LoRi: Low-Rank Distillation for Implicit Reasoning

Ryan Solgi[1], **Jiayi Tian**[1], **Zheng Zhang**[1]
[1]University of California-Santa Barbara, USA
solgi@ucsb.edu, zhengzhang@ece.ucsb.edu

Abstract

Implicit chain-of-thought (iCoT) methods aim to internalize reasoning in large language models, but often underperform explicit CoT prompting. We empirically find that hidden-state reasoning trajectories exhibit low-rank structure. Motivated by this observation, we propose a low-rank distillation framework that transfers reasoning by aligning teacher and student trajectories in a shared low-rank tensor subspace using first- and second-order statistics. The resulting formulation captures the global structure of reasoning while supporting a compact latent reasoning process. We evaluate the method across multiple model families, including LLaMA and Qwen, at different scales on mathematical reasoning benchmarks. Our approach consistently improves performance, especially on challenging multi-step tasks, approaching explicit CoT accuracy and outperforming prior iCoT distillation methods.

1 Introduction

Large language models (LLMs) exhibit strong reasoning abilities under explicit chain-of-thought (CoT) prompting (Wei et al., 2022; Zelikman et al., 2022). However, explicit CoT is computationally expensive, can encourage reliance on non-robust patterns, and is sensitive to inference procedures (Li et al., 2024; Lin et al., 2025; Wang et al., 2022; Yao et al., 2023). As a result, recent work has explored more efficient reasoning paradigms that reduce dependence on explicit textual rationales (He et al., 2026; Li et al., 2025b).

Implicit chain-of-thought (iCoT) encodes reasoning in latent representations rather than explicit text (Deng et al., 2024a; Hao et al., 2024). Recent iCoT distillation approaches transfer explicit rationales into latent reasoning states (Shen et al., 2025; Wu et al., 2025; Kuzina et al., 2025; Wei et al., 2025), but still underperform explicit CoT on difficult mathematical reasoning tasks. A key challenge is that explicit reasoning steps lack a clear correspondence to latent dynamics, making token-to-latent transfer inherently ill-posed. Existing methods rely on local supervision (Shen et al., 2025) or sampled intermediate states (Kuzina et al., 2025), which may not fully capture the global reasoning trajectory.

Prior work suggests that model representations exhibit low-dimensional geometric structure (Yu and Wu, 2023; Golowich et al., 2025; Modell et al., 2025; Park et al., 2024). While recent studies analyze reasoning through trajectory geometry and separability (Sun et al., 2026; Zhou et al., 2026; Li et al., 2025a), they do not directly examine the low-rank structure of hidden states across tokens and layers. By stacking hidden states across layers and CoT tokens, we find that the normalized cumulative singular values grow rapidly with rank (Appendix A), indicating that reasoning trajectories are well-approximated by low-dimensional subspaces.

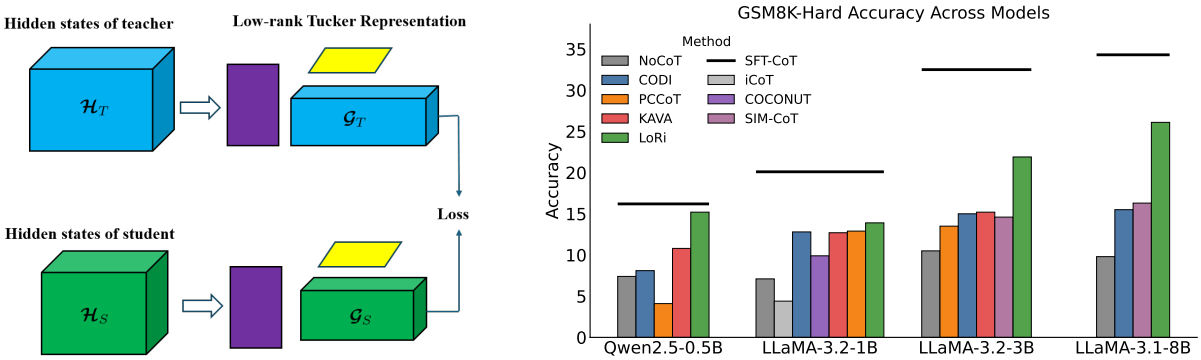
Motivated by this observation, we propose a low-rank iCoT distillation framework that aligns the global geometry of the teacher's reasoning trajectory through low-rank statistical representations. Instead of mimicking explicit reasoning steps, the student learns the low-dimensional subspace underlying the teacher's reasoning dynamics [Fig. 1 (a)]. This allows the student to capture the principal reasoning structure with a short latent trajectory, enabling length-invariant and efficient distillation.

Our main contributions are summarized below:

- **Low-rank iCoT distillation.** We propose an iCoT distillation framework that transfers reasoning by aligning teacher and student trajectories in a shared low-rank subspace using first- and second-order hidden-state statistics.

- **Efficient, length-invariant global reasoning transfer.** The formulation transfers long CoT reasoning into short latent trajectories independent

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49860–49871
October 24-29, 2026 ©2026 Association for Computational Linguistics



(a) High-level overview of LoRi.

(b) Representative reasoning results on GSM8K-Hard.

Figure 1: Overview of the proposed low-rank iCoT distillation framework and its reasoning performance. (a) Low-rank factors learned from the teacher hidden states are used to project teacher and student representations into a shared low-rank subspace. The projected low-rank representations are then matched through a loss function. (b) LoRi improves reasoning accuracy over prior implicit CoT baselines across model scales on GSM8K-Hard.

of sequence length, yielding efficient reasoning with a small number of latent steps and without requiring additional intermediate supervision.

- **Consistent gains across models and benchmarks.** Across models and scales, LoRi consistently outperforms prior iCoT methods, improving accuracy by up to ∼12% and achieving strong gains on GSM8K-Hard [up to ∼10%, as shown in Fig. 1 (b)]. At larger scales, LoRi substantially narrows the gap to explicit CoT.

