How Does RL Post-training Induce Skill Composition? A Case Study on COUNTDOWN

Simon Park [*1] **Simran Kaur** [*1] **Sanjeev Arora** [1]

Abstract

While reinforcement learning (RL) successfully enhances reasoning in large language models, its role in fostering compositional generalization (the ability to synthesize novel skills from known components) is often conflated with mere length generalization. To this end, we study what RL post-training teaches models about skill composition and how the composition structure affects the skill trasnfer. We focus on the COUNTDOWN task (given n numbers and a target, form an expression that evaluates to the target) and analyze model solutions using expression trees, where each subtree corresponds to a reusable subtask and thus can be viewed as a "skill." Tracking tree shapes and their success rates over training, we find: (i) out-of-distribution (OOD) generalization to larger n and to unseen tree shapes, indicating compositional reuse of subtasks; (ii) the order of skill acquisition depends upon the composition structure—models master shallow balanced trees (workload is balanced between subtasks) before deep unbalanced ones, with persistent fragility on right-heavy structures for a fixed composition depth. Our diagnostic reveals what is learned, in what order, and where generalization fails, clarifying how RL-only post-training induces OOD generalization beyond what standard metrics such as pass@k reveal. To study the implication of our findings for more real-world tasks, we present a preliminary study on formal theorem proving.

1. Introduction

A key objective in training large language models (LLMs) is to develop robust reasoning abilities that generalize to novel problem instances. Reinforcement learning (RL) is a widely adopted technique for this purpose (Ouyang et al., 2022; Grattafiori et al., 2024; OpenAI et al., 2024; DeepSeek-AI et al., 2025), yet the precise mechanism by which it improves reasoning remains poorly understood. The prevailing hypotheses suggest that RL either sharpens the model's probability distribution over known reasoning paths (Huang et al., 2025; Yue et al., 2025) or encourages in-context exploration via chaining in longer solution sequences (DeepSeek-AI et al., 2025; Setlur et al., 2025). However, these notions overlook a crucial aspect of intelligence: the ability to compose known skills in novel ways.

Current evaluations often conflate two fundamentally different modes of generalization. The first, *length generalization*, is the ability to solve problems of greater sequential depth by iterating a known algorithmic structure. The second, *compositional generalization*, is the ability to solve problems of a fixed depth by synthesizing a novel algorithmic structure from familiar components. While complex reasoning often involves both, we argue that a rigorous understanding requires operationalizing and testing these two modes of generalization independently. Without a clear way to distinguish them, it is difficult to ascertain the true algorithmic contribution of fine-tuning methods like RL.

To formalize this distinction and enable empirical investigation, a controlled setting is required. We utilize the COUNTDOWN[1] task, which serves as a testbed where the structure of a solution is isomorphic to a binary expression tree. This allows for an unambiguous, canonical representation of the reasoning pattern used to solve a problem.

This pattern framework allows us to move beyond a monolithic view of the solution to a nuanced skill-based view. Solving a complex pattern like $A \times (B - C/D)$ requires a composition of distinct abilities: *decomposition skills* to identify how to explore and correctly partition the solution (e.g., `factor_identification`: identifying the factors of the target to explore potential candidates) and the recursive *sub-problem skills* to solve the resulting sub-patterns (e.g., generating an expression for $B - C/D$ given

[1]Princeton Language and Intelligence (PLI), Princeton University. Correspondence to: <{juhyunp,skaur}@princeton.edu>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

[1]COUNTDOWN is a task where given n integers and a target, the model needs to output an expression using each integer once and operators $\{+, -, \times, \div\}$ that attains the target. For example, given $[4, 3, 2, 2]$ and a target 16, one solution would be $2\times2\times3+4 = 16$.

the remaining three numbers and a modified target).[2] The tree structure of the pattern determines which atomic skills are combined, in what order, and how they are grouped.

By tracking the development of complex skills for each pattern during training, we probe two fundamental questions: 1) To what extent does RL equip models for genuine structural composition? and 2) How does the underlying composition structure of a problem affect how quickly a model learns to solve it? We summarize our main contributions:

- **Pattern framework.** We introduce a framework to disentangle length and compositional generalization (Section 2). We propose a method to remove distributional bias in COUNTDOWN datasets (Section 2.3). This may inspire future studies on other domains with a subtask structure.
- **Models exhibit two forms of OOD generalization.** Using our framework, we separately show that models generalize to *larger puzzle sizes* (Section 4.1), where skills for smaller patterns need to be reused for more number of times, and to *unseen patterns* (Section 4.2), where skills for other patterns are combined to synthesize a held-out pattern and its higher-order extensions.
- **Pattern structure determines difficulty.** Upon fixing an input length n, we observe that the difficulty of a pattern depends on its structure: shallow balanced structures (with equal amount of "workload" across subtasks) are easier than deep unbalanced structures (Section 4.3). Within the same depth, right-heavy trees, which require committing to a plan ahead of the complex subroutine appearing later, are particularly hard (Section 4.4). Supervised finetuning (SFT) exhibits a different order of skill acquisition, suggesting that the hierarchy is inherent to RL, not an artifact of the task (Appendix B.2).
- **The effect of composition structure extends to formal theorem proving.** We show that an RL-trained model for formal theorem proving favors proofs with balanced subtasks, while avoiding proofs with right-heavy structure (Section 4.5). This suggests that RL post-training selectively amplifies a model's behavior based on the difficulty induced by the composition structure.

Overall, our findings show that RL post-training imparts length and compositional generalization, with difficulty governed primarily by composition structure.

[2]The atomic skills would be mastering the four operators: $a+b$, $a-b$, $a \times b$, and $a \div b$.

2. Methodology: Formalizing Length and Compositional Generalization

Standard metrics for reasoning, such as *pass@k*, *response length*, or *number of attempts* offer an incomplete picture of a model's capabilities. They fail to distinguish between a model's ability to iterate a known procedure and the more formidable challenge of synthesizing a novel one. To address this, we introduce a formal framework that disentangles these phenomena by analyzing the underlying computational structure of arithmetic solutions, which we apply to the COUNTDOWN task.

2.1. Canonical Patterns of Reasoning

We define an *atomic skill* as the application of a single binary operator from $\{+, -, \times, \div\}$. Then the ability to solve a COUNTDOWN puzzle with an arithmetic expression requires a *skill composition*. By replacing numbers with placeholder symbols, an arithmetic expression can be abstracted (e.g., $9/3 + 1 \rightarrow$ `A/B+C`) and parsed into a binary tree, which provides insights into how the atomic skills are combined.

To ensure that syntactically different but algorithmically identical expressions (e.g., $(A + B) + C$ and $C + (A + B)$ are treated as a single computational procedure, we apply a principled normalization process that maps them to a unique canonical *pattern* (Figure 1). This process, detailed in Appendix A.1, resolves ambiguities arising from properties like commutativity and associativity, allowing us to analyze the underlying computational graph and distinguish a simple iterative structure like $A + B + C + D$ from a complex, nested one like $A/(B - C/D)$.

We classify the structure of a computational pattern by its shape signature $[x] \circ [y]$, denoting the root operator $\circ$ and the number of leaf operands x and y in its left and right subtrees. This partitioning into left-heavy ($x > y$), balanced ($x \approx y$), and right-heavy ($x < y$) structures correlates with synthesis difficulty (see Figure 5). For instance, to generate a right-heavy pattern like $A/(B-C/D)$ (signature $[1] \div [3]$), a model cannot greedily select the root operator; it must first look ahead to the complex sub-expression $(B - C/D)$ and recognize its utility within the broader computation.

2.2. Disentangling Modes of Generalization

A primary contribution of our framework is its ability to formally disentangle two distinct and often-conflated axes of skill composition. We define **length generalization** as the ability to generalize to larger problem sizes than those seen during training, which typically requires using more atomic skills. We define **compositional generalization** as the ability to reconstruct novel combinations of known atomic skills at problem sizes that remain within the training range.

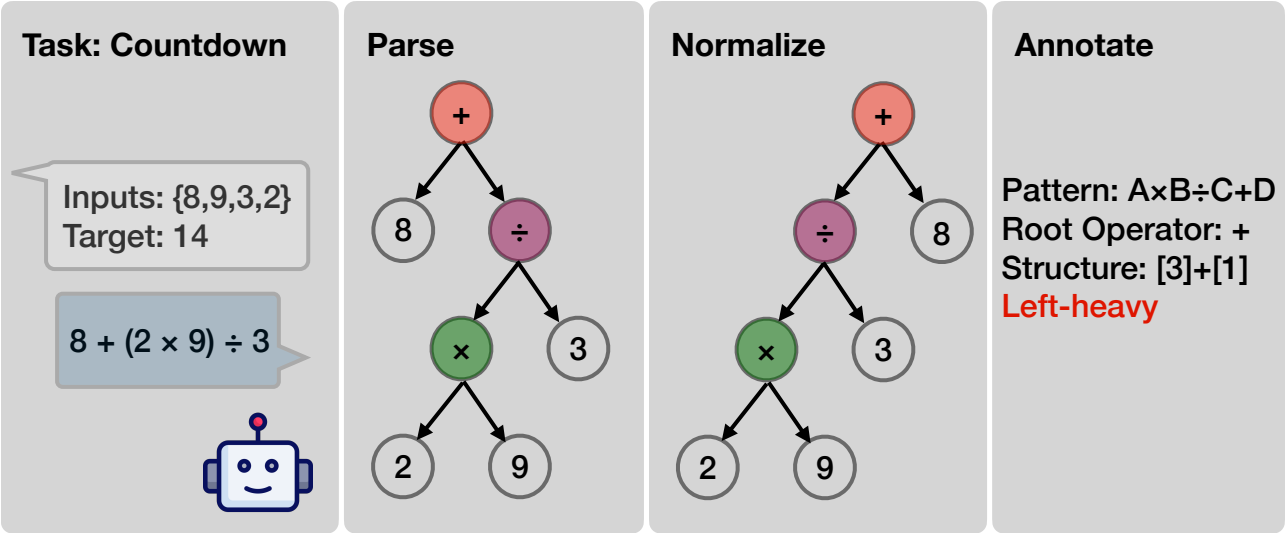

Figure 1. **From generated expression to canonical *pattern*.** The model-generated expression $8 + (2 \times 9) \div 3$ is parsed into an expression tree, normalized with respect to algebraic identities, and mapped to the unique canonical pattern $A \times B \div C + D$. Patterns are then grouped by their computational shape. Here, the root operator is "$+$", with a left sub-expression over three leaves (A, B, C) and a right sub-expression over one leaf (D). The signature $[3] + [1]$ serves as a compact representation of the pattern's structure at the root-level.

In COUNTDOWN, we quantify length generalization by the model's success rate at puzzles larger than those it was trained on, and compositional generalization by its success rate on held-out patterns (holding the puzzle size fixed).

Our detailed study reveals interesting phenomena that were missed by earlier works (e.g., Yuan et al. (2025)), which consider composition simply as a sequential application ("chaining") of atomic skills. For example, these works do not differentiate between the difficulty of atomic skills ($+$ and $\div$ are considered equal in difficulty, which violates intuition).

2.3. Controlling for Distributional Bias in Dataset Generation

A rigorous analysis of structural difficulty requires that the distribution of problem types be controlled. Standard approach for creating COUNTDOWN datasets (Pan et al., 2025b; Gandhi et al., 2024; Stojanovski et al., 2025) proceeds as follows: 1) choose a random binary tree; 2) choose random operators, numbers; 3) then filter for valid expressions that evaluate to an integer in $[1, 99]$. This introduces severe distributional biases, creating a confounding variable between a pattern's frequency and its intrinsic complexity. There are two problems with this design:

- **Structural bias.** $+$ and $\times$ are commutative and associative, and many expressions involving the two operators may have the same underlying pattern and have a higher probability of being generated. For instance, the same $A+B+C+D$ puzzle can be generated in $4! \times C_3 = 120$ ways (e.g., $((1+2)+3)+4$ and $1+(4+(3+2))$), whereas an asymmetric pattern like $A/(B - C/D)$ only has one expression.
- **Selection bias.** The second, more severe bias stems from the common practice of filtering for solutions that evaluate to an integer within a fixed range, such as $[1, 99]$. When random inputs are drawn from the set range, $+$ or $-$ yields a result in this range approximately 50% of the time, whereas $\times$ or $\div$ do so only about 5% of the time. Consequently, each $\times$ or $\div$ is approximately $10\times$ more unlikely to appear in an expression.

