Limits of Continuous Chain-of-Thought in Multi-Step and Multi-Chain Reasoning

Anonymous submission

Abstract

Chain-of-thought (CoT) improves LLM reasoning but can be slow and inefficient because it relies on discrete token generation. Recent *implicit* (answer-only) and *continuous* (latent-step) alternatives aim to reduce these costs, but evaluations so far have focused on simple benchmarks. We study how these models perform under two key challenges of compositional reasoning: robustness to increased chain length and the ability to exploit multiple valid reasoning paths. To probe these axes, we introduce a MATMUL benchmark with controllable depth and associativity-induced diversity, and a 100k-sample GSM8K corpus with 2–5 distinct chains per question. Across both settings, discrete CoT remains robust, while continuous and implicit models degrade as chains lengthen and fail to leverage reasoning diversity. We trace these failures to curriculum and alignment objectives that provide weak supervision for intermediate reasoning steps and can collapse diverse traces into a single representation. These results point to the need for training objectives that deliver intermediate credit assignment and preserve reasoning diversity in latent models.

Introduction

Chain-of-Thought (CoT) prompting and fine-tuning improve the ability of Large Language Models (LLMs) to answer challenging questions by deriving the answer step by step, for example by decomposing into sub-questions and revising when contradictions are detected. This *reasoning* is typically produced as discrete tokens, which has several limitations: (1) natural language is optimized for communication rather than internal computation, inducing verbosity and inefficiency; (2) computation per token is uniform regardless of information content; (3) tokens are generated sequentially even when parallel structure might be beneficial [Ye et al. 2024]; and (4) the process is discrete and therefore non-differentiable end-to-end.

Recent work explores *continuous* or *implicit* reasoning: either distilling a discrete-CoT teacher into a student that outputs the answer directly (thus reasoning only *implicitly*) [Deng, Choi, and Shieber 2024], or explicitly generating continuous vectors that stand in for reasoning steps (Coconut [Hao et al. 2024], CODI [Shen et al. 2025]). In GSM8K [Cobbe et al. 2021], these methods appear close to discrete CoT (e.g., 52.4% for CODI vs. 55.3% with CoT for a fine-tuned 1B model), but such parity is hard to interpret because directly predicting the answer (No-CoT) yields also relatively strong results (27.6% on GSM8K). To obtain a sharper test, we evaluate in settings where the performance difference between CoT and No-CoT models is more significant. For instance, on a matrix multiplication benchmark, a 1B model achieves near-perfect accuracy with CoT reasoning but near-zero without.

Research Questions. We ask how continuous and implicit approaches behave with respect to two fundamental aspects of compositional reasoning: robustness to depth and the ability to exploit multiple reasoning paths. **RQ1 (Depth):** how does accuracy scale with chain length, the number of (dependent) intermediate steps? **RQ2 (Multi-path):** when multiple valid CoTs exist for the same problem, does multi-chain supervision improve accuracy? To probe these axes we introduce MATMUL, a synthetic benchmark with controllable depth and associativity-induced diversity, and a GSM8K multi-chain corpus with 2–5 distinct traces per question.

Key findings. Across both benchmarks, we observe consistent patterns. **(i) Depth:** On the synthetic MATMUL task, discrete CoT maintains relatively high accuracy as chain length increases, whereas implicit and continuous variants show sharp drops once intermediate steps depend on earlier ones. **(ii) Multi-path:** Continuous reasoning models do not benefit to the same extent from multi-chain supervision as their discrete counterparts.

Related Works

Explicit chain-of-thought (CoT) prompting improves multi-step reasoning by supervising intermediate steps [Wei et al. 2022]. Test-time strategies mitigate single-trace brittleness by sampling multiple chains and voting via *self-consistency* [Wang et al. 2023], structuring subproblems with *least-to-most* prompting [Zhou et al. 2023], and searching over partial thoughts with *Tree-of-Thoughts* [Yao et al. 2023]. Program- or code-grounded variants disentangle computation from explanation—e.g., *Program-of-Thoughts* and *Chain-of-Code*—reducing arithmetic errors and enabling verifiable execution [Chen et al. 2022, Li et al. 2023]. Tool-augmented prompting further interleaves reasoning with actions (ReAct) or learns API calls automatically (Toolformer), externalizing parts of the computation while preserving a discrete, inspectable trace [Yao et al. 2022, Schick et al. 2023]. On math benchmarks such as GSM8K [Cobbe et al. 2021], these

approaches implicitly aggregate over diverse derivations but retain token-level supervision.