2 Background and Related Work

iCOT Distillation. In reasoning, a model predicts the final answer $\boldsymbol{y}$ conditioned on an intermediate reasoning process $\boldsymbol{\tau}$ given an input $\boldsymbol{x}$,

$$p(\boldsymbol{y}, \boldsymbol{\tau} \mid \boldsymbol{x}) = p(\boldsymbol{\tau} \mid \boldsymbol{x})\, p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\tau}),$$

where $\boldsymbol{\tau}$ may correspond to explicit textual rationales (explicit CoT) or implicit internal reasoning processes (iCoT). Following prior work (Shen et al., 2025; Kuzina et al., 2025; Wei et al., 2025), we consider a teacher–student distillation framework to transfer reasoning from CoT into iCoT.

- **Teacher Model.** The teacher generates a natural language reasoning trajectory $\boldsymbol{r} = (r_1, \cdots, r_N)$ and defines the joint conditional distribution

$$p_T(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{x}) = p_T(\boldsymbol{r} \mid \boldsymbol{x})\, p_T(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{r}).$$

- **Student Model.** The student constructs a latent reasoning trajectory $\{\boldsymbol{z}_t\}_{t=1}^{K}$, where $K \ll N$ and each $\boldsymbol{z}_t$ represents a hidden reasoning state. The trajectory is generated recursively as

$$\boldsymbol{z}_t = f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}_{1:t-1}), \qquad t = 1, \ldots, K,$$

where $f_{\boldsymbol{\theta}}$ is the student model, with $\boldsymbol{z}_1$ initialized from $\boldsymbol{x}$. The final answer is then generated autoregressively:

$$p_S(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K).$$

- **Training.** The student is trained through an objective of the form

$$\mathcal{L} = \mathcal{L}_{\text{reason}} + \lambda \mathcal{L}_{\text{task}},$$

where $\mathcal{L}_{\text{reason}}$ transfers reasoning behavior from explicit CoT teacher into latent reasoning dynamics, $\mathcal{L}_{\text{task}}$ supervises answer prediction, and λ balances the two objectives.

Although iCoT improves inference efficiency, it often reduces reasoning accuracy. Existing methods mainly differ in how they transfer reasoning from explicit CoT into iCoT. Stepwise Internalization (Deng et al., 2024a) gradually removes CoT tokens through iterative fine-tuning, while COCONUT (Hao et al., 2024) replaces textual reasoning with hidden-state dynamics. Recent approaches rely on distillation (Deng et al., 2024b): CODI (Shen et al., 2025) distills reasoning at the answer boundary, PCCoT (Wu et al., 2025) adds tokens for intermediate reasoning states, KAVA (Kuzina et al., 2025) aligns KV-cache dynamics, and SIM-CoT (Wei et al., 2025) uses step-level supervision to guide reasoning trajectories.

Low-Dimensional Structure. The manifold hypothesis states that high-dimensional data often lie on a low-dimensional manifold capturing their intrinsic structure (Bengio et al., 2013; Fefferman

et al., 2016; Chen et al., 2022). Recent work suggests that LLM representations exhibit similar low-dimensional geometry, including approximately low-rank structure of activations (Yu and Wu, 2023; Chen et al., 2024; Liu et al., 2025), linear semantic directions (Park et al., 2024), representation manifold structure (Modell et al., 2025), and low-rank behavior in output logits (Golowich et al., 2025).

Reasoning Geometry. In reasoning tasks, Sun et al. (2026) show that reasoning trajectories pass through step-specific subspaces that become increasingly separable with depth. Zhou et al. (2026) characterize reasoning as smooth flows shaped by logical structure, while Li et al. (2025a) propose a reasoning manifold where correct trajectories concentrate in low-dimensional regions and errors arise from deviations. Together, these studies suggest that high-dimensional reasoning trajectories can be effectively approximated by low-dimensional subspaces or manifolds.

3 The LoRi Method

We view the teacher's reasoning process as a trajectory in hidden-state space that lies near a low-dimensional subspace. From this perspective, distillation need not enforce step-by-step correspondence; instead, it should align the student's latent reasoning dynamics with the dominant structure of the teacher's trajectory. Motivated by this view, we propose **LoRi** (Low-Rank iCoT), a distillation framework that transfers reasoning from explicit chain-of-thought into a compact implicit latent process. LoRi combines two complementary objectives: rationale-level alignment, which preserves the global geometry of the teacher's trajectory, and anchor-level alignment, which regularizes the transition from latent reasoning to answer generation. Together with an efficient training scheme based on precomputed low-rank representations, LoRi enables scalable, length-invariant distillation from a CoT teacher to an iCoT student.

3.1 Low-Rank iCoT Distillation

Distillation loss. The goal of distillation is to transfer the reasoning structure encoded in the teacher's explicit rationales r into the student's latent reasoning dynamics z, while preserving the student's ability to generate explicit step-by-step solutions and final answers. We train the student with the composite objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{LR}} + \lambda\, \mathcal{L}_{\mathrm{CE}},$$

where $\mathcal{L}_{\mathrm{CE}}$ is a cross entropy that supervises the student's explicit output sequence, $\mathcal{L}_{\mathrm{LR}}$ aligns the student's latent reasoning states with the teacher's hidden representations in a low-rank subspace, and λ balances the two terms. Intuitively, $\mathcal{L}_{\mathrm{LR}}$ encourages both models to share the dominant low-rank structure of the teacher's reasoning trajectory, improving the student's stability and generalization.

Low-rank term. We define $\mathcal{L}_{\mathrm{LR}}$ as the sum of two complementary components,

$$\mathcal{L}_{\mathrm{LR}} = \mathcal{L}_{\mathrm{rationale}} + \beta\, \mathcal{L}_{\mathrm{anchor}},$$

where $\mathcal{L}_{\mathrm{rationale}}$ aligns the student's latent reasoning dynamics with the teacher's hidden states over the full reasoning trajectory, and $\mathcal{L}_{\mathrm{anchor}}$ aligns the student's representation at the answer prediction position with the teacher's corresponding hidden state, while β controlling its relative weight. The former captures the global low-rank structure of the teacher's reasoning process, while the latter provides a localized signal that guides the transition from internal reasoning to answer generation.