These two biases together produce a skewed dataset where aggregate performance metrics are dominated by a model's performance on simple, additive patterns. To eliminate this simplicity bias, we designed a different generation protocol. We generate examples by first selecting a target canonical pattern and then sampling numbers that satisfy it. This enforces a near-uniform distribution over all patterns, a methodological prerequisite for isolating and analyzing the true effect of compositional structure on reasoning difficulty. See Appendix A.2 for more details.

3. Experimental Setup

3.1. Training

We finetune `Qwen-2.5` models (1.5B, 3B, and 7B) (Qwen et al., 2025) (Apache-2.0) with GRPO (Shao et al., 2024) on problems of size $n \in \{3, 4\}$. Each rollout receives a reward of 0.1 for correctly formatting the final expression and an additional reward of 0.9 if the expression is correct

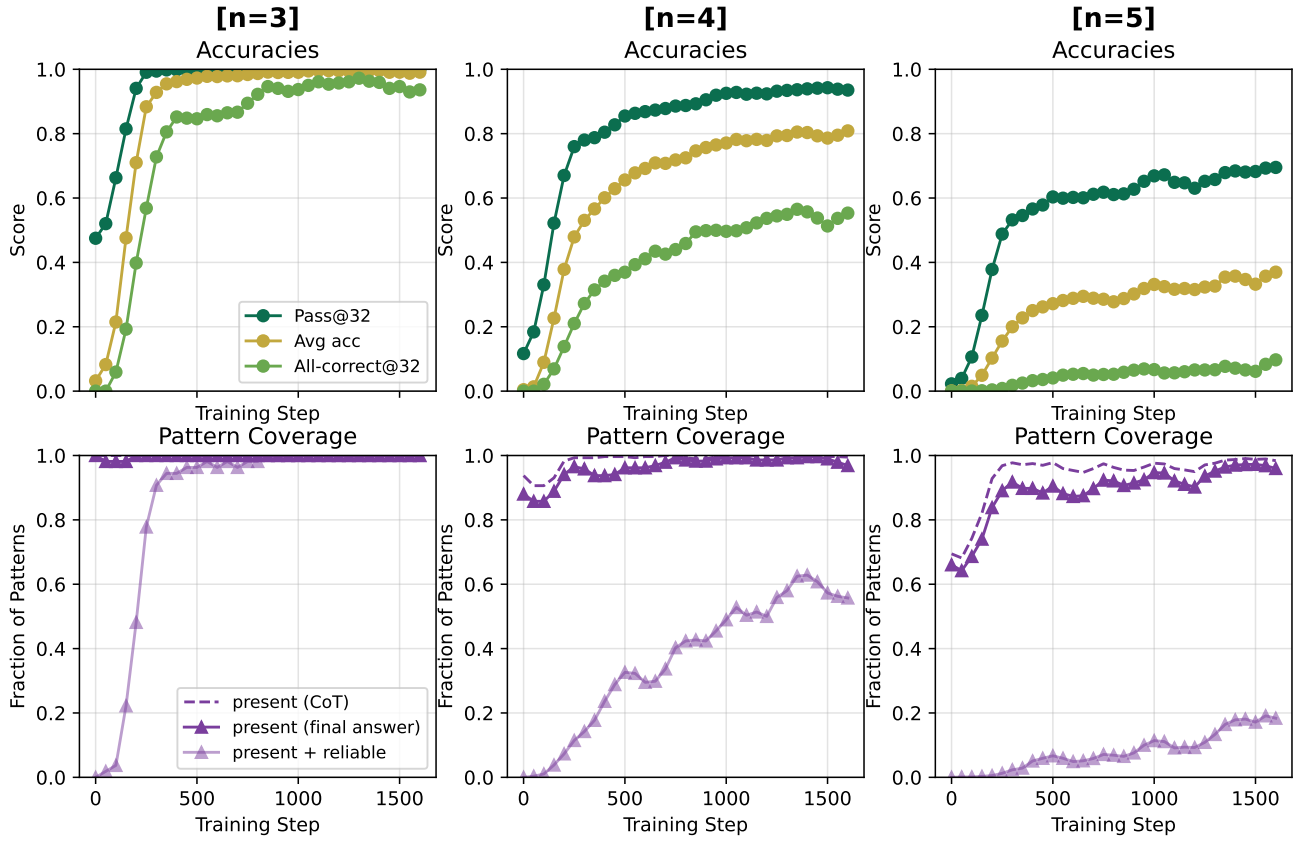

Figure 2. **Successful length generalization, but a gap between first producing a pattern and producing it reliably.** We train `Qwen2.5-1.5B` on problem sizes $n \in \{3, 4\}$ and evaluate on held-out data for $n \in \{3, 4, 5\}$. **(Top)** The model demonstrates strong *length generalization*: accuracy on the held-out $n = 5$ problems improves over the base model (leftmost data point). **(Bottom)** The model quickly learns to produce most $n = 5$ patterns at least once, but the ability to reliably produce the correct pattern (see Section 3.2 for the definition) lags significantly behind.

and valid.[3] We train for one epoch across three seeds, saving checkpoints every 50 gradient updates and discarding any unstable runs (see Appendix C.1). Full training details and results on other model families are available in Appendices A.2.3, A.3 and D.

3.2. Evaluation

We evaluate each checkpoint by sampling $k = 32$ outputs per held-out question (temperature $= 0.6$), with the maximum token length increasing for larger problem sizes (1024 for $n \in \{3, 4\}$, 2048 for $n = 5$, and 4096 for $n = 6$). We then compute the following metrics on the extracted final answers:

- **Pass@32 / All-correct@32:** The fraction of questions where at least one (Pass@32), or all (All-correct@32), of the k sampled answers is correct.
- **Average Accuracy:** The total number of correct answers across all questions and attempts, divided by the total number of answers generated (i.e., number of questions $\times\, k$).
- **Pattern Coverage:** The fraction of all possible canonical patterns that the model produces at least once across the evaluation set (such patterns will be labeled `present`).
- **Per-Pattern Precision:** The total number of correct solutions using a specific pattern, divided by the total number of times the model produced that pattern as a final answer across the entire evaluation set.
- **High-Precision Coverage:** The fraction of all possible canonical patterns for which the model's *per-pattern precision* is at least 80% (such patterns will be labeled `reliable`).[4]

[3] An expression is *valid* if each of the n provided numbers is used exactly once and operators are restricted to $\{+, -, \times, \div\}$. An expression is *correct* if it evaluates to the target number.

[4] The 80% threshold checks if the model can execute a pattern with a reasonable reliability, but allows some mistakes. The main findings are robust to reasonable variations of the threshold (70-90%). See Appendix G.

4. Main Results

We show that models achieve length (Section 4.1) and compositional generalization (Section 4.2) independently. We then demonstrate that this success is fundamentally governed by the composition structure of the task (Section 4.3). We identify a "lookahead bottleneck" as the primary barrier to skill composition (Section 4.4). We compare against a SFT baseline and find that the "lookahead bottleneck" is inherent to RL, not an artifact of the task (Appendix B.2). Finally, to show real-world application of our findings, we present extensions to formal theorem proving (Section 4.5).

4.1. Models Demonstrate Length Generalization

RL effectively teaches models to apply learned reasoning procedures to problems of greater lengths than seen during training. Models trained only on $n \in \{3, 4\}$ problems show substantial accuracy improvements on held-out $n = 5$ puzzles (Figure 2) over the base model, with final Pass@32 reaching nearly 70% for the 1.5B model (Table 1). This generalization extends to $n = 6$, where models achieve Pass@32 around 40%. These results confirm that models successfully achieve *length generalization*.

4.2. Models Demonstrate Compositional Generalization

We next test whether RL can enable a model to synthesize a novel reasoning pattern it has never been explicitly trained on, *while holding the length constant*. We remove the $n = 3$ pattern $A/B + C$ (a pattern rarely produced by the base model) and its entire family of $n = 4$ derivatives from the training dataset (e.g., $A/B+C+D$, $A/(B+C)+D$). See Appendix C.4 for the full list.

The results provide evidence for compositional generalization. `Qwen2.5-1.5B` trained on this dataset successfully learns to generate the held-out patterns despite zero training exposure (Figures 3, 23 and 24). The learning dynamics also reveal: coverage first emerges on the simpler $n = 3$ sub-pattern before the model learns to compose it into its more complex $n = 4$ dependent forms. This demonstrates that the model is not merely matching seen templates but is actively reusing learned substructures to assemble novel patterns, fulfilling our definition of *compositional generalization*.

4.3. Composition Structure Determines Difficulty

However, the improvement in an aggregate accuracy masks the true determinant of difficulty. The ability to generalize in length is fundamentally constrained by the composition structure of the pattern. This is evident in the out-of-distribution $n = 5$ evaluation in Figure 4, which reveals the same hierarchy of difficulty seen on the training distribution (with $n = 4$): balanced patterns are learned most readily, while right-heavy patterns remain the most challenging. This demonstrates that composition structure, not problem length, is the primary predictor of reasoning difficulty, even when generalizing to unseen lengths.

Table 1. **Performance across model size.** Pass@32 averaged across all successful seeds [5] (**final** checkpoints for 1.5B/3B and **best** checkpoints for 7B) [6] on n=3, 4 (trained) and n=5, 6 (held out). Models learn $n = 3, 4$ patterns almost perfectly and generalize to larger puzzle sizes $n = 5, 6$. See Appendix B.1 for a more complete version of the table.

Model	n=3	n=4	n=5	n=6
1.5B	1.00	0.94	0.68	0.36
3B	1.00	0.95	0.68	0.40
7B	1.00	0.98	0.75	0.39

4.4. The "Lookahead Bottleneck" Is a Fundamental Barrier to Skill Composition

To isolate the effect of composition structure while holding problem length constant, we analyze performance on $n = 4$ puzzles. Our findings reveal a clear difficulty hierarchy that points to a "lookahead bottleneck" as a fundamental challenge for autoregressive models.

Figure 4 ($n = 4$ column) shows that balanced patterns ($[2] \circ [2]$), which decompose a problem into two equal subtasks, are mastered far more reliably than unbalanced ones. Crucially, right-heavy patterns ($[1] \circ [3]$) are substantially harder to learn than left-heavy ($[3] \circ [1]$) ones. To generate a solution for a right-heavy pattern like $A/(B - C/D)$, the model must commit to the root operator ($\div$) before generating the complex, three-term subroutine it applies to. This need to plan ahead for a complex subsequent operation is the primary failure mode.

Table 2 quantifies this bottleneck across all model sizes: at the same depth of 3, $[1] \circ [3]$ generally underperforms $[3] \circ [1]$, especially with the root operator $\div$ and with smaller models.

4.5. Preliminary Evidence Beyond COUNTDOWN: Formal Theorem Proving

We next ask whether our observations on COUNTDOWN also appear in a more realistic reasoning domain. We consider formal theorem proving in Lean (de Moura and Ullrich), where many proof strategies naturally decompose into subtasks. In particular, we focus on tactics (e.g,.

[5] 2 seeds for 1.5B/3B and 3 for 7B. Due to compute budget, we evaluate only 1 seed per model for $n = 6$.

[6] 7B models are over-trained before the end of 1 epoch of training. We select the checkpoint right before the training reward declines. See Appendix B.1 for more details.

