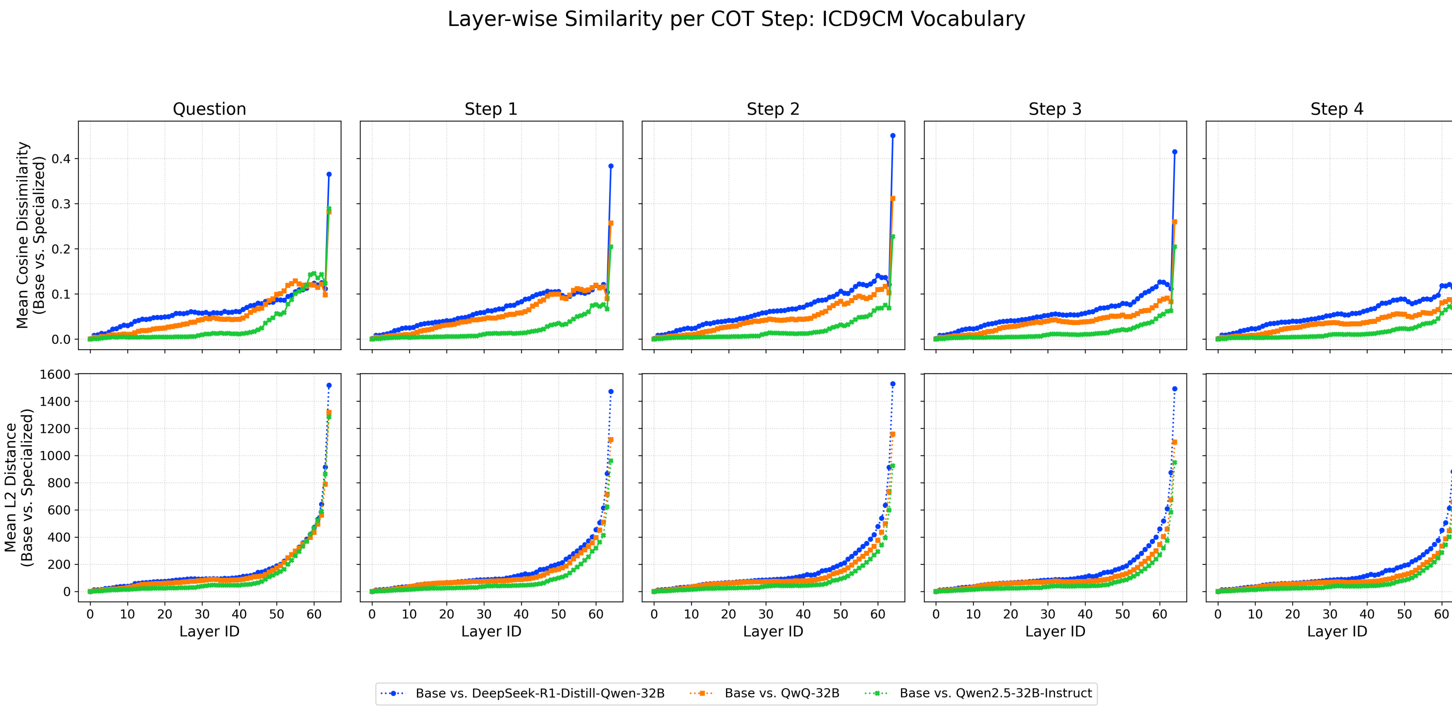
The number of CoT steps varies across vocabularies. To standardize this, we predefine all CoT sequences to be 5 steps long, with the exception of ICD10PROC, which uses 6 steps due to its more deeply embedded code structure (e.g., 0Q894Z). After grouping the activations by vocabulary for each layer, we compute the layerwise cosine similarity and L2 norm between the base and specialized models, following the methodology in Section 2.3.