3.2 Rationale Level Alignment ($\mathcal{L}_{\mathrm{rationale}}$)

To transfer reasoning capability, we align the student's hidden states with the teacher's reasoning trajectory in a low-rank subspace. Let $\boldsymbol{\mathcal{H}}_T \in \mathbb{R}^{N\times L\times H}$ denote the teacher's hidden states for rationale tokens, where L is the number of layers and H is the hidden dimension. Similarly, let $\boldsymbol{\mathcal{H}}_S \in \mathbb{R}^{K\times L\times H}$ be the student's hidden states for its implicit reasoning steps. Instead of directly matching these high-dimensional tensors, we project them into a shared low-rank subspace and align their low-dimensional representations through a statistical matching objective (Fig. 2).

Tucker Representation for Reasoning Trajectories. According to the Tucker decomposition (De Lathauwer et al., 2000), a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times L\times H}$ is factorized into a tensor core $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{r_N\times r_L\times r_H}$ and orthonormal factor matrices $\mathbf{U}_N \in \mathbb{R}^{N\times r_N}$, $\mathbf{U}_L \in \mathbb{R}^{L\times r_L}$, $\mathbf{U}_H \in \mathbb{R}^{H\times r_H}$:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{Q}} \times_1 \mathbf{U}_N \times_2 \mathbf{U}_L \times_3 \mathbf{U}_H,$$

where $\times_n$ denotes the mode-n tensor product (see Appendix F). We construct a low-rank Tucker-style representation of the teacher's hidden states along the layer and hidden dimensions by extracting low-rank factor matrices via SVD (top left singular vectors) of unfolded $\boldsymbol{\mathcal{H}}_T$ along the layer and hidden

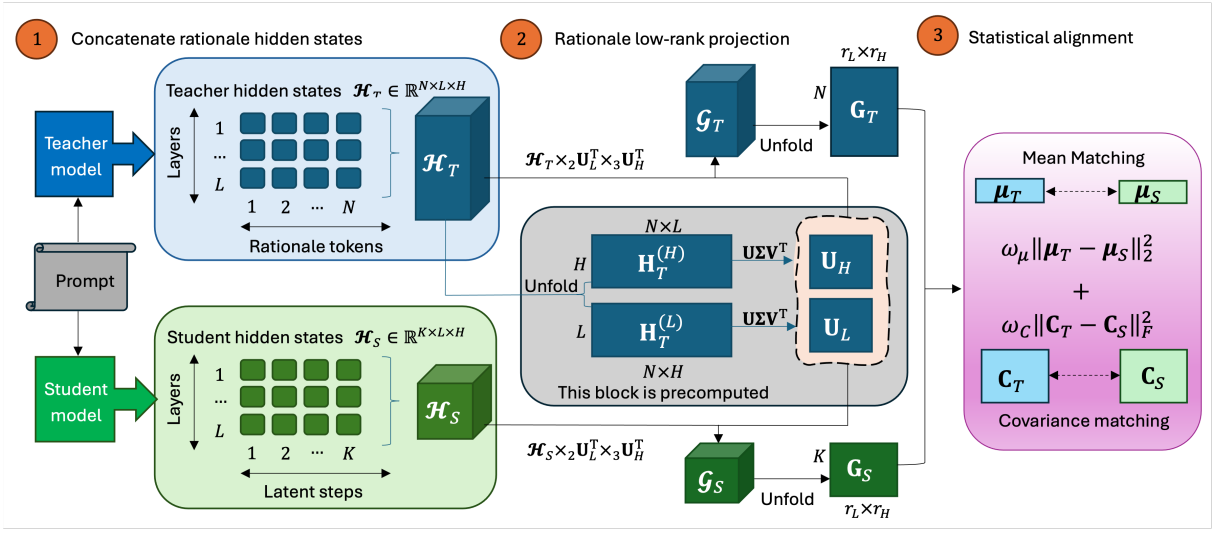


Figure 2: Overview of low-rank rationale-level alignment in LoRi.

dimensions, yielding orthonormal column matrices $\mathbf{U}_L \in \mathbb{R}^{L\times r_L}$ and $\mathbf{U}_H \in \mathbb{R}^{H\times r_H}$, respectively, which capture the dominant subspaces of the teacher's hidden-state representations. Each reasoning state is then projected as

$$\mathcal{G} = \mathcal{X} \times_2 \mathbf{U}_L^\top \times_3 \mathbf{U}_H^\top,$$

where $\mathcal{X}$ denotes either $\mathcal{H}_T$ or $\mathcal{H}_S$. This yields low-dimensional representations $\mathcal{G}_T \in \mathbb{R}^{N\times r_L\times r_H}$ for the teacher and $\mathcal{G}_S \in \mathbb{R}^{K\times r_L\times r_H}$ for the student. We then reshape these tensors along the last two modes to obtain matrix representations

$$\mathbf{G}_T \in \mathbb{R}^{N\times(r_L r_H)}, \qquad \mathbf{G}_S \in \mathbb{R}^{K\times(r_L r_H)},$$

where each row corresponds to a reasoning step represented in the shared low-rank subspace.

Statistical Alignment in Low-rank Space. We align student and teacher representations by matching the first- and second-order statistics of their projections in a shared low-rank subspace. For $\mathbf{G} \in \mathbb{R}^{M\times(r_L r_H)}$, where $M = N$ for the teacher and $M = K$ for the student, we define

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{G}_{m,:}, \qquad \mathbf{C} = \frac{1}{M}\mathbf{G}^\top\mathbf{G}.$$

Let $(\mathbf{C}_T, \boldsymbol{\mu}_T)$ and $(\mathbf{C}_S, \boldsymbol{\mu}_S)$ denote the statistics computed from $\mathbf{G}_T$ and $\mathbf{G}_S$, respectively. We define the rationale-level loss as

$$\mathcal{L}_{\text{rationale}} = \omega_C \, \|\mathbf{C}_S - \mathbf{C}_T\|_F^2 + \omega_\mu \, \|\boldsymbol{\mu}_S - \boldsymbol{\mu}_T\|_2^2,$$

where ω_C and ω_μ weight covariance and mean alignment. Matching covariance aligns the principal low-rank structure, while matching means aligns the trajectory centroids. Together, these constraints guide the student toward the teacher's reasoning geometry and discourage degenerate solutions such as collapsed latent trajectories.