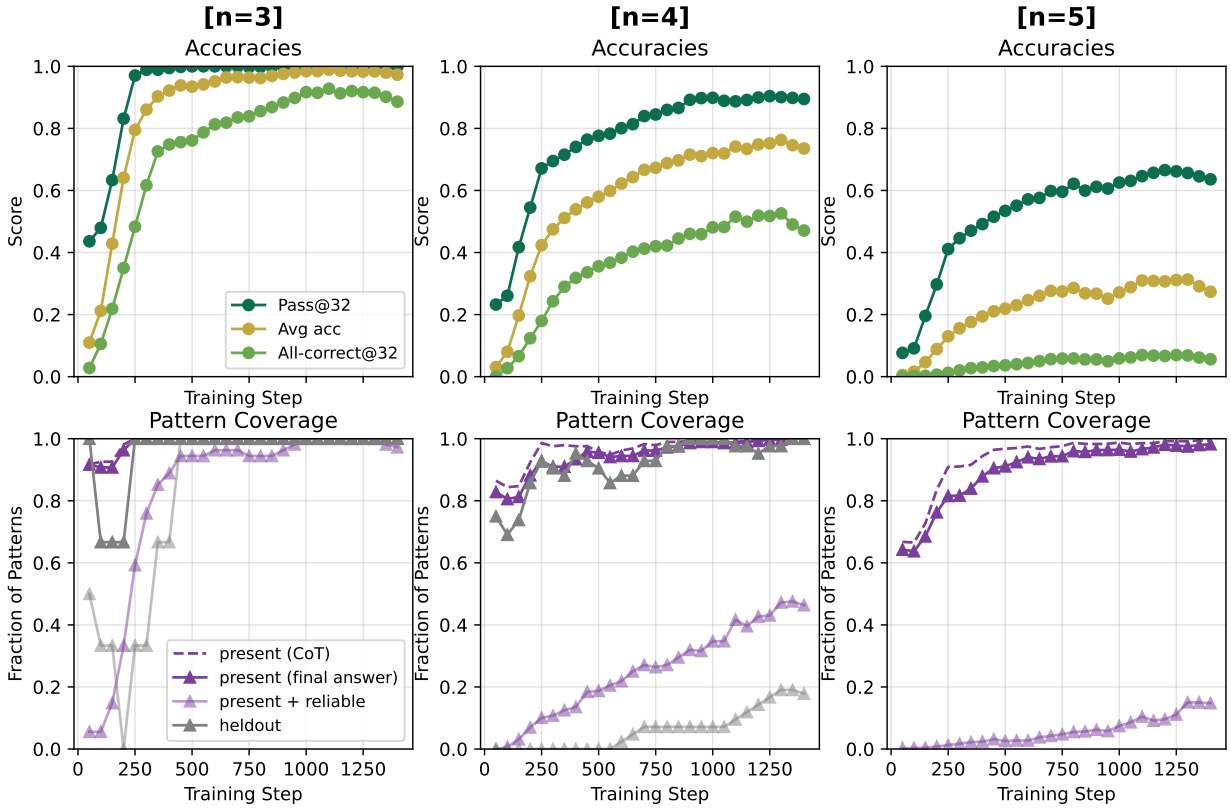

Figure 3. **Generalizes to entire families of held-out compositional patterns.** To evaluate compositional generalization, we removed an entire family of related patterns from the training set: the base $n = 3$ structure $A/B + C$ and all of its $n = 4$ extensions (e.g., $A/B + C + D$, $A/(B + C) + D$). The model successfully recovers these unseen patterns. Coverage first emerges on the held-out $n = 3$ subpattern before generalizing to its more complex $n = 4$ dependents, matching the typical learning hierarchy where presence precedes precision. This result demonstrates that the model can reuse learned substructures to synthesize novel operator-tree shapes it has never been explicitly trained on.

`constructor`) that require the model to generate exactly two subtasks. [7] A proof attempt can be categorized as balanced, left-heavy, or right-heavy based on the relative amount of workload in the two subtasks. To isolate the effect of RL post-training, we compare the behavior of the `Goedel-Prover-V2-32B` (Lin et al., 2026) (Apache-2.0) and its behavior before the final stage of RL training. [8] For more details, see Appendix B.3.

In Table 3, we provide three proof distributions: 1) SFT dataset (a proxy for the output before RL training); 2) output of the final checkpoint (after RL training) on the same SFT dataset (in-distribution); and 3) output of the final checkpoint on an out-of-distribution dataset. Compared to pre-RL training, the final checkpoint prefers balanced (51% $\rightarrow$ 56%) and left-heavy (14% $\rightarrow$ 16%) proofs and avoids right-heavy proofs (35% $\rightarrow$ 29%). This trend is consistent with the "lookahead bottleneck" identified in COUNTDOWN: if balanced proofs were easier to produce and right-heavy patterns were harder, RL post-training would have amplified the correct behavior (balanced), while inhibiting the incorrect behavior (right-heavy).

This trend is even more prominent on an out-of-distribution dataset, curated to contain an equal number of balanced, left-heavy, and right-heavy theorems (based on human reference proofs): the model attempts balanced proofs on 68% of the theorems, while generating right-heavy proofs on only 13% of the problems.

5. Ablations

In Appendix C, we run ablations on the different part of the training pipeline. Small models often find a formatting-only shortcut, but increasing the group size initially mitigates the instability (Appendix C.1). Normalizing by the standard deviation in the group advantage leads to longer responses and a lower final performance (Appendix C.2); increasing

[7] For example, if the original theorem is in the form of $P \wedge Q$, the `constructor` tactic decomposes the theorem into proving P and Q separately.

[8] Since the SFT checkpoint is not available, we use the distribution of the SFT dataset as the proxy for the distribution of proofs of the SFT checkpoint (before RL training).

Table 2. **Per-pattern precision averaged by pattern structure on n=4.** Averaged across all successful seeds (**final** checkpoints for 1.5B/3B and **best** checkpoints for 7B). Averaged by (a) root operator; (b) pattern with root operator $-$; (c) pattern with root operator $\div$. Models struggle more on $-$, $\div$ as the root operator. Shapes in the form of $[1] - [3]$ or $[1] \div [3]$ are particularly challenging. Within each column, the highest number is **boldfaced** and the second highest number is underlined.

(a) by root operator

Operator	1.5B	3B	7B
$+$	0.86	0.85	**0.91**
$-$	0.83	0.80	0.86
$\times$	**0.90**	**0.86**	**0.91**
$\div$	0.75	0.78	0.83

(b) by subtree shape ($-$)

Shape	1.5B	3B	7B
$[3] - [1]$	0.81	0.78	0.84
$[2] - [2]$	**0.92**	**0.89**	**0.90**
$[1] - [3]$	0.73	0.76	0.86

(c) by subtree shape ($\div$)

Shape	1.5B	3B	7B
$[3] \div [1]$	0.77	0.77	0.83
$[2] \div [2]$	**0.88**	**0.85**	**0.91**
$[1] \div [3]$	0.61	0.76	0.80

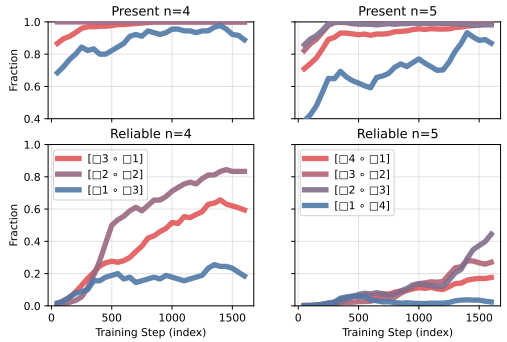

Figure 4. **Composition structure (not input length) determines difficulty.** Within $n = 4$, balanced patterns ($[2] \circ [2]$; purple) are discovered (top row) and mastered more reliably (bottom row) than the unbalanced ones (red/blue). Balanced patterns have shallow trees, with solution depth of 2 equal to that of $n = 3$ puzzles. Even when controlling for depth, right-heavy patterns ($[1] \circ [3]$; blue) are harder than left-heavy patterns ($[3] \circ [1]$; red). This evidence points to a "lookahead bottleneck," the challenge of committing to a solution ahead of a complex subroutine. The entire structural hierarchy persists for $n = 5$. Results shown for a representative run of `Qwen2.5-1.5B`. See Appendix G for equivalent plots for all other runs.

the number of rollouts may improve stability but leads to a lower final performance (Appendix C.3); models require harder and a diverse set of patterns to escape the formatting-only shortcut (Appendix C.4); models trained without a random subset of patterns can still generalize to unseen patterns and larger pattern sizes (Appendix C.4); training with PPO may improve All-correct@32 but hurts Pass@32 (Appendix C.5).

6. Related Work

Skill composition in LLMs Arora and Goyal (2023) establish a theoretical foundation for the emergence of skill composition in LLMs, and Yu et al. (2024) benchmark the ability of the models to compose skills. Regarding how to elicit skill composition, previous works focus mainly on pre-training and supervised fine-tuning: Chen et al. (2023) propose optimizing data ordering, whereas other works investigate in-context prompting (Chen et al., 2024) and supervised fine-tuning (Zhao et al., 2024). This work extends the analysis of skill composition to the RL post-training stage.

Table 3. **Distribution of proofs by composition structure (%).** After RL post-training, the final checkpoint prefers balanced (B) and left-heavy (L) proofs and avoids right-heavy (R) proofs. This trend is consistent with the "lookahead bottleneck" identified in COUNTDOWN: structures whose complex subroutine appears later are harder to produce reliably and are therefore less likely to be reinforced. The last row shows that the same qualitative shift persists on an out-of-distribution dataset which contains an equal number of reference proofs across balanced / left-heavy / right-heavy categories.

	B	L	R
SFT dataset	51	14	35
Model output (ID)	56 +5	16 +2	29 -6
Model output (OOD)	68 +17	19 +5	13 -22

Length and compositional generalization Previous works (Anil et al., 2022; Zhou et al., 2024; Lee et al., 2025) study the ability of LLMs to generalize to a sequence length longer than seen during supervised training. Similar length generalization capability has been studied for RL post-training, albeit sometimes conflated with compositional generalization. DeepSeek-AI et al. (2025); Setlur et al. (2025) claim that RL helps models generalize to more difficult problems by encouraging more attempts, considering length as a proxy for difficulty. Sun et al. (2025); Zhang et al. (2025); Yuan et al. (2025); Chu et al. (2025) study the "compositional ability" of RL, but their axis of generalization is limited to the number of skills or compositional depth, which coincides with length. While Cheng et al. (2025) study true compositional generalization, they do not study length generalization.

Our work instead proposes a framework to study length and compositional generalization separately. The abstract tree representation of each skill allows us to diagnose how the order of skill acquisition is determined by the composition structure, not just its length or depth.

Sharpening vs discovery and pass@k. Previous works (Yue et al., 2025; Dang et al., 2025) often use the metric of pass@k to label the effect of post-training as either *sharpening* (Huang et al., 2025)—concentrating probability on pre-existing good paths—or *discovery*—acquiring new knowledge. While useful, trajectory-level metrics like pass@k are too coarse to provide detailed insight into model behavior. Our work focuses on the learning dynamics, which are compatible with both possibilities. Instead of concluding whether RL is introducing a new type of reasoning or resurfacing it, we focus more on skill composition: *how* and *in which order* it happens.

COUNTDOWN and the game of 24. Prior works (Yao et al., 2023; Gandhi et al., 2024; Herr et al., 2025; Ni et al., 2025) consider COUNTDOWN, and its variant Game of 24, as a testbed for model reasoning due to its clean structure. They mainly focus on optimizing inference-time search or training on heuristics-based search traces, but these approaches can make it difficult to identify which skills acquired versus which are a part of the search algorithm / heuristics. By contrast, our work focuses on *natural RL* on COUNTDOWN—no handcrafted search heuristics or specialized decoding—so the only signal is the natural task reward. This design lets us attribute behavioral changes directly to RL and to analyze them with the pattern framework.

Design choices in RL training. RL training is sensitive to algorithmic choices. The loss function in GRPO (Shao et al., 2024) involves multiple terms, including group-level advantage normalization, KL regularization, and gradient clipping. Follow-up works (Liu et al., 2025; Yu et al., 2025; Shrivastava et al., 2025) analyze the effect of these choices; for example, arguing that removing standard-deviation normalization can improve stability and avoid biases toward easy prompts. Our work complements these works by providing additional empirical observations on the effect of the design choices in RL.

7. Discussion and Future Work

Our work studies the acquisition of skill composition during RL. On natural data, compositional complexity is confounded with length, making it impossible to isolate true failure modes. A synthetic testbed is therefore methodologically essential to control for these confounders and identify drivers of reasoning failure. Thus, the COUNTDOWN task introduces a formal lens through which to analyze the acquisition of compositional skills. By representing solutions as computational trees, we can disentangle the effects of a problem's length from its underlying compositional structure. This decomposition reveals a fundamental insight: the primary barrier to learning complex reasoning are not length-based but structural. Models do not fail simply because a task is long; they fail when its compositional form creates specific bottlenecks.

Our analysis reveals that problems decomposable into balanced sub-problems are learned most readily. In contrast, those requiring deep-sequential commitments (particularly "right-heavy" structures that demand significant lookahead before executing a complex subroutine) constitute a fundamental bottleneck for autoregressive models; crucially, these right-heavy structures are often harder than left-heavy ones, despite having the same compositional depth. This demonstrates that a task's compositional structure, rather than merely its size, is what dictates difficulty. This finding should be understood not as an artifact of our task, but as highlighting a plausible explanation of where difficulty in extending beyond the reasoning boundary arises. Our results suggest that the inherent challenge is tied to the structure of skill composition itself, not to incidental features of the data.