In contrast, *implicit* and *continuous* CoT aim to "think without speaking." Distillation-based implicit CoT trains students to output answers without visible steps [Deng, Choi, and Shieber 2024], while continuous variants replace text steps with latent vectors and either anneal from discrete to latent thoughts (Coconut) or align hidden states to a CoT teacher (CODI) [Hao et al. 2024, Shen et al. 2025]. Diffusion language models explore parallel, iterative refinement of "thoughts" without left-to-right tokens [Ye et al. 2024]. These methods target verbosity and latency but are less studied under increasing depth or multi-path supervision. From the lens of knowledge distillation, aligning student representations to a teacher at limited positions can average over multi-modal internal states, risking *trace collapse* unless objectives explicitly preserve mode structure [Hinton, Vinyals, and Dean 2015, Sun et al. 2019, Jiao et al. 2020].

Work on CoT faithfulness cautions that natural-language justifications may not always reflect internal computation [Turpin et al. 2023, Lanham, Chen et al. 2023]. Our study complements these lines by stress-testing reasoning *depth* and *multi-path diversity*: discrete CoT degrades gracefully and benefits from multi-trace supervision, whereas latent/implicit methods become brittle and tend to collapse across valid traces—suggesting distributional, contrastive, or latent-variable objectives that represent a *set* of traces rather than a single alignment target.

Background and Definitions

We view compositional reasoning as solving problems that require multiple dependent intermediate steps. We refer to the number of such steps as the reasoning depth. Some tasks, such as matrix multiplication, permit several equally valid ways to combine intermediate results; we call these multi-path problems.

Discrete, Implicit, and Continuous CoT. Let $t_{1:S}$ denote a sequence of natural-language thoughts, $z_{1:S}$ a sequence of latent vectors, and y the final answer. The three paradigms differ in what is produced and supervised:

$$\begin{aligned}
\text{No-CoT: } & x \to y \quad \text{(CE on } y\text{)} \\
\text{Discrete CoT: } & x \to (t_{1:S}, y) \quad \text{(CE on } t_{1:S} \text{ and } y\text{)} \\
\text{Implicit CoT: } & x \to y, \ \text{student aligned to teacher states} \\
\text{Continuous CoT: } & x \to (z_{1:S}, y), \ \text{no supervision on } z_{1:S}
\end{aligned}$$

In alignment-style objectives, the teacher generates both reasoning tokens $t_{1:S}$ and the final answer y, while the student predicts only y. We denote their respective output logits as y_{teach} and y_{stud} and their hidden representations at the answer boundary as $h_{\text{ans}}^{\text{teach}}$, $h_{\text{ans}}^{\text{stud}}$. The total loss combines (i) cross-entropy on the teacher's full discrete trace, (ii) cross-entropy on the student's answer prediction, and (iii) a hidden-state alignment term:

$$\begin{aligned}
\mathcal{L} = \mathrm{CE}(t_{1:S}^{\text{teach}}, t_{1:S}^{*}) + \mathrm{CE}(y_{\text{teach}}, y^{*}) + \mathrm{CE}(y_{\text{stud}}, y^{*}) \\
+ \lambda \| h_{\text{ans}}^{\text{stud}} - \mathrm{sg}(h_{\text{ans}}^{\text{teach}}) \|_2^2,
\end{aligned}$$

MATMUL Example

Question: Multiply $A \cdot B \cdot C \cdot D$ with $A = \left[\begin{smallmatrix} -9 & 2 \\ 8 & -2 \end{smallmatrix}\right]$, $B = \left[\begin{smallmatrix} 2 & 6 \\ -1 & -8 \end{smallmatrix}\right]$, $C = \left[\begin{smallmatrix} 7 & 0 \\ 5 & -4 \end{smallmatrix}\right]$, $D = \left[\begin{smallmatrix} 9 & 2 \\ -9 & -3 \end{smallmatrix}\right]$.