This objective transfers the global structure of the teacher's reasoning process without token-level alignment and remains invariant to reasoning length. The low-rank factors $\mathbf{U}_L$ and $\mathbf{U}_H$ are precomputed from the teacher and fixed during training, enabling efficient projection without repeated factorization.

3.3 Anchor-Level Alignment ($\mathcal{L}_{\text{anchor}}$)

While the rationale-level objective captures global reasoning structure, it does not explicitly constrain the transition from latent reasoning to answer generation. To address this, we introduce a localized alignment term at the answer prediction position. This position corresponds to a fixed prompt-aligned token (e.g., "The final answer is"), which serves as a consistent transition between reasoning and answer generation rather than part of the reasoning trajectory itself (Figure 3).

Let $\mathcal{H}_T^{\text{anchor}} \in \mathbb{R}^{B\times L\times H}$ denote the collection of teacher hidden states at the answer prediction position across B training samples. We construct a low-rank Tucker decomposition of this tensor, yielding factor matrices

$$\mathbf{U}_B^{(a)} \in \mathbb{R}^{B\times r_B}, \; \mathbf{U}_L^{(a)} \in \mathbb{R}^{L\times r_L}, \; \mathbf{U}_H^{(a)} \in \mathbb{R}^{H\times r_H}.$$

In our formulation, we use the hidden-mode factor $\mathbf{U}_H^{(a)}$, which captures the dominant subspace of the teacher's hidden representations at the anchor

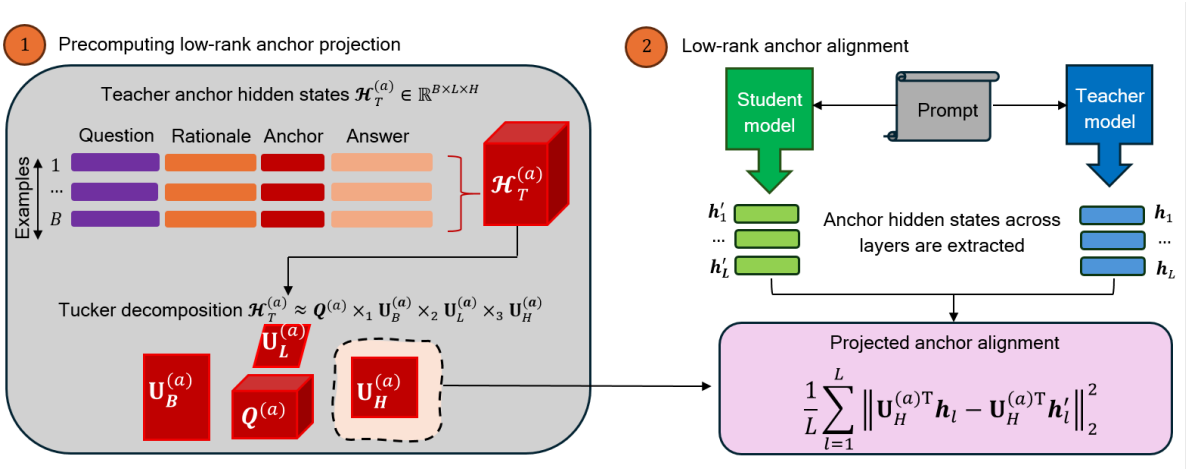


Figure 3: Overview of low-rank anchor alignment in LoRi.

position across the dataset. For a given sample, let $\boldsymbol{h}_\ell, \boldsymbol{h}'_\ell \in \mathbb{R}^H$, for $\ell = 1, \ldots, L$, denote the teacher and student hidden states at the answer token across layers. The anchor-level loss is defined as

$$\mathcal{L}_{\text{anchor}} = \frac{1}{L} \sum_{\ell=1}^{L} \left\| \mathbf{U}_H^{(a)\top} \boldsymbol{h}_\ell - \mathbf{U}_H^{(a)\top} \boldsymbol{h}'_\ell \right\|_2^2 .$$

This formulation leverages a shared low-rank subspace learned across training samples, ensuring that the student's representation at the answer prediction point aligns with the dominant structure of the teacher's hidden states at that location, while remaining complementary to the global alignment enforced by $\mathcal{L}_{\text{rationale}}$.

3.4 Implications for Reasoning

Reasoning compression. The low-rank structure provides a geometric explanation for why long chain-of-thought reasoning can be compressed into a shorter latent trajectory. Since the teacher's hidden-state trajectory lies near a low-rank subspace, its dominant variation can be represented using a limited number of degrees of freedom, largely independent of sequence length. By aligning the student within this subspace, the student is encouraged to reproduce the principal reasoning dynamics without explicitly matching each reasoning step.

Length invariance. Importantly, our formulation is invariant to the length of the reasoning trajectory. Since alignment is performed through aggregated statistics rather than step-wise correspondence, the student can construct a shorter latent trajectory that matches the global geometry of the teacher. This provides an intuitive justification for transferring reasoning from long explicit CoT sequences to compact implicit reasoning processes.

3.5 Relation to Prior Work

Our method differs from prior iCoT distillation approaches in both supervision and training. In the following we compare LoRi with representative prior work CODI (Shen et al., 2025), KAVA (Kuzina et al., 2025) and SIM-CoT (Wei et al., 2025):

- CODI aligns representations at a single boundary token, reducing supervision to a point-wise constraint and overlooking the structure of the reasoning trajectory. As a result, it struggles to transfer multi-step reasoning dynamics. In contrast, our method captures the full reasoning trajectory through a low-rank representation, supplemented by a localized anchor term that constrains the transition from reasoning to answer generation at a fixed answer-boundary position.

- KAVA distills teacher KV-cache trajectories through step-wise alignment in KV space using sampled rationale tokens. By contrast, LoRi models the reasoning process through low-rank hidden-state geometry, enabling length-invariant distillation without token-level sampling.

- SIM-CoT (Wei et al., 2025) applies step-level supervision by aligning latent states with explicit reasoning tokens via an auxiliary decoder. Our method instead avoids explicit step-wise alignment and supervises the student through low-rank trajectory structure.