Our preliminary observation on formal theorem proving hints that this structural view of difficulty likely extends to other domains–such as program synthesis or multi-step tool use–where solutions can be represented as computation or proof trees. In such settings, structural properties (balance, depth, and lookahead/sequential dependencies) may predict difficulty. We argue that the overall sample complexity of learning skill composition is not monolithic; instead, it is dominated by the challenge of mastering these structurally-hard instances.

This insight points to curriculum design as a promising direction for future work. We hypothesize that, once RL post-training yields signs of length and some compositional generalization, dynamically augmenting the training distribution with a minimal set of structurally hard instances may further advance the model's capability frontier. Verifying this hypothesis and developing such curricula remain a crucial direction for future work.

8. Limitations

The scope of the paper is primarily on COUNTDOWN. While we extend our analysis to real-world settings (formal mathematical reasoning), we only do a post-hoc evaluation on an existing RL checkpoint; there may be confounding factors that we were not able to remove. Also, to remove any bias from the dataset distribution, we train on a uniform random mix of a balanced dataset; we do not consider any data curriculum, which may further accelerate the rate of learning or reveal other learning dynamics of RL.

9. Acknowledgements

Supported by DARPA (AIQ program, DARPA CMO AWD8860). We thank Dr. Pat Shafto of AIQ Program for useful discussions. SA also acknowledges support from NSF, ONR, and the Schmidt Foundation. Additionally, SP is partially supported by the Gordon Wu Fellowship and the Kwanjeong Educational Foundation Scholarship. We thank Abhishek Panigrahi, Yun Cheng, Xingyu Zhu, Haoyu Zhao, Zixuan Wang, Noam Razin, and Sanghyuk Chun for helpful discussions and feedback.

10. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

4nums.com. Puzzle difficulties. https://www.4nums.com/game/difficulties/, 2025. Retrieved: August 14, 2025.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35: 38546–38556, 2022.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. Skills-in-context: Unlocking compositionality in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13838–13890, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.812. URL https://aclanthology.org/2024.findings-emnlp.812/.

Mayee F Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! A data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Sitao Cheng, Xunjian Yin, Ruiwen Zhou, Yuxuan Li, Xinyi Wang, Liangming Pan, William Yang Wang, and Victor Zhong. From atomic to composite: Reinforcement learning enables generalization in complementary reasoning, 2025. URL https://arxiv.org/abs/2512.01970.

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. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dYur3yabMj.

Xingyu Dang, Christina Baek, Kaiyue Wen, J Zico Kolter, and Aditi Raghunathan. Weight ensembling improves reasoning in language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=S2IKxulLT1.

Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language (system description).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia

Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D. Goodman. Stream of search (sos): Learning to search in language, 2024. URL https://arxiv.org/abs/2404.03683.

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.

Nathan Herr, Tim Rocktäschel, and Roberta Raileanu. Llm-first search: Self-guided exploration of the solution space. *arXiv preprint arXiv:2506.05213*, 2025.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WJaUkwci9o.

Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. Self-improving transformers overcome easy-to-hard and length generalization challenges. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ZtX0MBT6mf.

Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, Jiayun Wu, Jiri Gesi, Ximing Lu, David Acuna, Kaiyu Yang, Hongzhou Lin, Yejin Choi, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j4C0nALrgK.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Tianwei Ni, Allen Nie, Sapana Chaudhary, Yao Liu, Huzefa Rangwala, and Rasool Fakoor. Teaching large language models to reason through learning and forgetting. *arXiv preprint arXiv:2504.11364*, 2025.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz,

Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Jiayi Pan, Xiuyu Li, Long Lian, Charlie Snell, Yifei Zhou, Adam Yala, Trevor Darrell, Kurt Keutzer, and Alane Suhr. Learning adaptive parallel reasoning with language models. *arXiv preprint arXiv:2504.15466*, 2025a.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025b. Accessed: 2025-01-24.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

John Schulman. Approximating kl divergence. http://joschu.net/blog/kl-approx.html, 2020. Retrieved: August 25, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Amrith Setlur, Matthew Y. R. Yang, Charlie Snell, Jeremy Greer, Ian Wu, Virginia Smith, Max Simchowitz, and Aviral Kumar. e3: Learning to explore enables extrapolation of test-time compute for llms, 2025. URL https://arxiv.org/abs/2506.09026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Vaishnavi Shrivastava, Ahmed Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. Sample more to think less: Group filtered policy optimization for concise reasoning. *arXiv preprint arXiv:2508.09726*, 2025.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards, 2025. URL https://arxiv.org/abs/2505.24760.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization, 2025. URL https://arxiv.org/abs/2506.18880.

The mathlib Community. The Lean Mathematical Library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, CPP 2020, New Orleans, LA, USA, January 2020. ACM. doi: 10.1145/3372885.3373824. URL https://doi.org/10.1145/3372885.3373824.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. LeanDojo: Theorem proving with retrieval-augmented language models. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-mix: a flexible and expandable family of evaluations for AI models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From $f(x)$ and $g(x)$ to $f(g(x))$: Llms learn new skills in rl by composing old ones. *arXiv preprint arXiv:2509.25123*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models, 2025. URL https://arxiv.org/abs/2512.07783.

Haoyu Zhao, Simran Kaur, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Can models learn skill composition from examples? *Advances in Neural Information Processing Systems*, 37:102393–102427, 2024.

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024.

A. Experimental Setup (More Details)

A.1. Expressions to Patterns

Here are the heuristics (in plain English) to map an expression to its canonical form.

- Remove any unnecessary parentheses. For example, map $(A + B) + C$ to $A + B + C$.
- If possible, use $+$ and $\times$ instead of $-$ and $/$. For example, map $A - (B - C)$ to $A - B + C$.
- For commutative operations ($+$ and $\times$), the operands are sorted in decreasing order of the "weight" of the operand (i.e., the number of symbols involved in the intermediate term). For example, $(A + B) \times C$ is preferred over $A \times (B + C)$.
- If in the same intermediate level of computation, $+$ and $\times$ should appear before $-$ and $/$. For example, $A + B - C$ is preferred over $A - B + C$. For example, $(A + B) \times (A - B)$ is preferred over $(A - B) \times (A + B)$.

The full list of mapping is given in Tables 15 to 18. In Appendix F, we show that the choice of the canonicalization scheme does not affect the main findings of the paper.

A.1.1. A Tree View of Expressions and Patterns

Each **expression** can be viewed as a binary tree, with each node representing an operation and each leaf node representing a symbol. To define congruence between equivalent expressions, we convert them to a signed n-ary tree:

- Each node now represents a chain of addition/subtraction operations ($\oplus$) or a chain of multiplication/division operations ($\otimes$). Each node can now connect to any number of operands that are being computed in that chain. For example, $A + B + C$ is represented with one node with 3 children.
- Each edge is assigned a positive sign if it connects to an operand that is being added/multiplied in the chain or a negative sign if connects to an operand that is being subtracted/divided in the chain.
- To convert from a binary tree to a n-ary tree:
  - For nodes representing addition/multiplication, mark both edges as positive. For nodes representing subtraction/division, mark the left edge as positive but the right as negative.
  - Recursively merge any pair of parent/child if they belong in the same chain. The sign of the edges of the child node will be flipped if the edge between the parent and the child was negative.

If the binary trees have the same representation in the signed n-ary tree, they are considered equivalent. The **pattern** (canonical form) can be retrieved as follows:

- At any level, sort the children by
  1. sign of the edge between the shared parent and the child (positive comes first);
  2. the weight of the subtree rooted at the child (heavy comes first);
  3. the number of positive edges to the grandchildren (more positive edges comes first).
- Do an in-order traversal of the tree. Print the first node as A, the second node as B, and so forth. Include parentheses only if a child subtree has weight > 1 and is rooted in a $\oplus$ node.

A.1.2. Number of Expressions and Patterns

Given n, the number of distinct **expressions** is $C_{n-1} \times 4^{n-1}$ where C_n is the Catalan number. There are C_{n-1} ways to assign the order of $n - 1$ operations (equivalently, choose the structure of the binary tree). For example, when $n = 4$, there are 5 ways: $((AB)C)D$, $(A(BC))D$, $(AB)(CD)$, $A((BC)D)$, and $A(B(CD))$. There are additionally 4^{n-1} different ways to choose the operations.

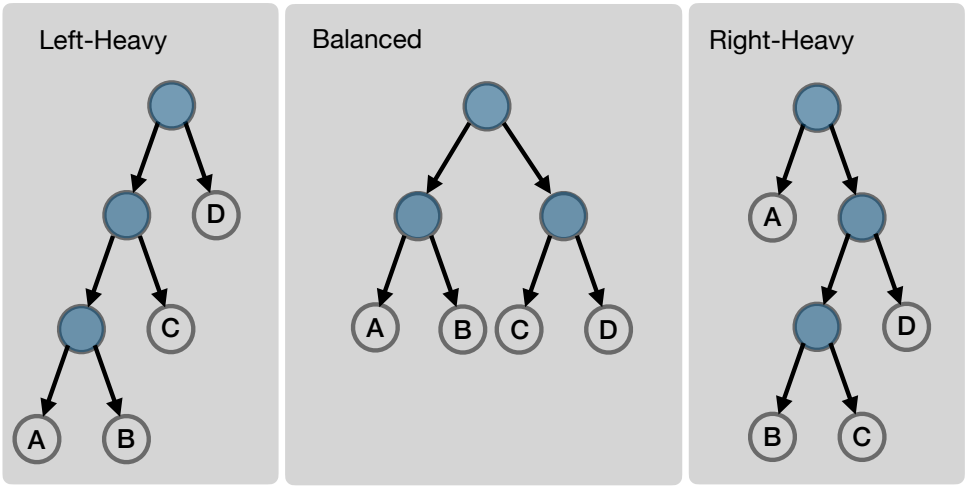

Figure 5. **Grouping canonical patterns by tree structure ($n = 4$).** Internal (blue) nodes are operators and leaves (A, B, C, D) are operands. We categorize patterns by the relative size of the root's subtrees: left-heavy, balanced, or right-heavy.

The number of distinct **patterns** T_n is equal to the number of trees with n leaf nodes with the following constraints: 1) any child of a $\oplus$ node must be a leaf node or a $\otimes$ node (and vice versa); 2) no node shall have all negative edges. We let A_n, M_n respectively denote the number of such trees with a $\oplus$ root node and a $\otimes$ root node. We let $A_1 = 1$ (for the single leaf A) and $M_1 = 0$. For $n \geq 2$, T_n, A_n, M_n can be computed with the following recursive definition.

$$\begin{cases} A_n = n + \sum\limits_{p \in \bigcup\limits_{2 \leq k < n} P(n,k)} \left[(g_1 + 1) \left(\prod\limits_{j=1}^{m_p} \binom{g_{p,j} + 2M_{d_{p,j}} - 1}{g_{p,j}} \right) - \left(\prod\limits_{j=1}^{m_p} \binom{g_{p,j} + M_{d_{p,j}} - 1}{g_{p,j}} \right) \right] \\ M_n = n + \sum\limits_{p \in \bigcup\limits_{2 \leq k < n} P(n,k)} \left[(g_1 + 1) \left(\prod\limits_{j=1}^{m_p} \binom{g_{p,j} + 2A_{d_{p,j}} - 1}{g_{p,j}} \right) - \left(\prod\limits_{j=1}^{m_p} \binom{g_{p,j} + A_{d_{p,j}} - 1}{g_{p,j}} \right) \right] \\ T_n = A_n + M_n \end{cases}$$

where

- $P(n, k)$ is the unordered set of partitions of n into a sum of k positive integers
- Each partition $p \in \bigcup\limits_{2 \leq k < n} P(n, k)$ is written as

$$p = \overbrace{(d_{p,1} + d_{p,1} + \cdots + d_{p,1})}^{g_{p,1} \text{ times}} + \cdots + \overbrace{(d_{p,m_p} + \cdots + d_{p,m_p})}^{g_{p,m_p} \text{ times}} + \overbrace{(1 + \cdots + 1)}^{g_1 \text{ times}}$$

  where

  - m_p is the number of distinct part sizes in p (other than 1)
  - $d_{p,j} > 1$ are the distinct part sizes in p
  - $g_{p,j}$ are their corresponding multiplicities, and g_1 is the multiplicity for 1

The first n corresponds to assigning $\{0, 1, \cdots, n-1\}$ negative edges to n leaf nodes directly connected to the root node (which we assume to be $\oplus$ without loss of generality). Then for any partition $p \in P(n, k)$ where $2 \leq k < n$, any $d_{p,j} > 1$ corresponds to a $\otimes$ child node with weight $d_{p,j}$, whereas $d_{p,j} = 1$ corresponds to a leaf node directly connected. If $d_{p,i} \neq d_{p,j}$, then the corresponding $\otimes$ child nodes are distinguishable, so we can independently count the number of choices for each child node.