Chains:
(1) $(A \cdot B) = \left[\begin{smallmatrix} -20 & -70 \\ 18 & 64 \end{smallmatrix}\right]$. $((A \cdot B) \cdot C) = \left[\begin{smallmatrix} -490 & 280 \\ 446 & -256 \end{smallmatrix}\right]$.
(2) $(A \cdot B) = \left[\begin{smallmatrix} -20 & -70 \\ 18 & 64 \end{smallmatrix}\right]$. $(C \cdot D) = \left[\begin{smallmatrix} 63 & 14 \\ 81 & 22 \end{smallmatrix}\right]$.
Answer: $\left[\begin{smallmatrix} -6930 & -1820 \\ 6318 & 1660 \end{smallmatrix}\right]$

GSM8K-multi-chain Example

Question: Sandy wants to lose as much as Joey. Joey loses 8 lbs in 4 weeks. Sandy needs 4 weeks to lose what Joey loses in one week.

Chains:
(1) Joey's weekly loss is $8/4 = 2$. Sandy needs 4 weeks to lose 2.
(2) Joey loses 8 in 4 weeks. Sandy needs $4\times$ the time for the same loss.

Answer: 16 weeks

Figure 1: Illustrative instances from MATMUL and GSM8K-multi-chain. For each question-answer pair, there are multiple valid CoT traces.

where $(t_{1:S}^{*}, y^{*})$ are the ground-truth reasoning tokens and answer, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. We refer to the tendency of the alignment term to average across multiple valid internal states as *trace collapse*.

Benchmarks

Our benchmarks are designed to isolate (i) how well models scale with deeper reasoning (more steps), and (ii) whether models can *represent or exploit* multiple correct chains when those exist.

Matrix Multiplication (MATMUL)

We propose a synthetic benchmark where the input is a sequence of k square matrices $A_1, \ldots, A_k \in \mathbb{R}^{n \times n}$ and the target is their product $A_1 A_2 \ldots A_k$. This task serves as an excellent proxy for structured, multi-step reasoning for several reasons:

- **Controllable Depth:** We can precisely control the number of sequential reasoning steps by increasing k. Each additional matrix requires another full computational step that depends on the previous result.
- **Algorithmic Diversity:** Matrix multiplication is associative, meaning the order of operations can be varied (e.g., $(A_1 A_2) A_3$ vs. $A_1 (A_2 A_3)$). The number of valid parenthesizations is given by the $(k-1)$-th Catalan number, C_{k-1}, providing a rich source of diverse yet valid reasoning chains. For $k = 4$, there are $C_3 = 5$ distinct paths.
- **Controllable Difficulty:** We can tune the per-step difficulty by changing the matrix dimension n or the distribution of the numbers within them.

For our experiments, we use $k \in \{3, 4\}$ with 2×2 matrices containing small integers. The dataset contains 400k training samples and 1k test samples, providing ample data to test generalization.

GSM8K Multi-Chain

Starting from the 378k augmented GSM8K corpus [Deng, Choi, and Shieber 2024], we prompt GPT-4.1 to generate 2 to 5 distinct CoT traces per question. The prompt instructs the model to produce as many *meaningfully different* traces as are sensible (not mere permutations of steps), and to omit the final answer from the trace so that the answer remains a non-trivial step derived from the CoT. This encourages the teacher to attend to the entire reasoning process when generating the answer, which was found to be beneficial in alignment-based distillation [Shen et al. 2025]. The prompted GPT-4.1 model is required to output both the reasoning chains and a final answer, without being shown the ground truth. We retain only those samples where the generated answer matches the ground-truth solution. This procedure is repeated until 100k valid instances are collected. Finally, we shuffle the order of chains to avoid positional bias, which matters for single-chain training.

