White Paper

Evolutionary Optimization of Chain-of-Thought Supervision for Small Language Models

A Data-Centric Framework on Improving Reasoning Robustness

Subha Fernando, Janaka Sendanayake

Predictiv.ai, 2026

Abstract

Small Language Models (SLMs) are increasingly deployed in environments constrained by cost, compute, latency, or privacy. Despite their practical advantages, SLMs often struggle with multi-step and compositional reasoning tasks. Chain-of-Thought (CoT) distillation—where reasoning trajectories generated by stronger teacher models are used as supervision—has emerged as an effective mechanism for improving observable reasoning behavior in SLMs. However, the effectiveness of CoT distillation is highly sensitive to the diversity, faithfulness, and structural quality of the reasoning traces used during training.

This white paper argues that evolutionary optimization, when combined with multi-teacher supervision, trajectory-aware loss shaping, and adaptive parameter-efficient layers, provides a principled, data-centric framework for improving reasoning robustness in SLMs.

Specifically, we propose:
(i) sampling CoT trajectories from multiple heterogeneous teacher models,
(ii) introducing penalty and repair mechanisms to guide student reasoning trajectories during training, and
(iii) leveraging LoRA-based adaptive layers alongside full fine-tuning to encode newly acquired reasoning patterns.

Together, these mechanisms improve the stability, faithfulness, and generalization of expressed reasoning behavior without claiming to expand the intrinsic reasoning capacity of the student model.

1. Introduction

Large Language Models (LLMs) have demonstrated strong performance on reasoning-intensive benchmarks, yet their computational and operational costs limit adoption in production, edge, and privacy-sensitive environments. Consequently, Small Language Models (SLMs) are increasingly favored for real-world deployment. However, SLMs frequently exhibit brittle reasoning behavior, particularly under distribution shift, increased reasoning depth, or domain transfer.

Chain-of-Thought (CoT) prompting and distillation aim to mitigate this gap by exposing student models to intermediate reasoning steps generated by more capable teacher models. While effective, recent studies show that CoT distillation is highly sensitive to the choice of teacher, the structure and diversity of reasoning traces, and the loss function used to enforce imitation.

Data-Centric Framing of Evolutionary CoT Supervision.

In contrast to model-centric approaches that primarily modify architectures or optimization algorithms, this work adopts a data-centric perspective on reasoning supervision. We treat the collection of teacher-generated chain-of-thought trajectories as a first-class dataset subject to explicit optimization. Under this view, the *trajectory dataset* constitutes the optimization population; *evolutionary operators* (mutation, repair, abstraction, pruning) act as structured data transformations; *fitness functions* define data selection criteria based on correctness, coherence, faithfulness, efficiency, and domain alignment; and *curriculum stages or niches* implement data stratification mechanisms that preserve structured diversity across reasoning styles and difficulty levels. Evolutionary optimization thus serves as a scalable data curation and refinement process over reasoning traces, enabling the student model to learn from a progressively filtered, corrected, and diversified supervision corpus rather than from raw, unstructured teacher outputs. This reframing positions evolutionary CoT not as a novel training algorithm per se, but as a disciplined methodology for constructing high-quality reasoning supervision data at scale.

Scope clarification:
This work does not claim to improve the fundamental reasoning capacity or representational limits of SLMs. Instead, it focuses on improving how reasoning behavior is supervised, selected, corrected, and internalized through better data curation, trajectory-level alignment, and training objectives.

Box 1: Terminology and Scope of Reasoning Objects

Terminology.
To avoid ambiguity, we standardize terminology used throughout this paper as follows:

- **Chain-of-Thought (CoT):** A textual representation of intermediate reasoning steps produced by a model while solving a task.
- **Reasoning Trajectory:** A complete ordered sequence consisting of (i) step-by-step Chain-of-Thought reasoning and (ii) the final answer. A trajectory is treated as the atomic unit of supervision and optimization.
- **Trajectory Segment:** A contiguous subset of steps within a reasoning trajectory corresponding to a partial reasoning process or sub-derivation.
- **Rationale / Trace:** Informal terms referring to reasoning steps; in this work, these are subsumed under the formal definition of reasoning trajectories or trajectory segments.
- **Repair Operation:** A targeted transformation applied at the trajectory-segment level to correct, prune, replace, or validate faulty reasoning steps while preserving the surrounding valid structure.
- **Fitness Evaluation:** A trajectory-level assessment function used to select, weight, or discard reasoning trajectories based on criteria such as correctness, coherence, faithfulness, efficiency, and domain alignment.