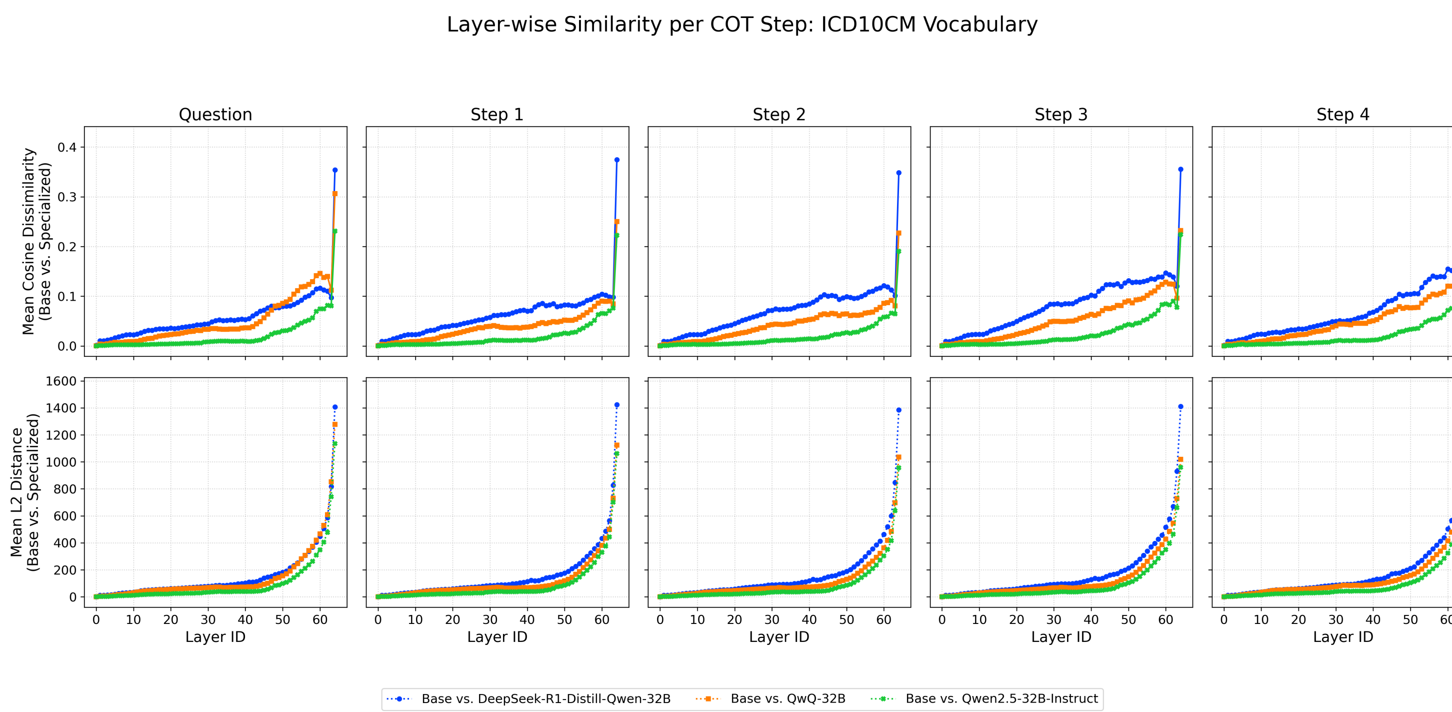
[t]0.9


Layer-wise Similarity per COT Step: ATC Vocabulary
Question
Step 1
Step 2
Step 3
Step 4
Mean Cosine Dissimilarity (Base vs. Specialized)
Mean L2 Distance (Base vs. Specialized)
Layer ID
Base vs. DeepSeek-R1-Distill-Qwen-32B
Base vs. QwQ-32B
Base vs. Qwen2.5-32B-Instruct

[t]0.9


Layer-wise Similarity per COT Step: ICD10PROC Vocabulary
Question
Step 1
Step 2
Step 3
Step 4
Step 5
Mean Cosine Dissimilarity (Base vs. Specialized)
Mean L2 Distance (Base vs. Specialized)
Layer ID
Base vs. DeepSeek-R1-Distill-Qwen-32B
Base vs. QwQ-32B
Base vs. Qwen2.5-32B-Instruct


Layer-wise Similarity per COT Step: ICD9CM Vocabulary
Question
Step 1
Step 2
Step 3
Step 4
Mean Cosine Dissimilarity (Base vs. Specialized)
Mean L2 Distance (Base vs. Specialized)
Layer ID
Base vs. DeepSeek-R1-Distill-Qwen-32B
Base vs. QwQ-32B
Base vs. Qwen2.5-32B-Instruct


Layer-wise Similarity per COT Step: ICD10CM Vocabulary
Question
Step 1
Step 2
Step 3
Step 4
Mean Cosine Dissimilarity (Base vs. Specialized)
Mean L2 Distance (Base vs. Specialized)
Layer ID
Base vs. DeepSeek-R1-Distill-Qwen-32B
Base vs. QwQ-32B
Base vs. Qwen2.5-32B-Instruct