Training Paradigms

We compare four families of models. **No-CoT** predicts the answer directly from the question with cross-entropy (CE) on the answer tokens. **Discrete CoT** generates a natural-language reasoning trace followed by the answer, with CE on both thoughts and answer. **Implicit CoT** distills a discrete-CoT teacher into a student that emits only the answer. **Continuous CoT** generates latent vectors as intermediate "thoughts" before the answer, with no direct supervision on these vectors.

For Implicit and Continuous CoT we evaluate two training paradigms. **Curriculum** begins from fully supervised discrete CoT and progressively (i) drops thought tokens for Implicit CoT or (ii) replaces them with unsupervised continuous vectors for Continuous CoT, ending with answer-only supervision. **Alignment** trains a discrete-CoT teacher alongside a student, aligning the student's hidden state at the answer boundary to the teacher's (stop-gradient on the teacher). We adopt prior models where available: Implicit CoT with curriculum [Deng, Choi, and Shieber 2024], Coconut for Continuous CoT with curriculum [Hao et al. 2024], and CODI for Continuous CoT with alignment [Shen et al. 2025]. For Implicit CoT with alignment we introduce a new variant, **Distill**.

Further details on training sequences, objectives, and inference procedures are given in Appendix.

Experiments and Results

We finetune *Llama-3.2-1B-Instruct* and *GPT-2* with LoRA [Hu et al. 2022], using rank 128, α=32, dropout 0.1, and a batch size of 128 sequences. For methods that involve both a teacher and student, we share model parameters between the two and stop gradients on the teacher side when computing alignment losses. For datasets with multiple valid chains per question, we sample one chain uniformly at random for each question per epoch. Evaluation is done with greedy decoding. We report exact-match accuracy.

Discrete CoT is the most robust and benefits from multiple traces. Across both models, discrete CoT is the most robust approach. For *Llama-3.2-1B-Instruct*, it achieves near-perfect performance on MATMUL with k=3 factors (97.2–95.6%) and remains strong at k=4 (60.9–60.5%). *GPT-2* follows a similar trend—CoT reaches 92.2–94.6% on MATMUL (k=3) but degrades at k=4 (43.9–46.9%). In most cases, incorporating chain diversity proves beneficial—for example, on GSM8K (100k)—yielding performance gains of 4.1% and 3.5% for Llama and GPT-2, respectively. This improvement arises because including multiple valid reasoning traces per question increases the effective dataset size.

Continuous models degrade under multi-chain supervision. Unlike CoT, which gains from diversity on the most dataset/settings, continuous methods generally degrade (e.g., CODI drops from 23.5% to 18.9% in MATMUL k=3 when trained with two chains for Llama). The main reason is that the reasoning process operates deterministically within a continuous space. Although sampling may be used during the final answer generation, every sample stems from the same reasoning trajectory, lacking the path diversity required. In the case of alignment objective, the student repeatedly sees the same question, but must match hidden states from different teacher traces. By forcing the student model to match a teacher's hidden states from different reasoning traces for the same question, the training objective incentivizes finding an average representation. Lacking a mechanism to handle multiple modes, the student collapses a rich, multimodal distribution of valid "thoughts" into a single, blurry, and ultimately incorrect mean, so additional chains are treated as variance to be suppressed instead of structure to be modeled.

No-CoT is weak and breaks on MATMUL. The No-CoT baseline confirms our hypothesis that standard benchmarks can obscure model weaknesses. While it achieves modest scores on GSM8K (17.1% for Llama and 8.2% for GPT-2 on our 100k set), it fails completely on MATMUL, scoring only (6.2%) for (k=3) (Llama) and (0%) for both (k=3) and (k=4) (GPT-2). This indicates a profound failure in compositional generalization that persists across architectures and scales, with both models unable to handle even simple structured reasoning without explicit chain-of-thought supervision.