For each $d_{p,j} > 1$, there could be multiple $\otimes$ child nodes that have the same weight and can have one of $M_{d_{p_j}}$ shapes. If they have the same shape, they are indistinguishable. For each $i = 1, 2, \cdots, M_{d_{p,j}}$, let $x_i \geq 0$ denote the number of the

child nodes that have the corresponding shape such that

$$x_1 + x_2 + \cdots + x_{M_{d_{p,j}}} = g_{p,j}$$

For each x_i, there are $x_i + 1$ distinguishable ways to assign a sign to each child node ($\{0, 1, \cdots, x_i\}$ negative edges). Therefore, we apply the stars and bars formula to get:

$$\begin{aligned}
&\sum_{\substack{x_1+\cdots+x_{M_{d_{p,j}}}=g_{p,j} \\ x_i \geq 0\ \forall i}} (x_1+1)(x_2+1)\cdots(x_{M_{d_{p,j}}}+1) \\
&= \sum_{\substack{y_1+\cdots+y_{M_{d_{p,j}}}=g_{p,j}+M_{d_{p,j}} \\ y_i \geq 1\ \forall i}} y_1 y_2 \cdots y_{M_{d_{p,j}}} \\
&= \binom{(g_{p,j}+M_{d_{p,j}})+M_{d_{p,j}}-1}{2M_{d_{p,j}}-1} \\
&= \binom{g_{p,j}+2M_{d_{p,j}}-1}{g_{p,j}}
\end{aligned}$$

When $d = 1$, the leaf children are all indistinguishable and there are $g_1 + 1$ ways to assign a sign to each leaf. We finally need to subtract the case where we assigned a negative edge to all possible $g_{p,j}$ child nodes of all possible weight of $d_{p,j}$. For each choice of x_i, there is exactly 1 way to assign a negative sign to all child nodes. Again, by the stars and bars formula, we get

$$\sum_{\substack{x_1+\cdots+x_{M_{d_{p,j}}}=g_{p,j} \\ x_i \geq 0\ \forall i}} 1 = \binom{g_{p,j}+M_{d_{p,j}}-1}{g_{p,j}}$$

A.2. Dataset

A.2.1. Why is the Existing Dataset Lacking?

Pan et al. (2025b) do not release the exact generation code for their dataset. So instead, we sample 10000 examples from the training split of the dataset and count the number of examples that can be solved with a given pattern (if one example can be solved with multiple patterns, increment the count of all such patterns). 18 of the 114 patterns never appear in the dataset (i.e., none of the examples can be solved with the patterns), and some patterns occur much more often than others (Figure 6). In the test split (the last 1000 examples), there are 36 out of 114 patterns that never appear.

One reason behind the imbalance is that almost half of the dataset consists of examples for $n = 3$, even though there are only 18 patterns for $n = 3$ compared to 96 for $n = 4$. Another explanation is that the requirement that the target number is an integer within $[1, 99]$ biases towards addition and subtraction—multiplication would frequently cause the expression to be larger than the bound; division would frequently make the final value non-integer.

Two other codebases that provide the generation code for Countdown (Gandhi et al., 2024; Stojanovski et al., 2025) wrongly assume that intermediate expressions cannot evaluate to non-integer values. This incorrectly removes puzzles like $8 \div (3 - 8 \div 3) = 24$ which are very difficult even for humans (4nums.com, 2025) and should be used to test against models.

A.2.2. What Happens When Training on the Existing Dataset?

We train `Qwen2.5-3B` on the dataset by Pan et al. (2025a) with the same set of hyperparameters and evaluate on both our and their held-out data. Since we do not explicitly check if our held-out data is in their training data, the results on our held-out data may be slightly higher than expected. In Figure 7, we recreate Figure 2.

Even though the model performance seems to smoothly improve, the final performance is much lower than training on our balanced dataset, which stems from the model failing to reliably learn most $n = 4$ patterns. This suggests that the diversity of the data and the presence of challenging patterns helps the model identify reusable components. Additionally, many patterns are not present when evaluating on the held-out data from Pan et al. (2025a), precisely because none of the puzzles test those patterns. This points to another deficiency in the existing datasets for Countdown that test simple additive skills, rather than a well-rounded portfolio of arithmetic reasoning.

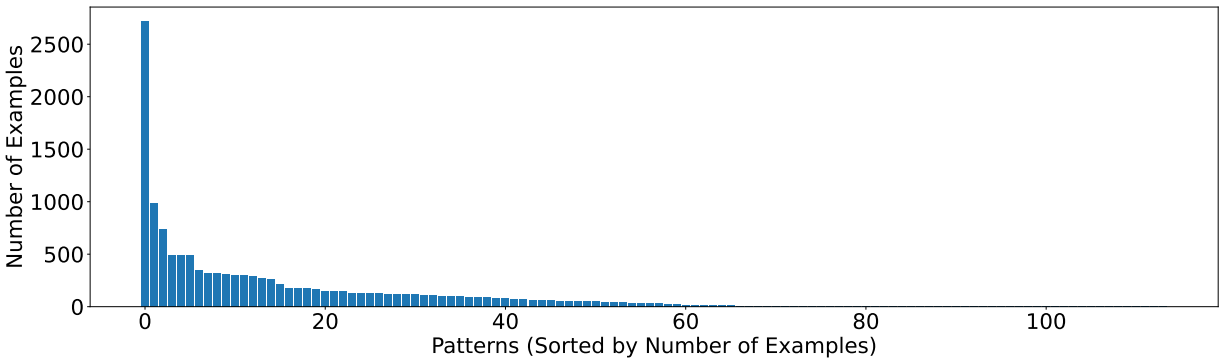

Figure 6. **Number of examples per pattern in Pan et al. (2025b).** 18 out of 114 patterns do not appear in a randomly selected 10000 examples from the dataset. Few patterns (mostly from $n = 3$) appear disproportionally frequently.

A.2.3. Creating a Balanced Dataset

To create our balanced dataset, we first choose a pattern and then plug in n numbers that are sampled independently and uniformly at random from $[1, 99]$. If the output of the expression is an integer in $[1, 99]$, we include it in our dataset. *For each pattern*, we repeat this process until we have 4000 distinct examples or have sampled 10 million combinations. There are only 15 patterns (4 from $n = 3$ and 11 from $n = 4$) where we sample less than 4000 examples. The exact numbers are given in Table 4. For each pattern, we select the first 10 examples as heldout examples. In total, we have 418619 examples for training and 1140 examples for testing. We modify the codebase from Pan et al. (2025b) (Apache-2.0 license).

For evaluation on $n = 5$, we similarly sample 10 examples for each of 558 patterns and for $n = 6$, 1 example for each of 4328 patterns.

Table 4. **Number of training examples per pattern.** Any pattern not listed has 4000 examples.

Pattern	Number of Examples	Pattern	Number of Examples
$(A + B) \times C$	3333	$A \times B \times C$	315
$A/(B + C)$	3905	$A/B/C$	791
$(A \times B + C) \times D$	2314	$(A + B) \times C \times D$	927
$A \times B \times C \times D$	188	$A/(B \times C + D)$	2704
$A/(B + C)/D$	1058	$A/B/C/D$	264
$A/B - C \times D$	2546	$(A + B) \times (C + D)$	647
$A/(B + C) - D$	932	$A/B/C - D$	1033
$(A/B - C)/D$	2802	-	-

A.3. Training

We train with batch size and mini batch size 256, prompt and response length 1024, learning rate 10^{-6}, and KL coefficient 10^{-3}, where the KL divergence is computed with the k_3 estimator (low variance) (Schulman, 2020). Following Liu et al. (2025), we do not normalize by the standard deviation in the GRPO advantage computation. We use $n = 4$ rollouts except during a warmup phase for 1.5B and 3B models (Appendix C.1). For the PPO experiments in Appendix C.5, we use a critic learning rate of 10^{-5} and PPO mini bach size of 64. For all other hyperparameters, we use the default values from the `verl` package version 0.5.0.dev0 (Sheng et al., 2024). See Appendix E for the prompt template used. Training the 1.5B model for 1 epoch takes roughly 8 hours on 4 H100 GPUs.

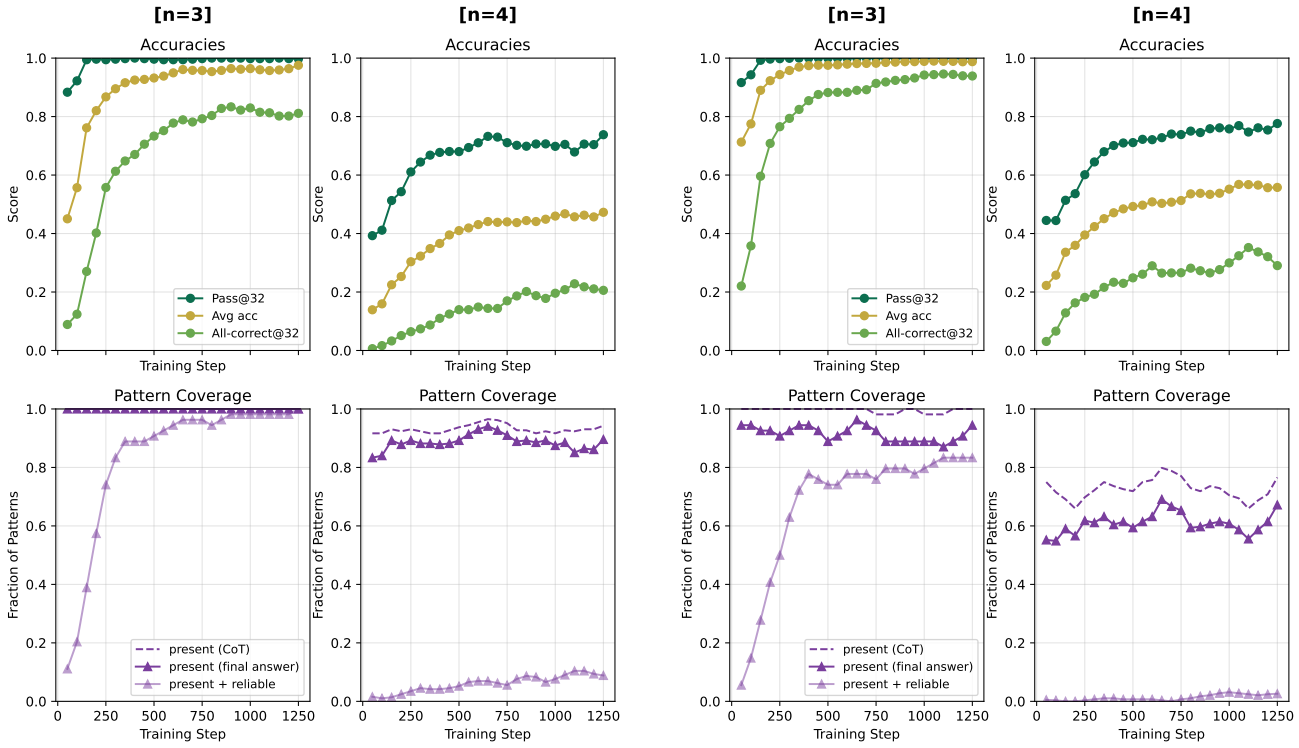

Figure 7. **Learning dynamics at a glance (Qwen2.5-3B)** trained on dataset from Pan et al. (2025a). **(Left)** evaluated on our dataset; **(Right)** evaluated on the held-out data from Pan et al. (2025a). Even though model accuracy smoothly increases (top), models do not reliably learn most $n = 4$ patterns and fails to generate at least 10% of $n = 4$ patterns (bottom). The held-out data from Pan et al. (2025a) do not include 36 possible patterns and fail to test the models on the corresponding skills.

B. Main Results (Continued)

B.1. Additional Tables

1.5B/3B models smoothly improve performance throughout training (1635 steps), whereas the 7B model reaches peak performance before 1 epoch. In the main paper, we reported the performance of the checkpoints right before the training reward declines (1350/950/650 steps respectively for each seed). For completeness, we also provide the performance of the final checkpoints of the 7B model in Tables 5 and 6. The 7B model can eventually improve on right-heavy patterns ($[1]-[3]$ or $[1] \div [3]$) but at the cost of sacrificing performance on balanced and left-heavy patterns. Additionally, we report the standard deviation when averaging across random seeds.