Finally, existing iCoT distillation methods rely on joint teacher–student training with online

Table 1: Accuracy (%) of different reasoning methods on math benchmarks across models. The best iCoT method is shown in bold, and the second-best method is underlined. Explicit CoT results are shown in gray. Parentheses denote CODI/KAVA results without their final-step dropping heuristic.

Model	Method	GSM8K	GSM8K-Hard	SVAMP	Average
Qwen2.5-0.5B	SFT-CoT	52.3	16.2	62.3	43.6
	NoCoT	31.5	7.4	34.5	24.5
	CODI	37.5	8.1	47.0	30.9
	PCCOT	20.5	4.1	33.0	19.2
	KAVA	46.9	10.8	50.6	36.1
	LoRi	**50.0**	**15.2**	**63.0**	**42.7**
	LoRi-E	52.0	16.9	62.7	43.9
LLaMA-3.2-1B	SFT-CoT	58.4	20.1	61.0	46.5
	NoCoT	30.9	7.1	44.1	27.4
	iCoT	19.0	4.4	40.9	21.4
	CODI	55.6 (47.2)	12.8	61.1	43.2
	COCONUT	45.3	9.9	48.8	34.7
	KAVA	**56.5** (51.2)	12.7	58.9	42.7
	PCCoT	54.2	12.9	57.7	41.6
	LoRi	54.9	**13.9**	**64.0**	**44.3**
	LoRi-E	58.1	14.9	64.3	45.8
LLaMA-3.2-3B	SFT-CoT	74.9	32.5	78.7	62.0
	NoCoT	41.7	10.5	56.9	36.4
	PCCoT	54.7	13.5	69.5	45.9
	CODI	61.0	15.0	72.4	49.5
	KAVA	65.7	15.2	72.7	51.2
	SIM-CoT	62.3	14.6	74.9	50.6
	LoRi	**70.0**	**21.9**	**75.0**	**55.6**
	LoRi-E	74.3	24.1	79.3	59.2
LLaMA-3.1-8B	SFT-CoT	76.6	34.3	81.0	64.0
	NoCoT	39.5	9.8	55.3	34.9
	CODI	61.1	15.5	78.1	51.6
	SIM-CoT	64.1	16.3	79.4	53.3
	LoRi	**79.3**	**26.1**	**83.3**	**62.9**
	LoRi-E	80.4	27.2	83.3	63.6

teacher inference. In contrast, we adopt a two-stage procedure: teacher-derived low-rank factors are precomputed once, after which the student is fine-tuned independently on a subset of the training data. This substantially reduces training cost while preserving effective reasoning transfer.

4 Results

4.1 Experimental Setup

Following prior work, we evaluate LoRi on standard mathematical reasoning benchmarks, which provide rigorous tests of multi-step reasoning through structured problem solving and verifiable answers. This focus is also motivated by prior findings that CoT is most effective for mathematical and symbolic reasoning (Sprague et al., 2024). Our additional analysis across math, coding, science, and commonsense tasks further supports this observation with CoT-SFT providing little or no advantage over standard SFT outside mathematical reasoning (Appendix C).

We compare LoRi against NoCoT, CODI (Shen et al., 2025), KAVA (Kuzina et al., 2025), SIM-CoT (Wei et al., 2025), PCCoT (Wu et al., 2025), COCONUT (Hao et al., 2024), and SFT-CoT across Qwen (Bai et al., 2023) and LLaMA (Grattafiori et al., 2024) model families. To study scalability, we evaluate models ranging from 0.5B to 8B parameters. SFT-CoT serves as an upper-bound reference since it relies on explicit reasoning at inference time, whereas LoRi targets implicit reasoning.

The teacher model is first fine-tuned on the full GSM8K-Aug dataset (Deng et al., 2024a,b) and then frozen during distillation. The student is initialized from the same base model. A random 128-sample subset of GSM8K-Aug is used to estimate the teacher-derived low-rank quantities, which are subsequently kept fixed during student training. We report final-answer accuracy based on extracted numerical outputs. Additional implementation details and hyperparameters are provided in Appendix E.

4.2 Accuracy Comparison

Table 1 summarizes performance on several benchmarks and model scales. Overall, LoRi consistently outperforms prior implicit reasoning methods and substantially narrows the gap to full CoT models. This suggests that low-rank alignment captures global reasoning dependencies that token-level methods fail to model.

Performance w.r.t. Model Scale. For small models, LoRi provides strong gains over prior implicit reasoning methods. On Qwen2.5-0.5B, LoRi achieves an average accuracy of 42.7%, substantially outperforming CODI (30.9%), KAVA (36.1%), and PCCoT (19.2%). On LLaMA-3.2-1B, LoRi reaches the best average accuracy among iCoT methods (44.3%), although its advantage over CODI and KAVA is more nuanced: CODI and KAVA obtain slightly higher accuracy on GSM8K, while LoRi performs better on GSM8K-Hard and SVAMP. Moreover, when CODI and KAVA are evaluated without their final-step dropping heuristic (Kuzina et al., 2025), LoRi outperforms them by a clear margin. This is notable because LoRi does not use step-level supervision or heuristic rationale truncation.

For larger models, the gains are pronounced. On LLaMA-3.2-3B, LoRi achieves 55.6% average accuracy, outperforming all implicit baselines by a clear margin. On LLaMA-3.1-8B, LoRi reaches 62.9%, approaching the performance of full CoT models (64.0%). These results suggest that the proposed method scales effectively, substantially narrowing the gap between implicit and explicit reasoning.

Hard Reasoning Tasks. The improvements are particularly notable on GSM8K-Hard, which requires more complex multi-step reasoning. For example, on LLaMA-3.2-3B, LoRi improves performance from 15.2% (KAVA) to 21.9%; on LLaMA-3.1-8B, LoRi achieves 26.1%, substantially outperforming CODI (15.5%) and SIM-CoT (16.3%). This suggests that capturing the global structure of reasoning through low-rank alignment is especially beneficial for harder reasoning tasks.