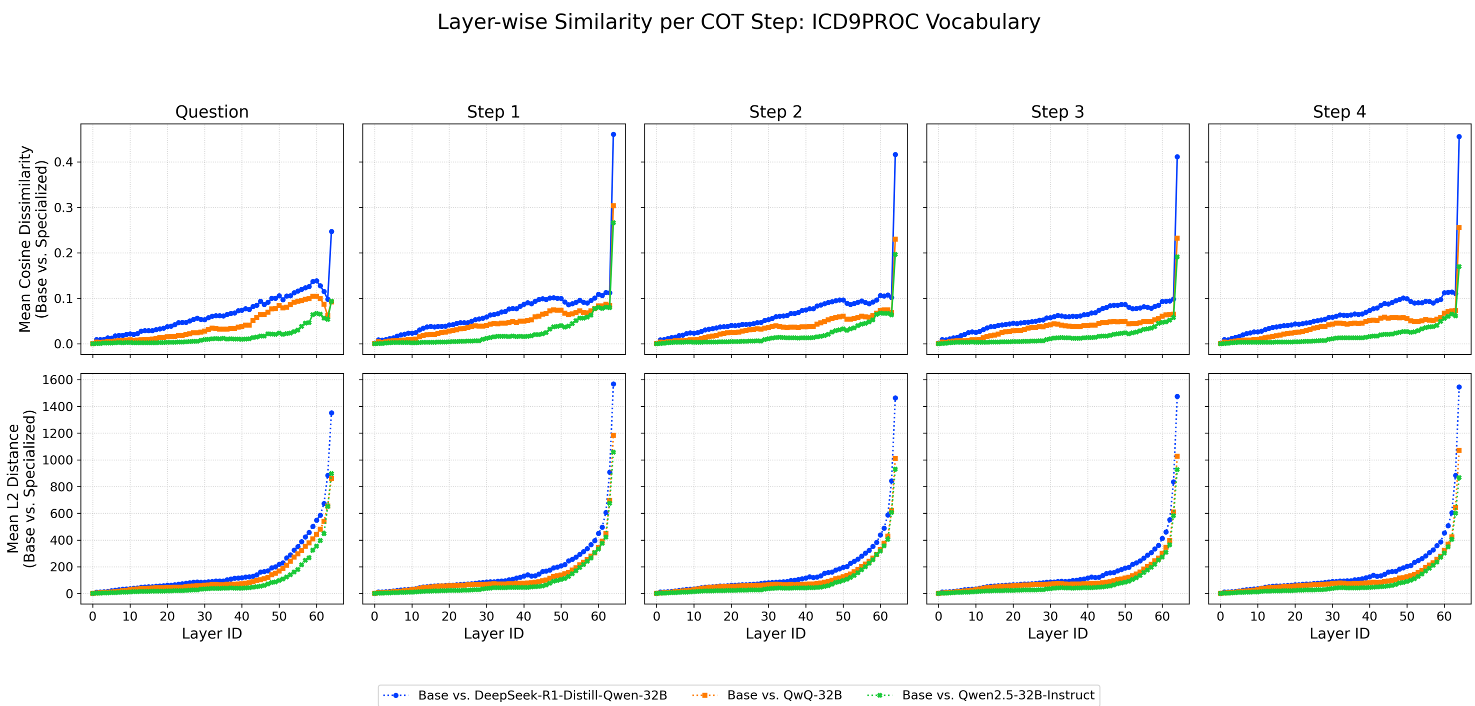
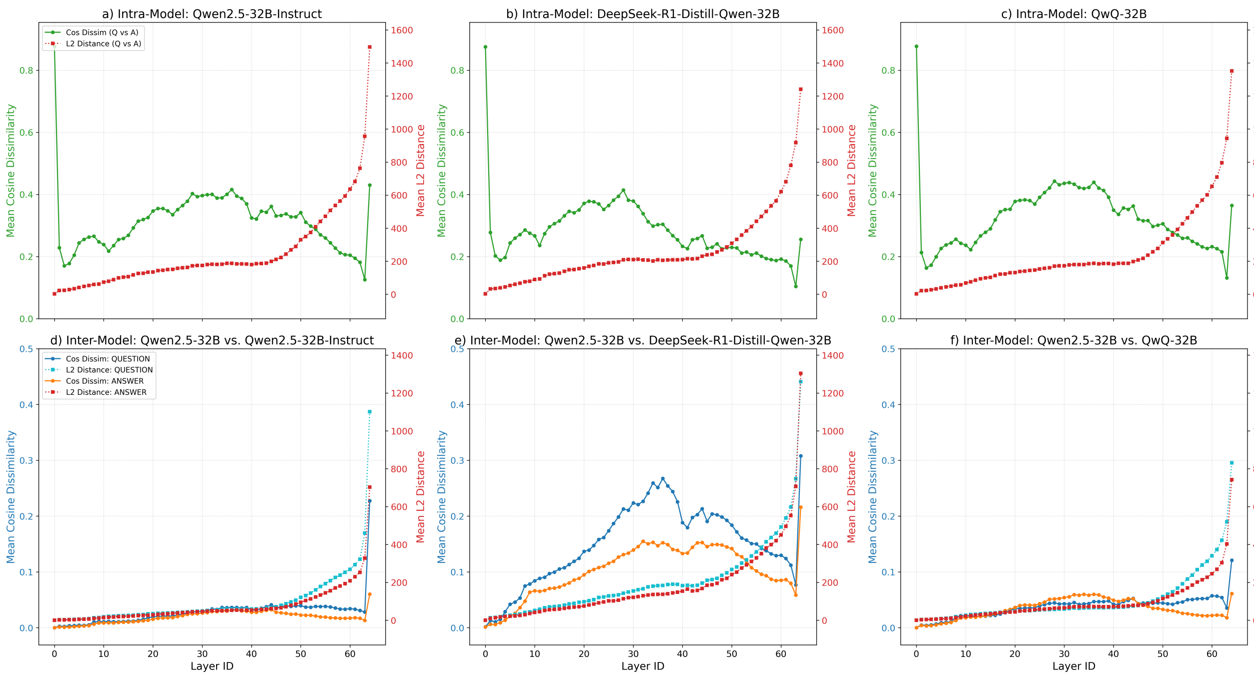