Continuous/Implicit methods collapse like No-CoT on MATMUL. The failure of No-CoT extends to its latent counterparts. Continuous and implicit methods, behave more like No-CoT than CoT—especially with *GPT-2*. While they attain modest scores on GSM8K (e.g., 10–20%), they all collapse on MATMUL. This consistent breakdown across GPT-2 indicates that latent objectives provide no meaningful compositional signal: without explicit intermediate supervision, the model's hidden-space alignment fails to sustain multi-step reasoning, mirroring the brittleness of No-CoT rather than the robustness of discrete CoT.

Llama-3.2-1B-Instruct						
Dataset / Setting	**CoT**	**No-CoT**	**Implicit**	**Distill**	**Coconut**	**CODI**
GSM8K-aug (400k)	55.3	27.6	9.7	27.7	45.0	52.4
GSM8K-our (100k), single chain	36.2	17.1	9.1	19.0	18.6	24.8
GSM8K-our (100k), all chains	40.3	—	9.1	19.1	16.1	24.1
MatMul $3 \times (2\times2)$, single chain (400k)	97.2	6.2	0.0	21.8	5.7	23.5
MatMul $3 \times (2\times2)$, two chains (400k)	95.6	—	0.0	27.7	5.1	18.9
MatMul $4 \times (2\times2)$, single chain (400k)	60.5	0.0	0.0	1.0	0.0	0.0
MatMul $4 \times (2\times2)$, five chain (400k)	60.9	—	0.0	0.0	0.0	0.0
GPT-2						
Dataset / Setting	**CoT**	**No-CoT**	**Implicit**	**Distill**	**Coconut**	**CODI**
GSM8K-aug (400k)	42.3	20.2	3.6	8.5	22.2	42.1
GSM8K-our (100k), single chain	16.5	8.2	3.5	6.8	10.7	15.6
GSM8K-our (100k), all chains	20.0	—	3.6	7.5	10.1	14.8
MatMul $3 \times (2\times2)$, single chain (400k)	92.2	0.0	0.0	0.0	0.0	0.0
MatMul $3 \times (2\times2)$, two chains (400k)	94.6	—	0.0	0.0	0.0	0.0
MatMul $4 \times (2\times2)$, single chain (400k)	43.9	0.0	0.0	0.0	0.0	0.0
MatMul $4 \times (2\times2)$, five chain (400k)	46.9	—	0.0	0.0	0.0	0.0

Table 1: Accuracy (%) for **Llama-3.2-1B-Instruct** and **GPT-2** across datasets and training settings. *Single chain* = training includes exactly one CoT trace per question per epoch; *two/five chains* = two/five distinct traces per question are included in the training set (we still sample one per epoch); *all chains* = all available 2–5 traces per question are included (again sampling one per epoch).

Continuous/implicit methods are brittle to chain length. Distillation-style approaches (Distill, CODI) manage low-20s accuracy on MATMUL at k=3 but collapse once chains lengthen (Distill $\sim$1% and CODI 0% at k=4) for Llama. Curriculum-style approaches like Coconut is unstable (5.7% at k=3, 0% at k=4) too. On GSM8K, these methods reach middling accuracy (20–25% for Llama, 10–15% for GPT-2) but far below CoT. This brittleness reflects limitations of the training objectives. Under *curriculum* learning, the number of training stages grows with chain length, while supervision on earlier steps is progressively weakened, allowing errors to accumulate. Under *alignment*, the model receives signal only at the answer boundary, providing no guidance on how to generate or connect intermediate steps. As a result, longer chains expose the lack of effective credit assignment, leading to sharp degradation with depth.

Conclusion

Discrete chain-of-thought (CoT) remains the most reliable interface for compositional reasoning: it degrades gracefully as chains lengthen and, in low-data regimes, benefits from multi-chain supervision. In contrast, current continuous and implicit approaches are efficient but brittle: they deteriorate with depth and fail to exploit reasoning-path diversity.

Our analysis points to objective-level causes. Curriculum schedules dilute supervision on later steps, while alignment losses concentrate signal at the answer boundary. Both provide weak guidance for generating and composing intermediate steps and tend to average over distinct traces, collapsing multimodality.