Capability of Explicit Reasoning. To evaluate whether LoRi preserves explicit reasoning ability, we test LoRi-E, which uses the same student model but performs standard CoT generation at inference time. LoRi-E achieves performance comparable to SFT-CoT across all models and benchmarks. For instance, on LLaMA-3.1-8B, LoRi-E attains an average accuracy of 63.6%, close to 64.0% for SFT-CoT. These results suggest that the proposed distillation method preserves explicit reasoning capability while enabling both implicit and explicit reasoning at inference time.

Table 2: Training cost on NVIDIA RTX 6000 GPUs. FT: teacher fine-tuning. Distill: reasoning distillation.

Model	Method	Stage	GPU-h
LLaMA-1B	CODI	Distill	16.5
	LoRi	FT + Distill	**4.75 + 0.02**
LLaMA-3B	CODI	Distill	30.9
	LoRi	FT + Distill	**8.8 + 0.03**

4.3 Training Efficiency

Beyond accuracy, LoRi also improves training efficiency over prior implicit reasoning methods. Existing iCoT distillation approaches, including CODI, KAVA, and SIM-CoT, rely on joint teacher–student training with repeated teacher forward passes to obtain intermediate representations, resulting in substantial computational overhead. In contrast, LoRi uses a two-stage procedure: teacher-derived low-rank factors are estimated once from a small subset of the training data and subsequently kept fixed during student optimization, avoiding repeated teacher forward passes. Despite this simplified pipeline, LoRi maintains strong benchmark performance while significantly improving training efficiency and scalability as summarized in Table 2. In practice, the distillation overhead becomes effectively negligible, making LoRi a more scalable and accessible approach for implicit reasoning distillation. We use CODI as a representative baseline for training-cost reference since related iCoT distillation methods (e.g., PCCoT, KAVA, and SIM-CoT) employ similar joint teacher–student training procedures with repeated teacher forward passes, leading to very similar computational overhead.

4.4 Inference Latency

We compare the inference latency of LoRi-based iCoT and explicit CoT. Latency is measured as average wall-clock time per sample. CoT generates full reasoning sequences autoregressively, whereas iCoT uses a fixed number of latent steps ($K = 5$) followed by short answer generation, significantly reducing decoding cost. As shown in Figure 4, iCoT consistently achieves lower latency across all model scales. Specifically, it is about $6.9\times$

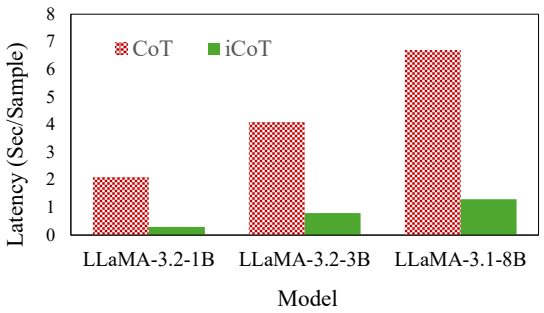


Figure 4: Inference latency comparison. Note: LoRi and prior iCoT methods have almost the same inference cost since they use the same latent-step inference procedure.

faster for LLaMA-3.2-1B and 5.1× faster for both LLaMA-3.2-3B and LLaMA-3.1-8B. These results confirm the efficiency advantages of iCoT while showing that LoRi preserves this benefit alongside improved reasoning performance. Since LoRi and prior iCoT methods use the same latent-step inference procedure, their inference complexity is almost the same; the main difference lies in the distillation strategy during training.

4.5 Ablation Studies

We study the effect of the number of latent reasoning steps K and the number of training samples used for distillation on model performance.

Ablation on Latent Reasoning Steps. Figure 5 reports benchmark accuracy as the number of latent steps K varies from 2 to 8. Performance consistently improves from $K = 2$ to $K = 5$, suggesting that a minimum number of latent reasoning iterations is needed to capture the underlying reasoning dynamics. The best results are achieved at $K = 5$ across all datasets, indicating that this setting provides sufficient capacity to represent the dominant low-rank reasoning structure. Increasing K beyond 5 yields no further gains and occasionally causes slight degradation, implying that additional steps introduce redundancy rather than useful computation. Overall, these results support the view that reasoning trajectories can be compressed into a small number of latent steps without sacrificing essential structure.

Ablation on Training Sample Size. Figure 5 shows that accuracy improves rapidly in the low-data regime and saturates after roughly 128 samples across all model scales, indicating that strong performance can be achieved with only a small subset of the training data. One explanation is that the effective degrees of freedom of the distillation problem are limited: although the teacher produces long chain-of-thought trajectories, the information needed for the student may lie in a low-dimensional subspace. Under this view, distillation only needs to recover the dominant reasoning structure rather than the full trajectory. Consequently, relatively few samples are sufficient to estimate this structure, leading to rapid gains followed by saturation. This observation supports our low-rank formulation and suggests that the reasoning signal relevant for distillation is highly compressible.

5 Conclusion

We have proposed LoRi, a low-rank formulation for iCoT distillation. Empirically, LoRi outperformed prior implicit reasoning methods across model families, scales, and benchmarks and reduced the gap of iCoT with explicit CoT while preserving the efficiency advantages of implicit reasoning. In addition, the proposed method significantly reduced computational overhead, making distillation lightweight and scalable. The results also support a geometric perspective of reasoning: long chain-of-thought trajectories can be effectively compressed by preserving their low-rank structure. This perspective provides a principled and efficient alternative to existing iCoT distillation approaches and suggests that the essential dynamics of reasoning are governed by a low-dimensional subspace that can be exploited for both learning and inference.