2. Background and Problem Statement

A. Chain-of-Thought Distillation

Traditional CoT distillation trains a student model to imitate teacher-generated reasoning traces using token-level cross-entropy loss. Variants include single-teacher distillation, multi-teacher ensembles, and self-consistency approaches. These methods implicitly assume that teacher rationales are uniformly high-quality and suitable for direct imitation.

B. Limitations of Existing Approaches

Single-teacher bias.
Distillation from a single teacher often induces stylistic overfitting and narrow reasoning priors, reducing robustness across tasks and domains.

Unstructured multi-teacher inconsistency.
While multi-teacher ensembles increase diversity, they frequently produce conflicting or incoherent reasoning trajectories without principled mechanisms for selection or reconciliation.

Trajectory-level misalignment.
Token-level imitation losses fail to distinguish between structurally valid and logically flawed reasoning paths, allowing misleading or post-hoc rationales to be learned.

Rigid parameter adaptation.
Full fine-tuning alone may entangle newly learned reasoning behaviors with existing representations, leading to catastrophic interference or unstable generalization.

These limitations motivate a shift toward trajectory-centric supervision, explicit correction mechanisms, and adaptive parameterization strategies.

3. Evolutionary Optimization of CoT Supervision

A. Multi-Teacher Prompt and Trajectory Sampling

We extend evolutionary CoT supervision by sampling reasoning trajectories from multiple heterogeneous teacher models—differing in architecture, scale, or training regime. Each teacher contributes distinct reasoning styles, abstractions, and inductive biases.

Teacher-generated trajectories form an initial population of candidate CoTs. Evolutionary selection mechanisms are then applied to:

- filter low-quality or inconsistent trajectories,
- preserve complementary reasoning styles, and
- reduce over-reliance on any single teacher's inductive bias.

This structured multi-teacher sampling mitigates single-teacher bias while avoiding naive ensemble averaging. Multi-teacher sampling increases diversity; evolutionary refinement makes that diversity usable.

B. Evolutionary Selection, Mutation, and Refinement

Following CoT-Evo-style frameworks, reasoning trajectories are treated as evolutionary individuals subject to selection, mutation, and refinement. Mutation operators include paraphrasing, step reordering, abstraction, or compression, while refinement mechanisms validate, prune, or correct invalid segments.

Fitness functions extend beyond answer correctness to include coherence, efficiency, novelty, and domain alignment, enabling principled trajectory-level optimization.

4. Trajectory-Guided Loss with Penalty and Repair

A. Limitations of Pure Imitation Loss

Token-level cross-entropy enforces surface imitation but cannot penalize logically invalid intermediate steps, encourage recovery from partial errors, or reward structurally faithful reasoning. As a result, students may learn how teachers *write* rather than how they *reason*.

B. Penalty- and Repair-Augmented Objectives

We augment standard imitation loss with **trajectory-level penalty and repair mechanisms**. Penalties down-weight trajectories exhibiting contradictions, unsupported claims, hallucinated entities, or invalid logical transitions. Repair mechanisms automatically correct or prune faulty steps, replace invalid segments with validated sub-trajectories, and enforce domain constraints.

The resulting **guided trajectory loss** softly constrains student reasoning toward valid reasoning manifolds rather than exact token-level reproduction.

5. Adaptive Knowledge Integration with LoRA Layers

A. Motivation

While full fine-tuning enables global adaptation, it risks overwriting previously learned representations, entangling task-specific reasoning with general language knowledge, and reducing modularity and interpretability.

B. LoRA as Reasoning Adaptation Layers

We introduce **LoRA-based low-rank adaptation layers** as additional reasoning adapters trained sequentially with full fine-tuning. These layers specialize in encoding newly acquired reasoning patterns, domain-specific reasoning styles, and evolved CoT structures.