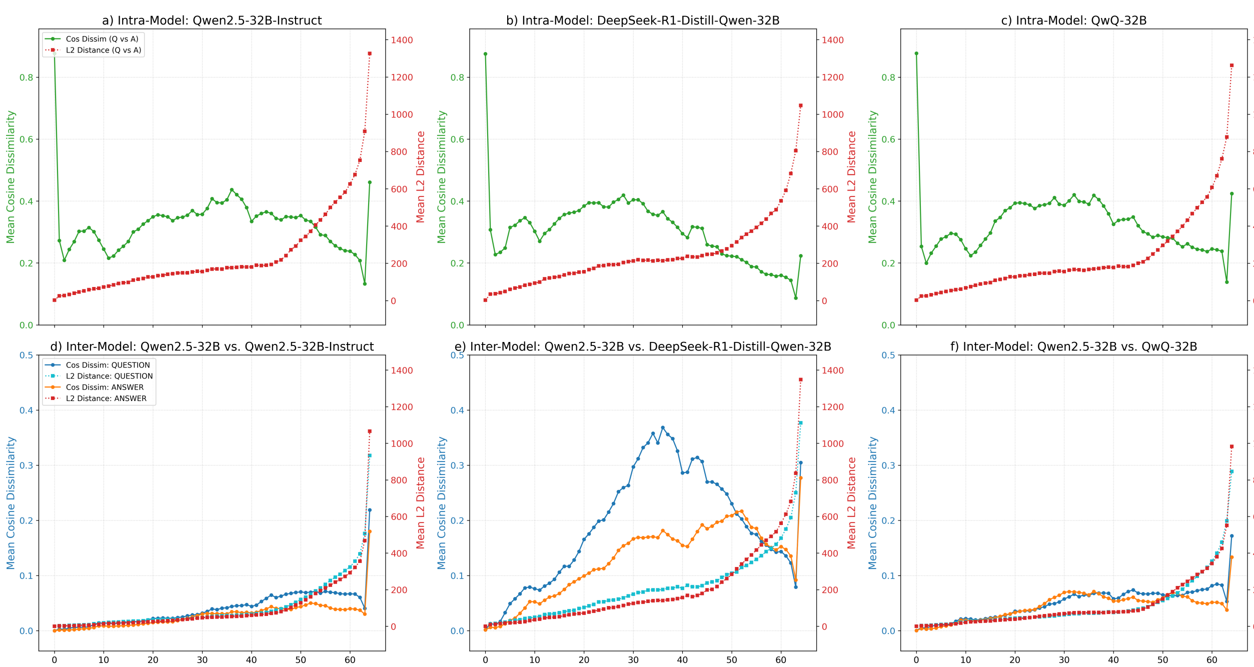