Limitations

The proposed formulation is motivated by an empirical observation that hidden-state reasoning trajectories exhibit strong low-rank structures across models and benchmarks. While LoRi improves implicit reasoning, the theoretical relationship between low-rank representation geometry and reasoning capability is still not fully understood. Our empirical evaluation is primarily based on a single mathematical reasoning training setting; although the coding analysis supports the broader low-rank observation, the usefulness of CoT varies across tasks, and LoRi's effectiveness may therefore depend on the dataset and teacher training setting. Future work could further investigate the geometry of reasoning trajectories and the role of low-dimensional structure in CoT reasoning from a theoretical perspective. Additionally, undisclosed pre-

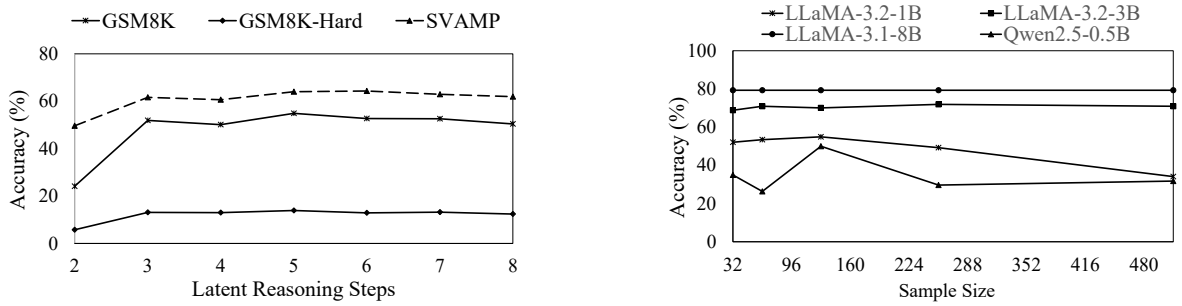


Figure 5: Ablation studies on reasoning steps (left) and training sample size (right).

training data may influence the observed CoT and non-CoT behaviors.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Jinze Bai et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhuotong Chen, Qianxiao Li, and Zheng Zhang. 2022. Self-healing robust neural networks via closed-loop control. *Journal of machine learning research*, 23(319):1–54.

Zhuotong Chen, Zihu Wang, Yifan Yang, Qianxiao Li, and Zheng Zhang. 2024. PID control-based self-healing to improve the robustness of large language models. *Transactions on machine learning research*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024a. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.

Yuntian Deng, Kiran Prsad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2024b. Implicit chain-of-thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv: 2211.10435*.

Noah Golowich, Allen Liu, and Abhishek Shetty. 2025. Sequences of logits reveal the low rank structure of language models. *arXiv preprint arXiv:2510.24966*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Ankur Pandey, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Zhenghao He, Guangzhi Xiong, Bohan Liu, Sanchit Sinha, and Aidong Zhang. 2026. Reasoning beyond chain-of-thought: A latent computational mode in large language models. *arXiv preprint arXiv:2601.08058*.

Anna Kuzina, Maciej Pioro, Paul N. Whatmough, and Babak Ehteshami Bejnordi. 2025. Kava: Latent reasoning via compressed kv-cache distillation. *arXiv preprint arXiv:2510.02312*.

Bo Li, Guanzhi Deng, Ronghao Chen, Junrong Yue, Shuo Zhang, Qinghua Zhao, Linqi Song, and Lijie Wen. 2025a. Rema: A unified reasoning manifold framework for interpreting large language models. *arXiv preprint arXiv:2509.22518*.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025b. Implicit reasoning in large language models: A comprehensive survey. *arXiv preprint arXiv:2509.02350*.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2025. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Mingsong Yan, Zi Yang, Paul D Hovland, Bogdan Nicolae, Franck Cappello, Sui Tang, and Zheng Zhang. 2025. CoLA: Compute-efficient pre-training of llms via low-rank activation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4627–4645.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. 2025. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Svamp: A dataset of verbally perturbed math word problems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. 2026. Llm reasoning as trajectories: Step-specific representation geometry and correctness signals. *arXiv preprint arXiv:2604.05655*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. 2025. Sim-cot: Supervised implicit chain-of-thought. *arXiv preprint arXiv:2509.20317*.

Haoyi Wu, Zhihao Teng, and Kewei Tu. 2025. Parallel continuous chain-of-thought with jacobi iteration. *arXiv preprint arXiv:2506.18582*.

Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. 2025. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6980–7008, Vienna, Austria.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*.

Hao Yu and Jianxin Wu. 2023. Compressing transformers: Features are low-rank, but weights are not! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11007–11015.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R. Zhang. 2026. The geometry of reasoning: Flowing logics in representation space. In *International Conference on Learning Representations (ICLR)*.

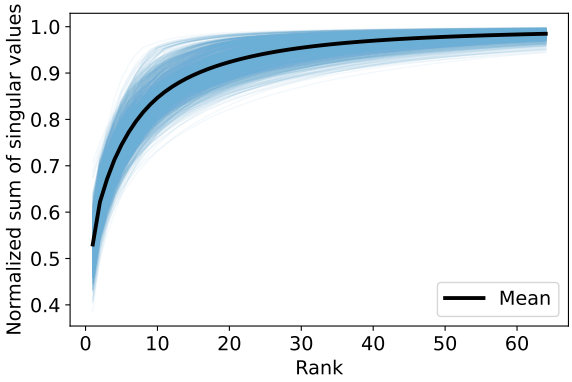


Figure 6: Normalized sum of singular values versus rank for stacked hidden states of all layers and all CoT tokens of LLaMA-3.2-1B model, where each curve corresponds to a single GSM8K example.

A Low-rank Analysis of CoT Hidden States

For this analysis, we stack the hidden states corresponding to all transformer layers and all CoT rationale tokens into a single matrix and compute its singular value decomposition (SVD). Figure 6 shows the normalized cumulative singular values across GSM8K examples for LLaMA-3.2-1B. Despite the model having a hidden dimension of 2048, the singular values decay rapidly and are largely concentrated within a relatively small rank. This strong low-rank structure suggests significant compressibility in hidden-state reasoning dynamics, motivating the use of low-rank representations for implicit reasoning distillation.

We further analyze CoT hidden states for coding task using LLaMA-3.2-1B. As seen in Table 3, CoT fine-tuning on KodCode (Xu et al., 2025) substantially concentrates the spectrum such that the dimensions required to capture 95% and 99% of the energy decrease from 293.3 to 85.3 and from 615.2 to 205.5, respectively, while the tail energy beyond rank 64 drops from 0.276 to 0.077. Meanwhile, accuracy improves on KodCode, MBPP (Austin et al., 2021), and HumanEval (Chen et al., 2021), suggesting that the increased spectral concentration reflects more structured representations rather than collapse. These results indicate that the low-rank structure may also emerge in non-mathematical reasoning.