In our framework, LoRA-based adaptation and full fine-tuning are applied in an *alternating curriculum-driven schedule*, where each curriculum stage introduces newly evolved or domain-specialized reasoning trajectories through dedicated LoRA layers, followed by controlled consolidation phases using limited full fine-tuning to stabilize and integrate these reasoning patterns into the base model without destabilizing prior representations.

6. Future Directions: Toward Faithful, Robust, and Deployable CoT Supervision

A. Faithfulness-Aware Fitness Functions

Future work should develop multi-objective fitness functions balancing answer correctness, internal coherence, reasoning efficiency, diversity, and faithfulness proxies such as contradiction detection, counterfactual sensitivity, and causal dependency tests.

B. Domain-Aware Diversity Preservation

Niche-preserving evolutionary strategies—such as fitness sharing, novelty-based selection, and domain-conditioned mutation—can maintain structured diversity aligned with task semantics while avoiding incoherent reasoning collapse.

C. Guardrails and Automated Repair

Future pipelines should integrate symbolic validation, retrieval augmentation, rule-based checking, and automated repair prior to trajectory selection, ensuring only structurally valid reasoning traces supervise the student.

D. Hybrid Optimization Objectives

Promising directions include combining imitation loss with trajectory-level rewards, reinforcement learning over reasoning paths, and alignment with human-evaluated metrics such as plausibility, informativeness, and efficiency.

E. Deployment-Aware Evaluation

Evaluation protocols should include out-of-distribution testing, ablation studies, and cost-normalized metrics such as accuracy per training, inference latency, and robustness–efficiency trade-offs.

7. Discussion

The framework presented in this white paper advances a unifying perspective: the dominant gains we target come from supervision/alignment rather than architectural change." By treating reasoning trajectories as first-class optimization objects, evolutionary methods expose failure modes that remain invisible under token-level training objectives.

The introduction of multi-teacher sampling addresses inductive bias and stylistic overfitting, while evolutionary selection provides a principled mechanism for reconciling heterogeneous reasoning styles. Penalty and repair mechanisms further constrain optimization toward faithful and logically valid trajectories, reducing the propagation of hallucinations and post-hoc rationales. The use of LoRA layers as adaptive reasoning overlays enables controlled integration of newly acquired reasoning behaviors without destabilizing base representations.

However, several limitations remain. Fitness functions inevitably encode heuristic proxies for faithfulness and coherence, and misalignment between these proxies and true causal reasoning may lead to over-optimization artifacts. Repair mechanisms introduce additional system complexity and may reduce transparency if not carefully designed. Moreover, while the framework improves expressed reasoning robustness, it does not address deeper questions of reasoning generalization beyond the support of available supervision.

These considerations underscore the importance of conservative claims, careful objective design, and rigorous evaluation. Evolutionary CoT supervision should be viewed not as a path toward emergent reasoning capability, but as a disciplined strategy for making existing reasoning capacity more reliable, stable, and deployable.

Bibliography

[1] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,”

in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.

[2] X. Wang et al., “Self-consistency improves chain-of-thought reasoning in language models,”

in Proc. Int. Conf. Learn. Represent. (ICLR), 2023.

[3] X. Chen et al., “Unveiling the key factors for distilling chain-of-thought reasoning,”

in Findings of the Assoc. Comput. Linguistics (ACL), 2025.

[4] X. Li et al., “Teaching small language models to reason for knowledge-intensive multi-hop question answering,” in Findings of the Assoc. Comput. Linguistics (ACL), 2024.

[5] S. Li et al., “SCOTT: Self-consistent chain-of-thought distillation,” in Proc. Assoc. Comput. Linguistics (ACL), 2023.

[6] S. Adarsh et al., “SIKeD: Self-guided iterative knowledge distillation for mathematical reasoning,” in Findings of the Assoc. Comput. Linguistics (ACL), 2025.

[7] S. Wiegreffe and Y. Pinter, “Attention is not explanation,” in Proc. Assoc. Comput. Linguistics (ACL), 2019.

[8] K. Feng, et al., “CoT-Evo: Evolutionary distillation of chain-of-thought for scientific reasoning,” OpenReview, ICLR submission, 2026.

[9] J. Jakubik et al., “Data-centric artificial intelligence,” accepted for publication in Bus. Inf. Syst. Eng.

[10] R. Zelikman et al., “STaR: Bootstrapping reasoning with reasoning,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.