Figure 4: **Layer-wise Representation Divergence Across CoT Steps for All MedConceptsQA Vocabularies.** This figure shows the divergence analysis results for the ATC, ICD9PROC, ICD10PROC, ICD9CM, and ICD10CM vocabularies. The top and bottom rows correspond to mean cosine similarity and L2 distance, respectively. Each column represents a distinct step in the Chain-of-Thought (CoT) process, from Step 0 (the original question) to the final step (the original question plus the complete hierarchical traversal to the correct answer).

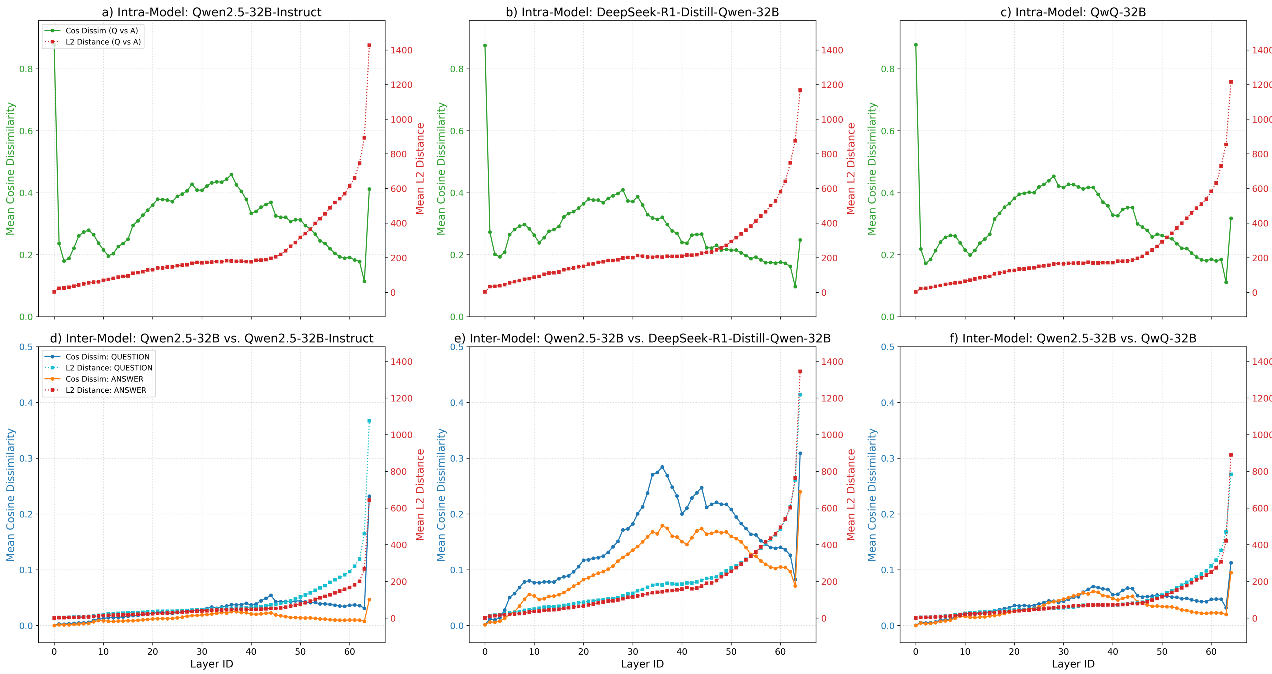
[t]0.9 **ATC**



[t]0.9 **ICD10PROC**



[t]0.9 **ICD9CM**



[p]0.9 **ICD10CM**

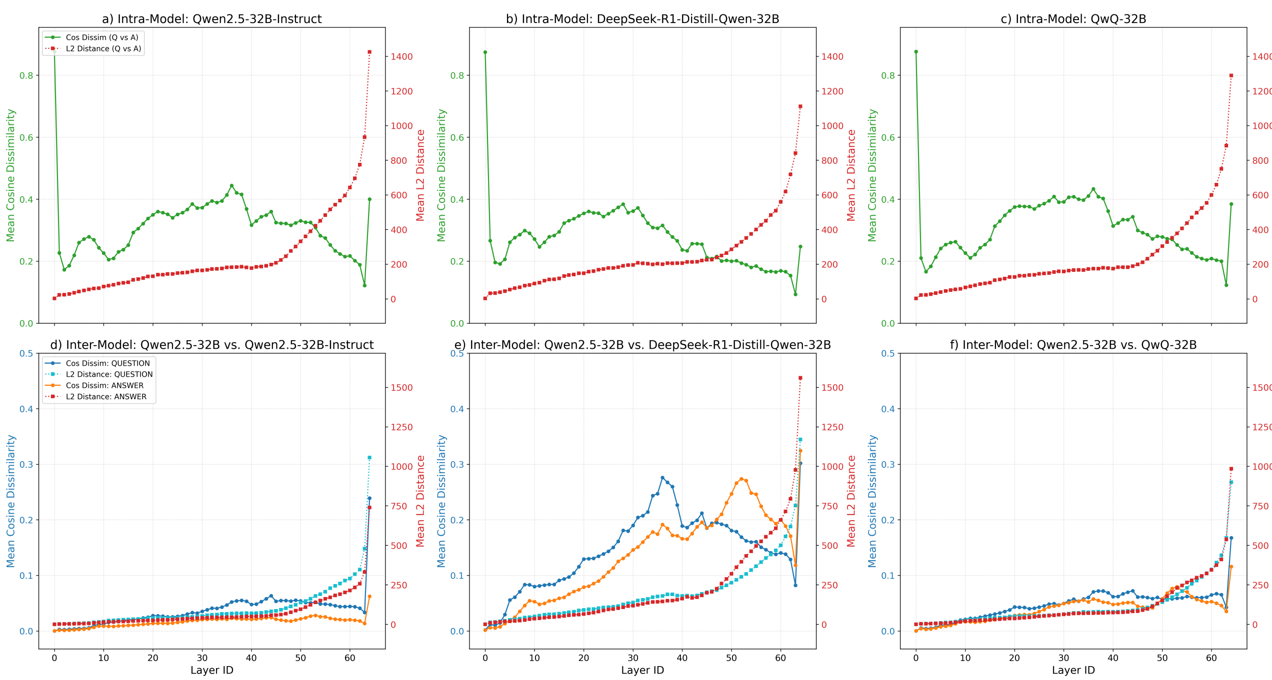
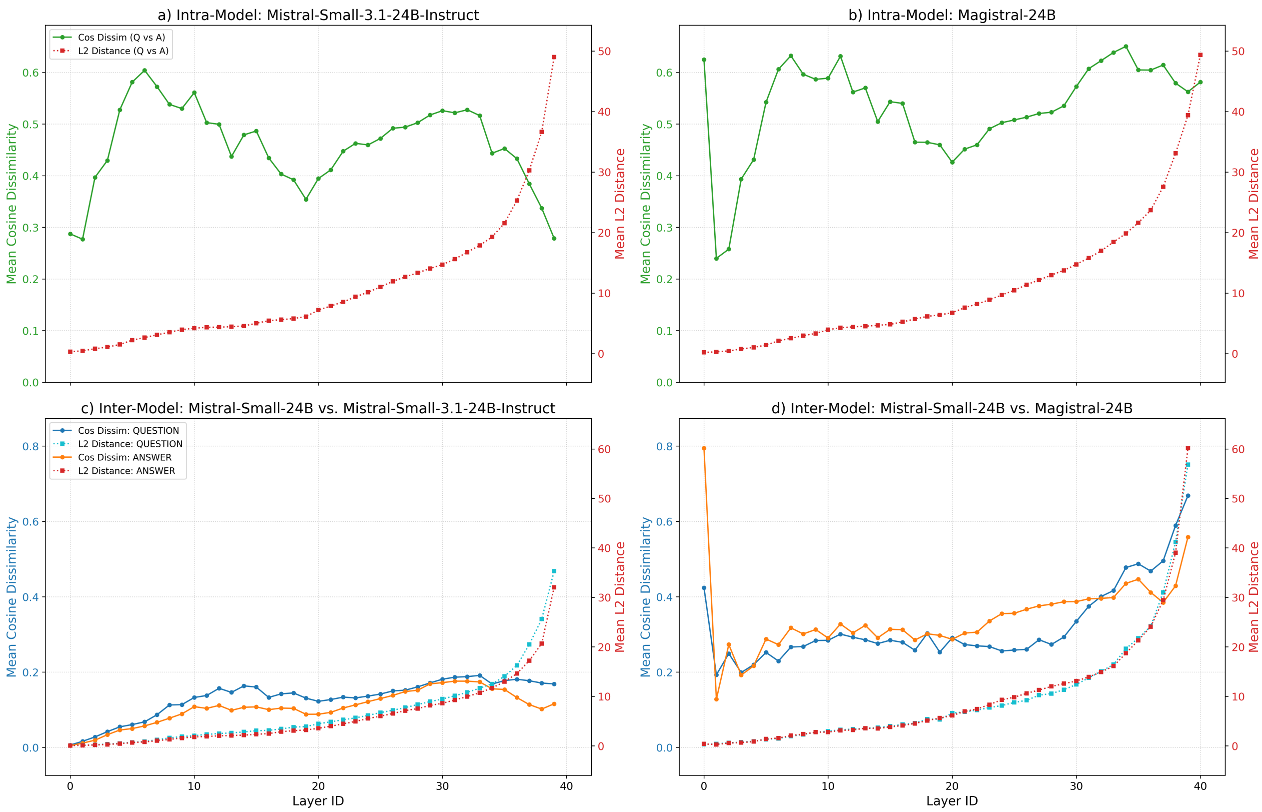
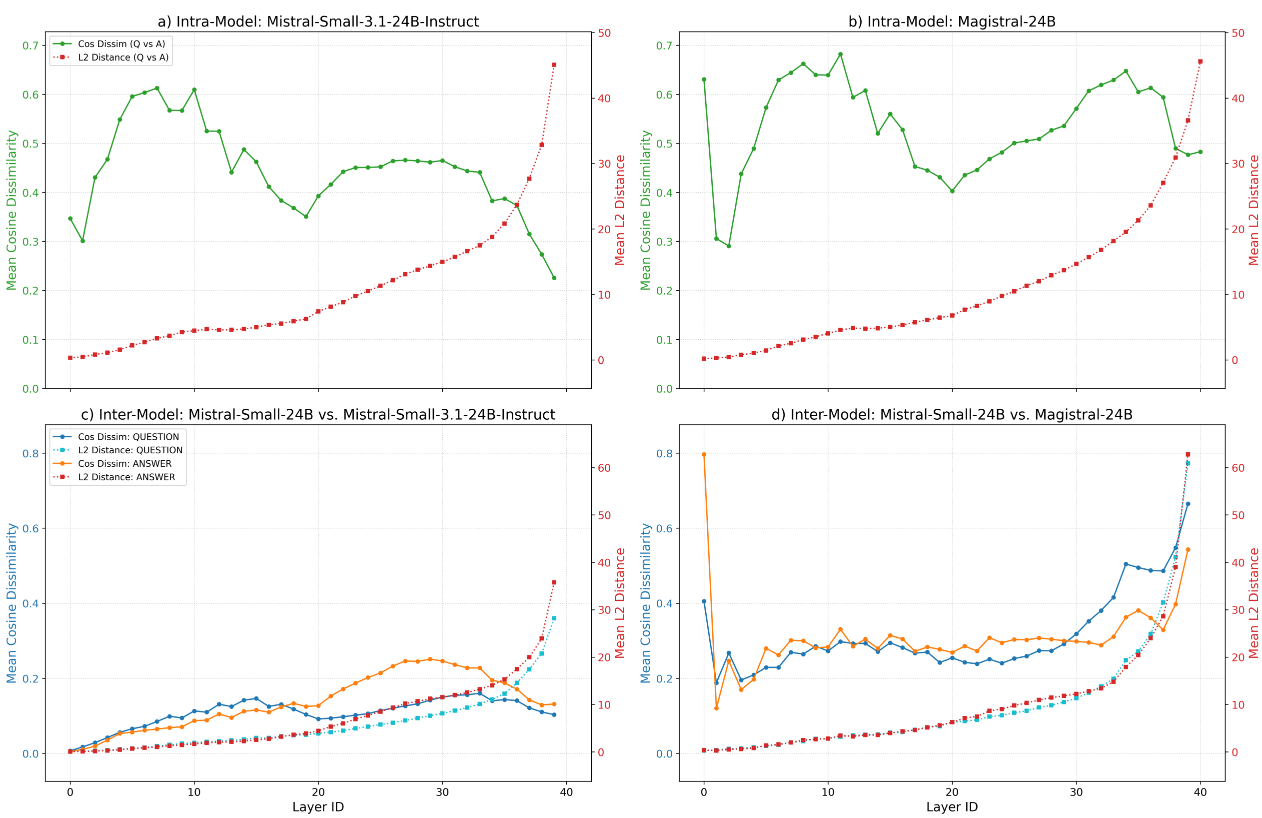