B Ablation of the Anchor Loss Term

Table 4 shows that the anchor-level loss $\mathcal{L}_{\text{anchor}}$ generally improves performance, with the largest gains observed for smaller models. For Qwen2.5-0.5B, it yields substantial improvements on both GSM8K and SVAMP, indicating a strong additional supervision signal. For larger models, the effect is less pronounced, with improvements on some benchmarks and minor degradations on others. Notably, the magnitude of these degradations is considerably smaller than the observed gains, suggesting that the anchor term acts as a stable complementary signal to the low-rank rationale alignment. Overall, the largest benefits are observed in smaller models, while the effects become more nuanced at larger scales. Importantly, the proposed distillation method remains effective even without the anchor term, indicating that the global low-rank alignment alone captures a substantial portion of the reasoning signal.

C Extension Beyond Mathematical Reasoning

The effectiveness of LoRi depends on settings where CoT itself provides a meaningful advantage. To further examine the scope of CoT across task domains, we compare standard SFT and CoT-SFT on math, coding, science, and common sense tasks using LLaMA-3.2-1B. As shown in Table 5, CoT-SFT substantially improves performance on GSM8K, while providing little or no advantage over standard SFT on KodCode, ScienceQA (Lu et al., 2022), and CommonsenseQA (CSQA) (Talmor et al., 2019). This trend is consistent with prior findings that CoT is most effective for mathematical and symbolic reasoning (Sprague et al., 2024). These results further motivate our primary focus on mathematical reasoning, where CoT provides a meaningful advantage to distill.

D Extension to Reasoning-Oriented Models

To study LoRi's performance for a reasoning-oriented model, we additionally evaluate it on Qwen3-0.6B. As seen in Table 6, LoRi follows the same overall pattern observed with the non-thinking models: its latent reasoning recovers most of the explicit-CoT accuracy relative to the no-reasoning baseline. In particular, LoRi achieves 63.5% on GSM8K, 74.0% on SVAMP, and 24.8% on GSM8K-Hard, compared with 67.3%, 73.0%, and 30.3% for CoT-SFT.

Table 3: Low-rank structure of coding reasoning trajectories using LLaMA-3.2-1B. We report the number of dimensions required to capture 95% and 99% of the spectral energy, the residual energy beyond rank 64, and accuracy (%) on coding benchmarks.

Variant	Dims (95%)	Dims (99%)	Tail Energy ($r > 64$)	KodCode	MBPP	HumanEval
Base CoT	293.3	615.2	0.276	15.5	35.5	21.3
CoT-SFT (KodCode)	85.3	205.5	0.077	32.5	37.0	35.4

Table 4: Effect of the anchor loss term on accuracy across models.

Model	**Setting**	**GSM8K**	**SVAMP**
Qwen2.5-0.5B	No Anchor	39.8	54.7
	With Anchor	**50.0**	**63.0**
LLaMA-3.2-1B	No Anchor	52.8	**64.3**
	With Anchor	**54.9**	64.0
LLaMA-3.2-3B	No Anchor	**71.3**	77.3
	With Anchor	70	**83.3**

Table 5: Accuracy (%) of LLaMA-3.2-1B with and without chain-of-thought across different domains, before and after supervised fine-tuning.

Method	GSM8K	KodCode	ScienceQA	CSQA
Base NoCoT	4.9	**23.8**	62.6	**56.9**
Base CoT	**34.8**	15.0	**66.9**	49.1
NoCoT-SFT	30.9	**41.2**	**91.1**	**90.7**
CoT-SFT	**58.4**	32.5	86.3	74.3

Table 6: Accuracy (%) of LoRi on Qwen3-0.6B.

Method	GSM8K	SVAMP	GSM8K-Hard
NoCoT-SFT	25.1	23.5	5.6
CoT-SFT	**67.3**	73.0	**30.3**
LoRi	63.5	**74.0**	24.8

Table 7: Hyperparameter settings.

Hyperparameter	**Value**
R_L	8
R_H	256
$R_L^{(a)}$	8
$R_H^{(a)}$	128
$R_B^{(a)}$	64
λ	1
β	10^{-5}
ω_C	1
ω_μ	10^{-2}
Learning rate	$2 \cdot 10^{-5}$
Epochs	1
Samples	128

E Implementation Details and Hyperparameter

We assume a pretrained teacher model that produces high-quality chain-of-thought rationales. The teacher is kept fixed and used to extract hidden states for both the rationale tokens and the answer prediction position. From these, we precompute the low-rank projection factors and anchor-level targets as described in Sec. 3.2 and Sec. 3.3. The student model is then trained using the proposed loss in Eq. (3.1).

We evaluate the model on standard mathematical reasoning benchmarks, including GSM8K (Cobbe et al., 2021), GSM8K-Hard (Gao et al., 2022), and SVAMP (Patel et al., 2021). Hyperparameter settings for LoRi corresponding to the results in Table 1 is reported in Table 7. Here, R_L and R_H denote the ranks used for the low-rank decomposition of rationale representations, while $R_L^{(a)}$, $R_H^{(a)}$, and $R_B^{(a)}$ are the Tucker ranks for anchor factorization. We use a shared hyperparameter configuration across all models, except for $R_L^{(a)}$, which is increased for larger models (e.g., LLaMA 3B and 8B) to 16 due to their greater depth, while all other parameters remain fixed.

F Tensor Contraction

The operation $\times_n$ denotes contraction of a tensor with a matrix along mode n. In index form, the expression

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Q}} \times_1 \mathbf{U}_N \times_2 \mathbf{U}_L \times_3 \mathbf{U}_H$$

corresponds to

$$X_{n,t,h} = \sum_{m=1}^{r_N} \sum_{i=1}^{r_L} \sum_{j=1}^{r_H} Q_{m,i,j} \, (\mathbf{U}_N)_{n,m} \cdot (\mathbf{U}_L)_{t,i} \, (\mathbf{U}_H)_{h,j} \tag{1}$$

which contracts $\boldsymbol{\mathcal{X}}$ along the first, second and third modes.