Implications. For practitioners, discrete CoT is the safer choice in settings demanding robust multi-step reasoning; apparent parity of latent methods on easy benchmarks should not be mistaken for reliability. For researchers, closing the gap likely requires objectives that (i) deliver intermediate credit assignment throughout the chain and (ii) explicitly preserve and select among multiple valid reasoning modes rather than averaging them.

Limitations & Opportunities. Our study highlights core behaviors using smaller models and math-heavy tasks, providing a controlled setting that clarifies key dynamics; extending to larger models or domains with richer semantics may further enhance robustness. Promising directions include contrastive or distributional alignment over full traces, latent-variable or mixture-of-traces objectives, and supervision that attaches signal to multiple intermediate locations rather than only at the answer boundary.

References

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *arXiv:2211.12588*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Deng, Y.; Choi, Y.; and Shieber, S. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.

Hao, S.; Sukhbaatar, S.; Su, D.; Li, X.; Hu, Z.; Weston, J.; and Tian, Y. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning Workshop*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jiao, X.; et al. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *EMNLP Findings*.

Lanham, T.; Chen, A.; et al. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv:2307.13702*.

Li, C.; et al. 2023. Chain of Code: Reasoning with a Language Model-Augmented Code Interpreter. *arXiv:2312.04474*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv:2302.04761*.

Shen, Z.; Yan, H.; Zhang, L.; Hu, Z.; Du, Y.; and He, Y. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient Knowledge Distillation for BERT Model Compression. In *EMNLP*.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *NeurIPS*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *ICLR*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv:2210.03629*.

Ye, J.; Gong, S.; Chen, L.; Zheng, L.; Gao, J.; Shi, H.; Wu, C.; Jiang, X.; Li, Z.; Bi, W.; et al. 2024. Diffusion of thought: Chain-of-thought reasoning in diffusion language models. *Advances in Neural Information Processing Systems*, 37: 105345–105374.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.; and Chi, E. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *ICLR*.

Model Cards

Here we provide detailed model cards for all paradigms described in the main text. We group them into four families: No-CoT, Discrete CoT, Implicit CoT, and Continuous CoT. For the latter two families we include both curriculum-based and alignment-based variants. Each card specifies the input–output sequences, training objectives, and inference procedures.

No-CoT (Direct Answer)

Sequence: Question Answer
Training: CE on Answer.
Inference: given Question, sample Answer.

CoT (Discrete Reasoning)

Sequence: Question CoT Answer
Training: CE on CoT and Answer.
Inference: given Question, sample CoT and Answer.

Implicit CoT [Deng, Choi, and Shieber 2024] (via curriculum learning)

Sequence: Question CoT Answer
Training: curriculum: start with CE on CoT+Answer; progressively drop CoT tokens.
Inference: given Question, sample Answer.

Distill (Implicit CoT via Alignment)

Teacher Sequence: Question CoT Answer
Student Sequence: Question Answer
Training: teacher: CE on CoT+Answer; student: CE on Answer; $+\,\lambda\,$L2 alignment of hidden states at the Answer-start position (stop-grad on teacher).
Inference: given Question, sample Answer.

Coconut [Hao et al. 2024] (Curriculum to Continuous CoT)

Early Sequence: Question CoT Answer

Final Sequence: Question cont. CoT Answer

Training: curriculum: start with CE on CoT+Answer; progressively replace CoT steps with cont. CoT (no supervision on cont. CoT); end with CE on Answer only.

Inference: given Question, generate cont. CoT and sample Answer.

CODI [Shen et al. 2025] (Continuous CoT via Alignment)

Teacher Sequence: Question CoT Answer

Student Sequence: Question cont. CoT Answer

Training: teacher: CE on CoT+Answer; student: CE on Answer; $+\ \lambda$ L2 alignment of hidden states at the Answer-start position (stop-grad on teacher).

Inference: given Question, generate cont. CoT and sample Answer.