In Table 5, we also report the performance of the base models before GRPO training. Note that the Pass@32 for these models is very close to $1-(1-p)^{32}$ where $p =$ Average Accuracy, suggesting that the models are randomly guessing the output. At the same time, within each output, the model tends to randomly guess multiple expressions, until it runs out of the generation budget. Since there is only a small subset of expressions that are valid for $n = 3, 4$ (up to $18 \times 3! = 108$ for $n = 3$ and $96 \times 4! = 2304$), the random baseline for the base models is quite high (assuming 10 guesses per output, $p \approx 0.1$ for $n = 3$ and $p \approx 0.05$ for $n = 4$). The observed performance of the base models is not far from the random baseline.

Table 5. **Performance across model size (Full version).** Averaged across all successful seeds (**final** (F) checkpoints for 1.5B/3B and **best** (B) and **final** (F) checkpoints for 7B). All-correct@32 (All), Average Accuracy (Avg), and Pass@32 (Pass) on n=3, 4 (trained) and n=5, 6 (held out). Models learn $n = 3, 4$ patterns almost perfectly and generalize to larger puzzle sizes $n = 5, 6$, but the gap between Pass@32 and Average Accuracy increases with n, signaling that the generalization is not perfect.

Model	All	Avg	Pass	All	Avg	Pass
	n=3			n=4		
1.5B	0.00	0.02	0.42	0.00	0.00	0.00
1.5B (F)	0.97 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.60 ± 0.09	0.82 ± 0.02	0.93 ± 0.00
3B	0.00	0.05	0.66	0.00	0.01	0.20
3B (F)	0.93 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	0.56 ± 0.06	0.81 ± 0.03	0.94 ± 0.01
7B	0.00	0.10	0.88	0.00	0.02	0.36
7B (B)	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.61 ± 0.07	0.86 ± 0.03	0.98 ± 0.01
7B (F)	0.95 ± 0.02	0.99 ± 0.00	1.00 ± 0.00	0.56 ± 0.03	0.80 ± 0.06	0.92 ± 0.06
	n=5			n=6		
1.5B	0.00	0.00	0.02	0.00	0.00	0.01
1.5B (F)	0.13 ± 0.05	0.36 ± 0.03	0.68 ± 0.00	0.00 ± 0.00	0.09 ± 0.00	0.36 ± 0.00
3B	0.00	0.00	0.03	0.00	0.00	0.00
3B (F)	0.12 ± 0.02	0.39 ± 0.01	0.68 ± 0.02	0.02 ± 0.00	0.15 ± 0.00	0.40 ± 0.00
7B	0.00	0.00	0.08	0.00	0.00	0.02
7B (B)	0.15 ± 0.02	0.46 ± 0.04	0.75 ± 0.03	0.02 ± 0.00	0.15 ± 0.00	0.39 ± 0.00
7B (F)	0.10 ± 0.02	0.35 ± 0.05	0.61 ± 0.10	0.01 ± 0.00	0.12 ± 0.00	0.36 ± 0.00

Table 6. **Per-pattern precision averaged by pattern structure on n=4 (Full version).** Averaged across all successful seeds (**final** checkpoints for 1.5B/3B and **best** and **final** checkpoints for 7B). Averaged by (a) root operator; (b) pattern with root operator $-$; (c) pattern with root operator $\div$. Models struggle more on $-$, $\div$ as the root operator. Shapes in the form of $[1] - [3]$ or $[1] \div [3]$ are particularly challenging. Within each column, the highest number is **boldfaced** and the second highest number is underlined.

(a) by root operator

Operator	**1.5B**	**3B**	**7B** (best)	**7B** (final)
$+$	$\underline{0.86} \pm 0.08$	$\underline{0.85} \pm 0.01$	$\mathbf{0.91} \pm 0.01$	$\underline{0.83} \pm 0.03$
$-$	0.83 ± 0.09	0.80 ± 0.04	0.86 ± 0.06	0.76 ± 0.11
$\times$	$\mathbf{0.90} \pm 0.01$	$\mathbf{0.86} \pm 0.03$	$\mathbf{0.91} \pm 0.01$	$\mathbf{0.84} \pm 0.08$
$\div$	0.75 ± 0.04	0.78 ± 0.06	0.83 ± 0.04	0.81 ± 0.08

(b) by subtree shape ($-$)

Shape	**1.5B**	**3B**	**7B** (best)	**7B** (final)
$\boxed{3} - \boxed{1}$	$\underline{0.81} \pm 0.13$	$\underline{0.78} \pm 0.03$	0.84 ± 0.04	0.72 ± 0.12
$\boxed{2} - \boxed{2}$	$\mathbf{0.92} \pm 0.01$	$\mathbf{0.89} \pm 0.00$	$\mathbf{0.90} \pm 0.07$	$\mathbf{0.86} \pm 0.10$
$\boxed{1} - \boxed{3}$	0.73 ± 0.00	0.76 ± 0.09	$\underline{0.86} \pm 0.12$	$\underline{0.82} \pm 0.10$

(c) by subtree shape ($\div$)

Shape	**1.5B**	**3B**	**7B** (best)	**7B** (final)
$\boxed{3} \div \boxed{1}$	$\underline{0.77} \pm 0.05$	$\underline{0.77} \pm 0.05$	$\underline{0.83} \pm 0.05$	0.83 ± 0.05
$\boxed{2} \div \boxed{2}$	$\mathbf{0.88} \pm 0.10$	$\mathbf{0.85} \pm 0.00$	$\mathbf{0.91} \pm 0.07$	$\mathbf{0.89} \pm 0.14$
$\boxed{1} \div \boxed{3}$	0.61 ± 0.03	0.76 ± 0.12	0.80 ± 0.17	0.80 ± 0.05

B.2. Isolating the Effect of RL

To test whether the effect of the composition structure on the speed of skill acquisition is an artifact of the task or an effect of RL post-training, we additionally compare with supervised finetuning (SFT). To prepare the answer column, we replace the placeholder symbols in the desired pattern with the correct list of numbers. For the optional chain of thought text, we 1) randomly select an integer $k \in [2, n]$ based on the puzzle size n; 2) for each of k incorrect attempts, randomly choose a pattern different from the desired pattern and replace the placeholder symbols with a randomly shuffled list of the input numbers; 3) evaluate the expression and append "`(Incorrect)`"; 4) append the correct expression and append "`(Correct)`." The CoT is explicitly designed to be random to remove any bias from hand-coded heuristics. See Appendix E for example prompts.

Final accuracies. Table 7 reports the final accuracies of `Qwen2.5-1.5B` trained with SFT. The model performs better without the CoT, most likely because including the random patterns confuses the model. Note that these results are only provided as reference and are not a core part of the analysis. Instead, we focus on how the order of skill acquisition shifts when the model is trained on the same task, without access to its own exploration.

SFT exhibits different learning dynamics by pattern structure. The model trained without CoT fails to reliably learn most $n = 4$ patterns. To probe the order in which the model learns the patterns to an intermediate level, we instead relax the threshold of high-precision to 50% (Figure 8). Even though the coverage of patterns for $n = 5$ patterns decrease in a similar order as in RL: right-heavy > left-heavy > balanced, the model instead masters $n = 4$ patterns (in-distribution) in the exact opposite order: right-heavy < left-heavy < balanced. This highlights that the clear hierarchy of difficulty caused by the subtree structure is inherent to the RL training, not an artifact of the dataset.

Table 7. **Final performance of SFT-trained models.** Pass@32 on n=3, 4 (trained) and n=5 (held out). We present the performance of the base model and the GRPO-trained model as a basis for comparison.

Training	n=3	n=4	n=5
SFT+No CoT	0.92	0.82	0.33
SFT+Random CoT	0.59	0.25	0.04
Base Model	0.42	0.00	0.02
GRPO	1.00	0.94	0.68

B.2.1. Training

We train with batch size 256, prompt and response length 1024, max learning rate 10^{-6} with a linear warmup of 0.03 and cosine decay to zero. For all other hyperparmeters, we use the default values from the `verl` package. We train the `Qwen2.5-1.5B` model with and without the random chain of thought text, each with 1 random seed for dataset shuffling.

B.2.2. Learning dynamics

In Figures 8 to 10, we present the learning dynamics plots for the `Qwen2.5-1.5B` model trained with SFT (equivalent to Figure 4).

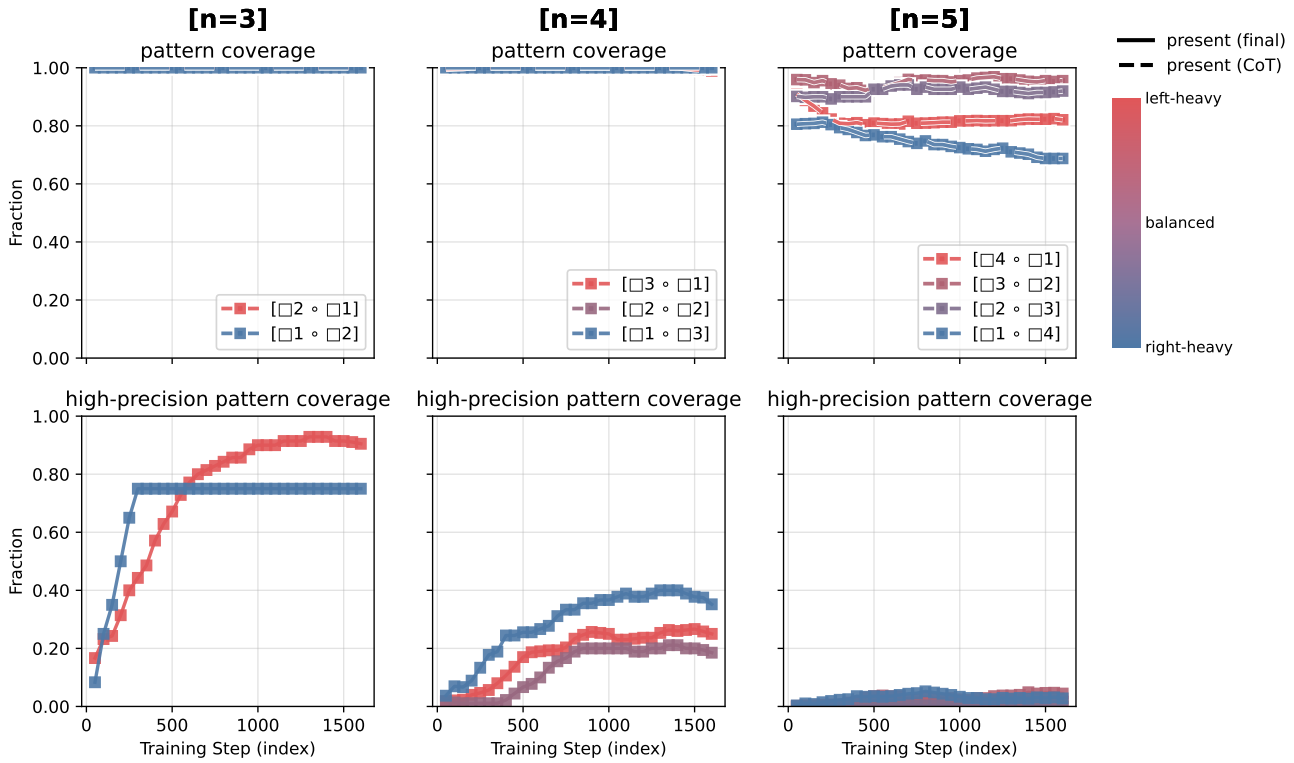

Figure 8. **Learning dynamics at a glance (Qwen2.5-1.5B with SFT)**. Here, the high-precision threshold has been lowered to 50%. Even though the coverage of $n = 5$ patterns (OOD) decrease in the following order: balanced > left-heavy > right-heavy, the model masters $n = 4$ patterns (in-distribution) in the opposite order: right-heavy > left-heavy > balanced. This shows a fundamental difference in the learning dynamics from RL.