Figure 5: **Layer-wise Representation Divergence Across Remaining MedConceptsQA Vocabularies.** Same visualization format as Figure 3, showing results for ATC, ICD10PROC, ICD9CM, and ICD10CM vocabularies. Top and bottom rows correspond to intra- and inter-model divergence, respectively.

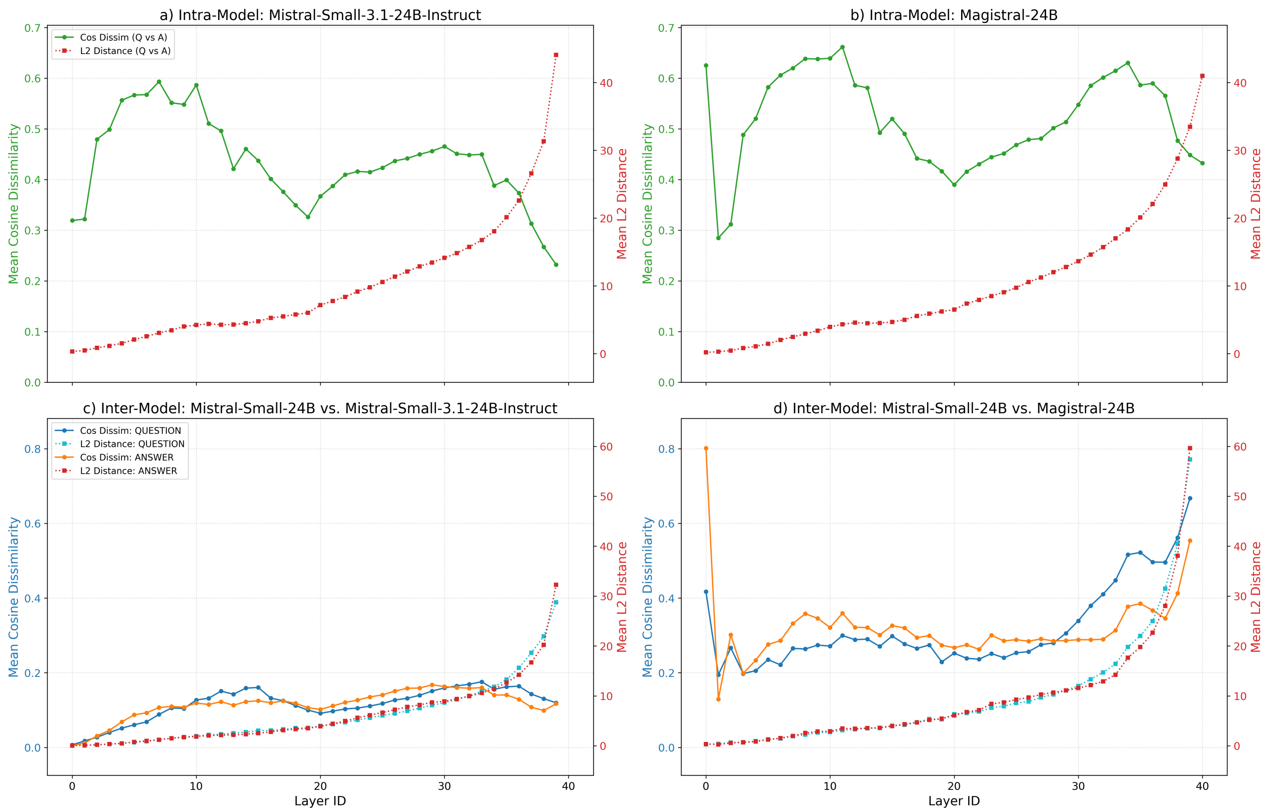
[t]0.9 **ATC**



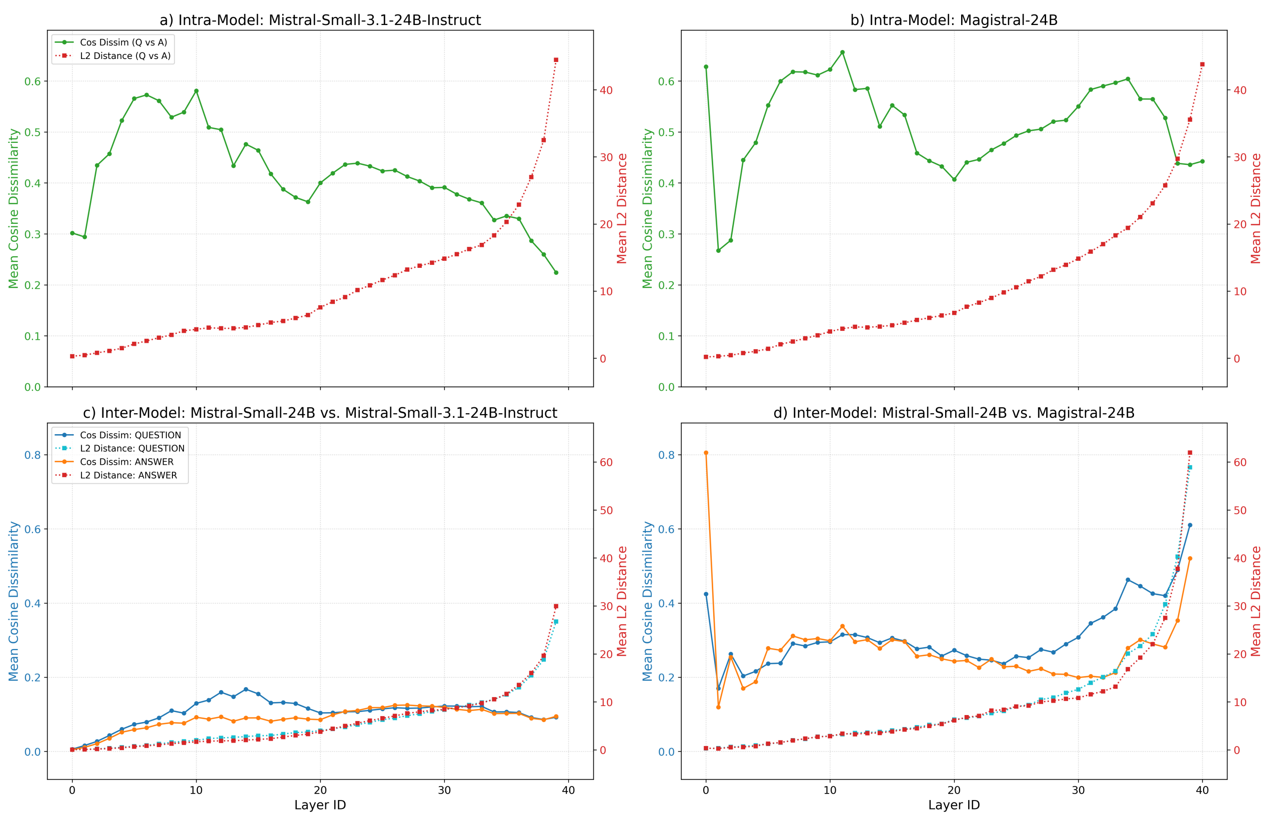
[t]0.9 **ICD9PROC**



[t]0.9 **ICD10PROC**