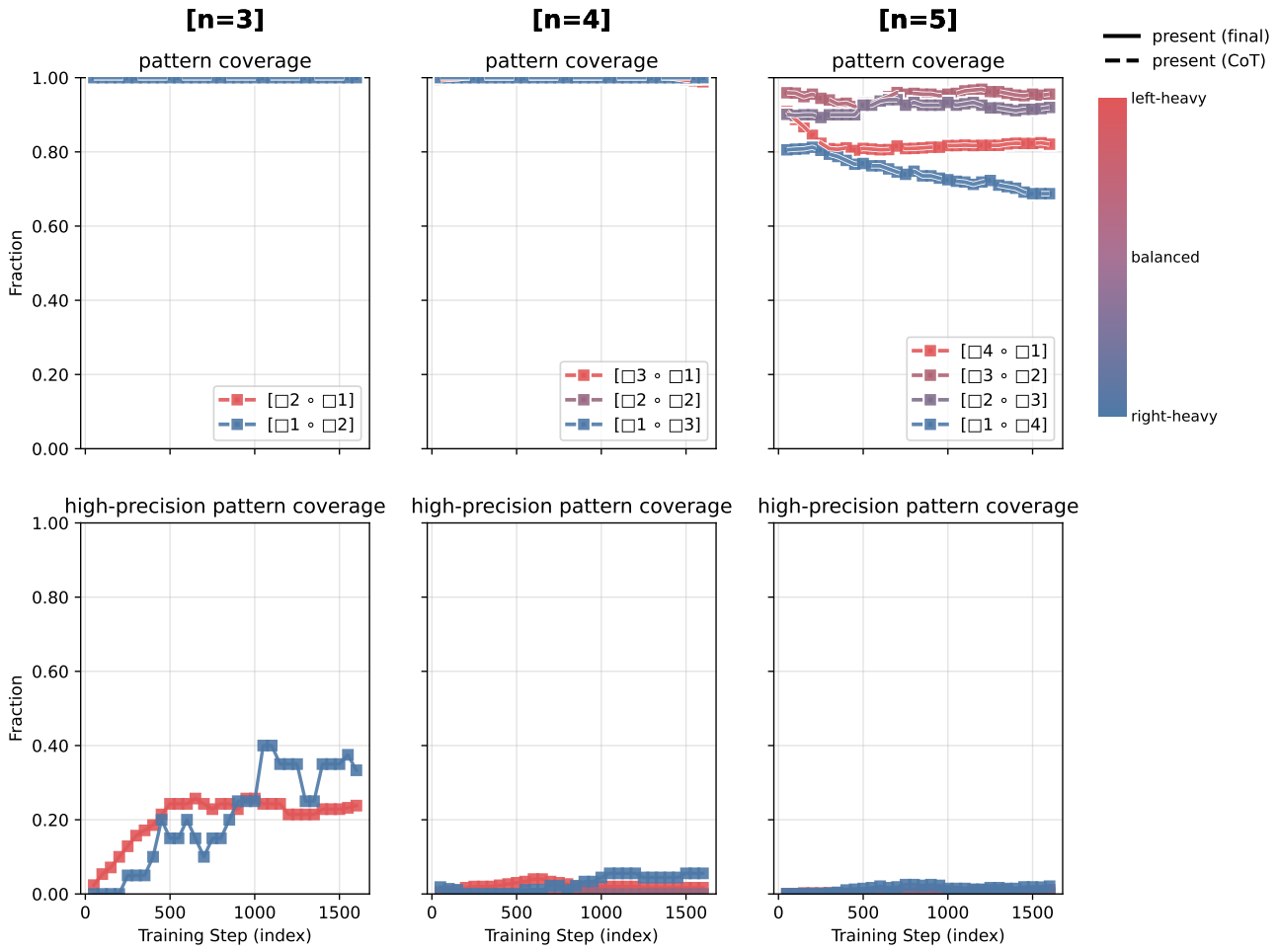

Figure 9. **Learning dynamics at a glance (Qwen2.5-1.5B with SFT)**. The model does not reliably learn any $n = 4$ pattern.

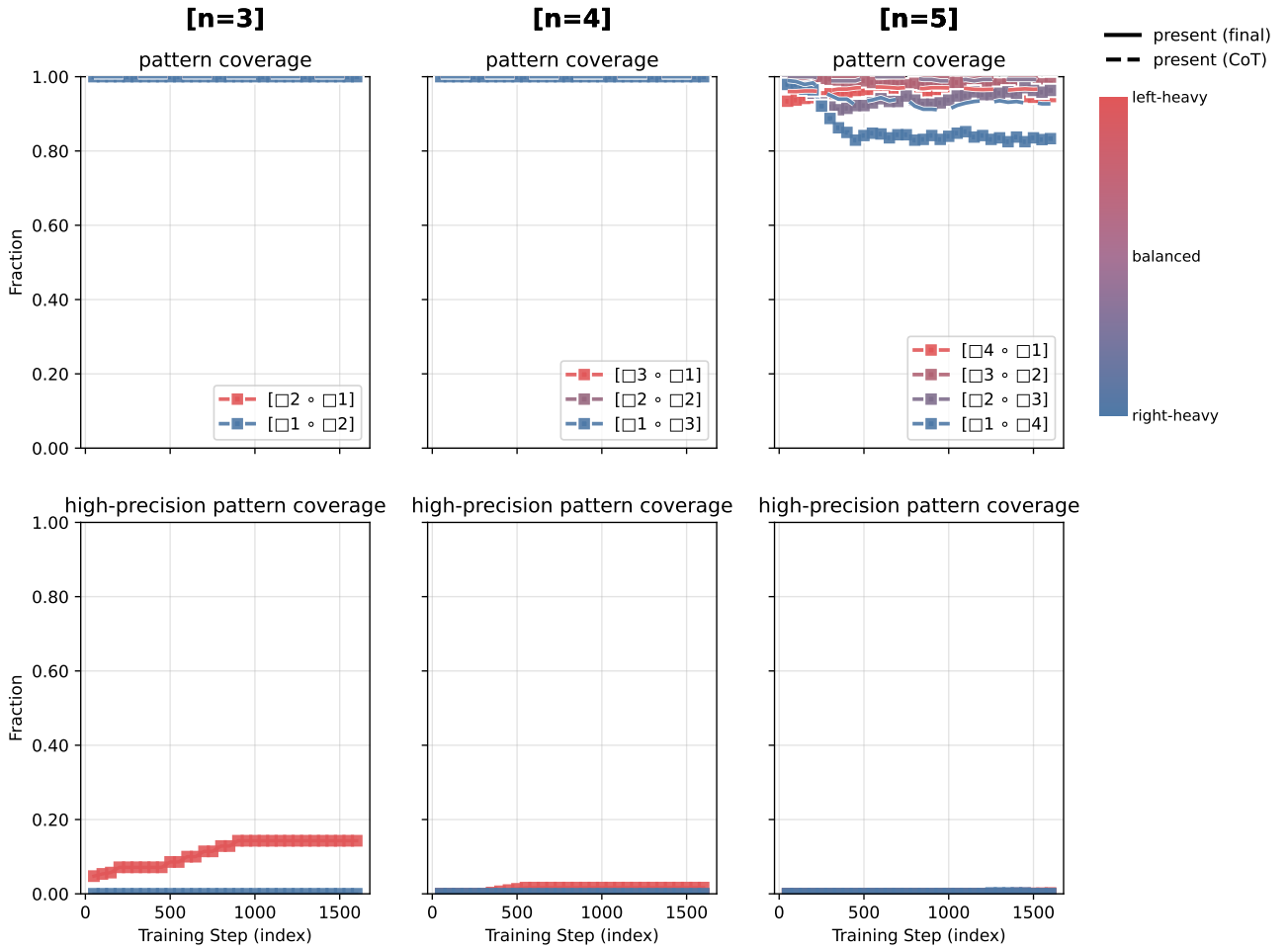

Figure 10. **Learning dynamics at a glance (Qwen2.5-1.5B with SFT+CoT)**. The model does not reliably learn any $n = 4$ pattern.

B.3. Extension to Formal Mathematical Reasoning (More Detailed)

Lean 4 ((de Moura and Ullrich)) is an interactive theorem prover in which proofs are constructed by writing *tactics* that incrementally transform the current proof state until all goals are discharged. Many proofs naturally decompose into subtasks: for an iff or conjunction, the model must prove both directions; for a case analysis, each case is a separate subtask. The `constructor` tactic, for example, opens both directions of an iff as two child goals, and the proof writer then dispatches each one. This hierarchical structure makes Lean a natural setting for the same compositional analysis we apply in COUNTDOWN: each binary-split tactic produces a *left* subtask and a *right* subtask, and the model's effort can be quantified as the number of tactics it spends on each.

B.3.1. CHOICE OF TACTICS

We limit our analysis to proofs that open with a binary-split tactic at the top level of the proof body. Specifically, we include the following tactics:

- `constructor` and `fconstructor`
- `refine ⟨?_, ?_⟩`
- `apply Iff.intro` and `apply And.intro`
- `refine Iff.intro ?_ ?_` and `refine And.intro ?_ ?_`

We deliberately exclude case-analysis tactics such as `induction`, `cases`, and `rcases`, since these often produce a trivial base case and a substantial inductive step, a confounder that would inflate the number of right-heavy patterns.

Example Proof: `Set.preimage_subset_preimage_iff` (mathlib4)

```
theorem preimage_subset_preimage_iff {f : $\beta$ → $\alpha$}
    (hs : s $\subseteq$ range f) :
    f$^{-1}$' s $\subseteq$ f$^{-1}$' t $\leftrightarrow$ s $\subseteq$ t := by
  constructor
  -- forward branch (3 tactics, $w_L$ = 3)
  intro h x
  apply h
  intro h x hx
  -- backward branch (2 tactics, $w_R$ = 2)
  rcases hs hx with ⟨y, rfl⟩
  exact h hx
```

The top-level `constructor` produces two subgoals; subsequent tactics are partitioned into the left and right branches based on which subgoal they discharge. With $w_L = 3$ and $w_R = 2$, this proof has $r = 3/(3+2) = 0.6$ and is classified as *balanced* (see Appendix B.3.2).

B.3.2. Definition of pattern structure

For a proof body that opens with a binary-split tactic, we count the number of top-level tactics used in each of the two immediate subtasks and write them as (w_L, w_R). Subtasks are delimited by the Lean 4 bullet syntax $(\cdot)$ or by `case`-tagged blocks; nested `by`-blocks within a single `have`/`show` are counted as one tactic at the parent level. Comments and blank lines are ignored. We let

$$r = \frac{w_L}{w_L + w_R},$$

and categorize the shape of each proof as

$$\text{shape}(r) = \begin{cases} \text{left-heavy} & \text{if } r \geq 0.6 \\ \text{right-heavy} & \text{if } r \leq 0.4 \\ \text{balanced} & \text{otherwise.} \end{cases}$$

The interval $(0.4, 0.6)$ for the balanced bucket matches the convention used for $n = 5$ Countdown puzzles. In Appendix B.3.5, we verify that the results are invariant to the choice of the threshold values.

B.3.3. Dataset

Distribution of the SFT dataset. We take 323K samples (19%) from the full dataset (1.74M samples) and find that roughly 47K (14% of the sampled data) use binary-split tactics. Among those, 51% use balanced proofs, 14% use left-heavy proofs, and 35% use right-heavy proofs.

Preparing the in-distribution dataset. We sample 100 questions from `Goedel-LM / SFT_dataset_v2` (Lin et al., 2026) (Apache-2.0) while ensuring that the dataset's structural distribution among proofs with binary-split tactics is maintained: 14 left-heavy, 51 balanced, and 35 right-heavy.

Preparing the out-of-distribution dataset. We filter for questions in the LeanDojo 4 Benchmark (Yang et al., 2023) (CC BY 2.0) whose mathlib4 (The mathlib Community, 2020) (Apache-2.0) reference solutions use the binary-split tactics. Next, among questions that whose reference solutions have balanced / left-heavy / right-heavy proofs, we respectively sample 43 / 46 / 46 questions.

B.3.4. EVALUATION

Generation hyperparameters. We run `Goedel-Prover-V2-32B` with $k = 32$ samples per question, temperature 0.6, top-p 0.95, and maximum new tokens 8192. The prompt is wrapped with the model's Qwen3 chat template with the user prompt directly from the SFT data.

Parsing the proof body. The model's raw output typically contains a chain-of-thought paragraph followed by one or more fenced code blocks. We extract the proof body by (i) locating all fenced blocks with an explicit language tag (`lean`, `lean4`, `tactic`, `tactics`), (ii) preferring the final `lean*` block over any `tactics*` scratch block, and (iii) taking the substring after the final `:= by` marker inside the chosen block. Outputs that do not contain a parseable Lean code block are treated as empty proof bodies and are discarded.

B.3.5. CHOICE OF THRESHOLD

We test whether the results in Section 4.5 are sensitive to the choice of threshold values for determining what counts as balanced, left-heavy, or right-heavy. In Table 8, we present the result when re-categorizing the proofs with a different threshold: (1/3, 2/3). The results still hold.