[t]0.9 **ICD9CM**



[t]0.9 **ICD10CM**

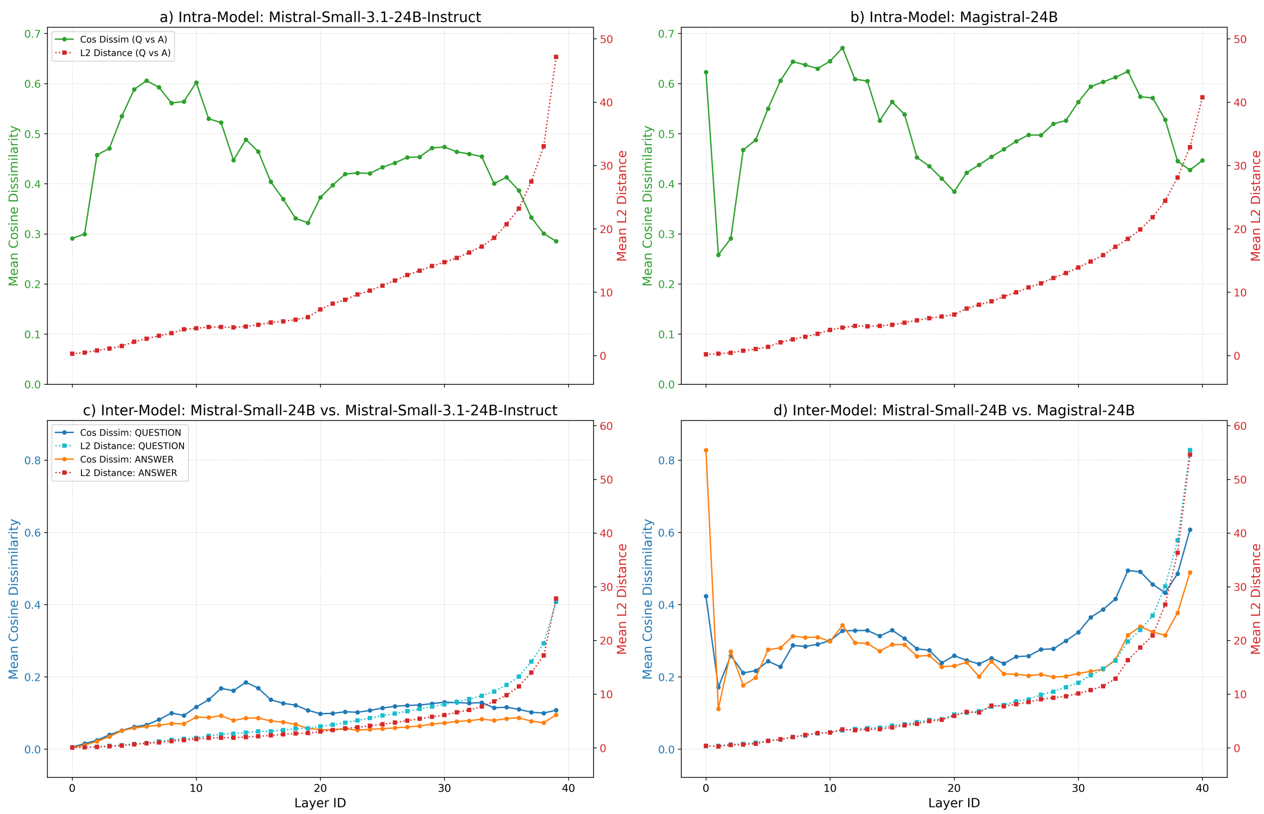


Figure 6: **Layer-wise Representation Divergence Across Remaining MedConceptsQA Vocabularies.** Same visualization format as Figure 3, showing results for ATC, ICD10PROC, ICD9CM, and ICD10CM vocabularies. Top and bottom rows correspond to intra- and inter-model divergence, respectively.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.