Table 8. **Distribution of proofs by compositional structure (%) under different thresholds (1/3, 2/3).** The results from Table 3 still hold.

	balanced	left-heavy	right-heavy
Present in SFT dataset	58	11	31
Generated by RL model (ID)	65 +7	11	25 -6
Generated by RL model (OOD)	77 +19	14 +3	9 -22

C. Additional Ablations

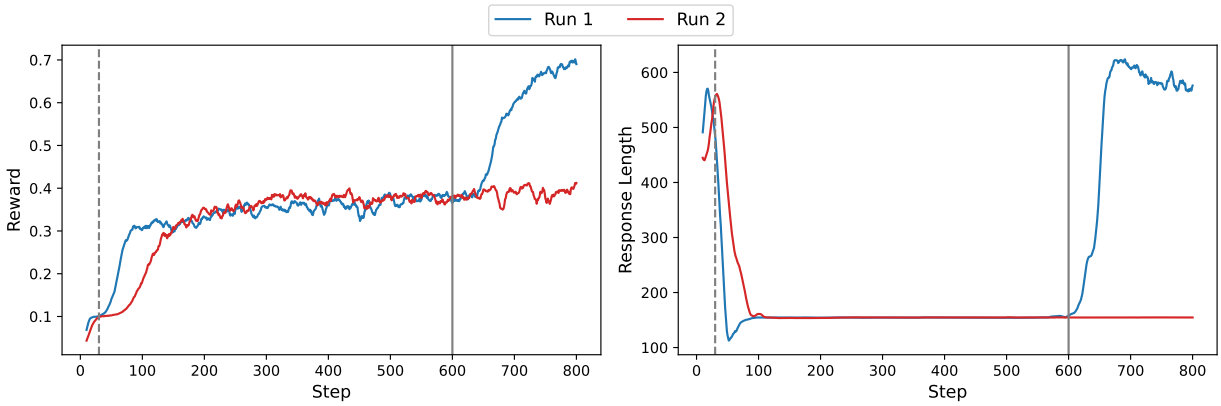

Figure 11. **Qwen2.5-1.5B often learns a formatting-only shortcut: (Left)** training reward; **(Right)** training response length of two distinct training runs. Once the model learns to collect the formatting reward (grey dotted vertical), the model attempts the puzzle without chain of thought, characterized by a sharp decrease in the response length. During some but not all runs (blue), we observe a phase transition (grey solid vertical), where the model starts generating chain of thought again.

C.1. A Formatting-Only Shortcut, and How to Avoid It

Each answer can receive 0.1 points for correct formatting, plus up to 0.9 points for correctness. Smaller models (1.5B and 3B) often exploit the format-only reward first (Figure 11), whereas the larger 7B model does not. This shows that pretraining endows the larger model with more advanced reasoning circuits. For example, if it can collect reward 0.9 with probability > 0.1 from the start, then it has no incentive to overfitting to the formatting reward.

Intervention (large-to-small group size schedule) To stabilize training, we warm-start training with a larger group size to encourage exploration of different patterns [9] and reduce the group size to the default value of $G = 4$ whenever the models make the phase transition and exit the shortcut.

C.2. Effect of Standard Deviation Normalization

For our main experiments, we do not apply std normalization to the group advantage, following Liu et al. (2025). Here, we present the results when we do normalize the advantages by the standard deviation in each group (Shao et al., 2024). All results are from the 1.5B model.

C.2.1. Negative Gradients Induce Longer Responses

When initializing training with the normalization, the response length increases faster and stays higher, compared to when training without normalization (Figure 12). We observe a similar trend with a different model (Appendix D).

To identify the cause of the phenomenon, we separately train with positive gradients only (set the group advantage $A_i = \max(A_i, 0)$ after the standard deviation normalization) and negative gradients only ($A_i = \min(A_i, 0)$). The response length of the positive-only training run follows the trend of training without the normalization, whereas negative-only training mimics the trend of training with the normalization. This shows that the negative gradients are the sole cause behind the initial increase in the response length.

This behavior can be understood as the effect of the negative gradients on the token probabilities. Initially, weak models do not collect the final output reward, and the possible rewards are within $\{0, 0.1\}$. However, when the rewards are normalized by the standard deviation, the absolute value of the advantage becomes at most 20 times larger. Updates using the negative gradients greatly reduces the probability of all tokens generated, and in particular, the generation of the EOS character (which usually occurs after the `<answer></answer>` tags). After the probability of generating EOS token is lowered sufficiently, the model tends to continue generating tokens after the `<answer></answer>` tags.

[9] We find that $G = 16$ for the 1.5B and $G = 8$ for the 3B model are sufficient to induce the phase transition.

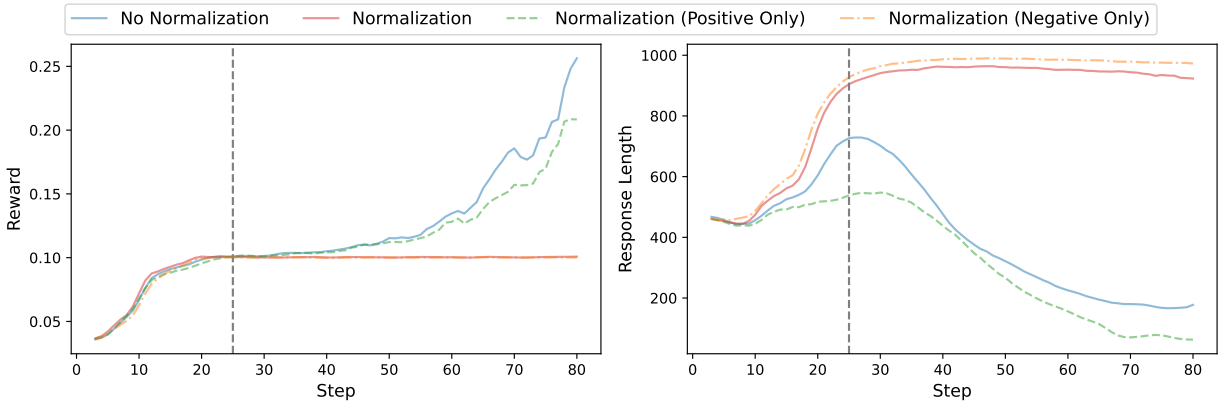

Figure 12. **When normalizing group rewards by standard deviation, negative gradients make responses longer: (Left)** training reward; **(Right)** training response length. In the first 25 steps, training reward increases to 0.1 at the same pace across all runs. However, with standard deviation normalization (red), the response length peaks high, and the training reward stalls. Training with only positive gradients (green dotted) and negative gradients (yellow dotted) respectively mimic the behavior of training without (blue) and with (red) normalization.

C.2.2. Standard Deviation Normalization Leads to Worse Performance

We additionally notice that normalizing by the standard deviation leads to a worse performance at the end of the training (Table 9).

Table 9. **Effect of standard deviation normalization on final performance.** All-correct@32 (All@), Average Accuracy (Avg), and Pass@32 (Pass@) with or without standard deviation normalization during group rewards computation. Normalizing generally hurts final performance.

	No Normalization			Normalization		
	All@	Avg	Pass@	All@	Avg	Pass@
1.5B	0.66	0.85	0.94	0.36	0.60	0.86
3B	0.61	0.84	0.95	0.66	0.86	0.96

C.3. Effect of Number of Rollouts

In Appendix C.1, we increase the number of rollouts only during the initial phase of training to induce the phase transition necessary for learning. Here, we explore maintaining the increased number of rollouts ($G = 16$ for the 1.5B model and $G = 8$ for the 3B model) until the end of the training.

We notice that entropy remains slightly higher than training with a small group size ($G = 4$) and the training reward may start decreasing before 1 epoch (around 0.5 epoch for 1.5B model and 0.75 epoch for 3B model). We stop training whenever we observe a decline in the training reward, and evaluate the final performance. The performance of these models is slightly worse than training switching to a smaller rollout mid-training and significantly worse than fully training with $G = 4$ (Table 10). Other than the initial training stability, we observe no benefit of training with a larger group size.

C.4. Effect of Held-out Patterns

C.4.1. Generalization to Held-Out Patterns (More Details)

Here, we continue the discussion from Section 4.2. When we remove the pattern $A/B+C$, we must also remove all patterns which directly include $A/B+C$ as a subpattern, or patterns that can collapse into $A/B+C$, once some subtasks have been solved. See Table 11 for the full list.

Table 10. **Effect of group size on final performance.** All-correct@32 (All@), Average Accuracy (Avg), and Pass@32 (Pass@) on different choices of group sizes. A larger number of rollouts generally hurts final performance.

	Small			Large			Switch		
	All@	Avg	Pass@	All@	Avg	Pass@	All@	Avg	Pass@
1.5B	0.66	0.85	0.94	0.46	0.78	0.93	0.46	0.79	0.95
3B	0.61	0.84	0.95	0.46	0.78	0.94	0.48	0.83	0.98

Table 11. **The list of patterns removed, when holding out $A/B + C$.** We additionally remove 6 patterns that contain $A/B + C$ as a subpattern and 8 more that can be collapsed into $A/B + C$ once a subtask has been solved.

Pattern	Explanation	Canonical Pattern	Structure
$A/B + C$	Itself	$A/B + C$	$[2] + [1]$
$(A/B + C) + D$	Direct subpattern	$A/B + C + D$	$[3] + [1]$
$(A/B + C) - D$	Direct subpattern	$A/B + C - D$	$[3] - [1]$
$(A/B + C) \times D$	Direct subpattern	$(A/B + C) \times D$	$[3] \times [1]$
$(A/B + C)/D$	Direct subpattern	$(A/B + C)/D$	$[3] \div [1]$
$A - (B/C + D)$	Direct subpattern	$A - B/C - D$	$[1] - [3]$
$A/(B/C + D)$	Direct subpattern	$A/(B/C + D)$	$[1] \div [3]$
$A/B + (C \times D)$	Combine $C \times D$	$A \times B + C/D$	$[2] + [2]$
$A/B + (C/D)$	Combine C/D	$A/B + C/D$	$[2] + [2]$
$A/(B + C) + D$	Combine $B + C$	$A/(B + C) + D$	$[3] + [1]$
$A/(B - C) + D$	Combine $B - C$	$A/(B - C) + D$	$[3] + [1]$
$(A + B)/C + D$	Combine $A + B$	$(A + B)/C + D$	$[3] + [1]$
$(A - B)/C + D$	Combine $A - B$	$(A - B)/C + D$	$[3] + [1]$
$(A \times B)/C + D$	Combine $A \times B$	$A \times B/C + D$	$[3] + [1]$
$(A/B)/C + D$	Combine A/B	$A/B/C + D$	$[3] + [1]$

C.4.2. What Patterns are Necessary for the Model to Learn?

We further explore the effect of heldout patterns during training. We train the `Qwen2.5-1.5B` model with the following heldout strategies:

- Only 3 / Only 4: Train only with patterns from specific puzzle sizes.
- No Hard: Train without patterns of shape $[1] - [3]$ or $[1] \div [3]$ (Section 4.3)
- Random p: Train only with randomly selected $p\%$ of the patterns

Following Appendix C.1, we warm-start the training with a larger group size $G = 16$ and check how many runs are successful out of 3; i.e., model escapes the formatting-only shortcut after something "clicks" (Table 12). When hard patterns are not a part of the training ("Only 3" or "No hard"), the model does not have access to learn from high-entropy, high-reward rollouts and fail to learn. However, training only with $n = 4$ patterns still leads to successful training. Additionally, the diversity of the patterns also seems to be helpful for the model — the probability of a successful run increases with the number of unique patterns in the dataset.

C.4.3. Models Can Generalize to Unseen Patterns (Random 50)

Even when the model is trained with only a randomly selected 50% of the patterns, more than 95% of the patterns are `present` in the final checkpoint, and the model scores near 70% average accuracy. The 3B model is able to generalize to more patterns, whereas 4% of patterns are `Absent` from the reasoning of the 1.5B model and another 4% are always applied incorrectly (Table 13). Additionally, the 1.5B model is able to generalize to $n = 5$ patterns, but the strength of the generalization is weaker than having been trained on the full set of patterns (Table 1).

Table 12. **Number of successful runs (out of 3) per heldout strategy:** The existence of harder patterns and an overall wide range of patterns seems necessary for the 1.5B model to learn.

	Only 3	Only 4	No Hard	Random 30	Random 40	Random 50
# Successful Runs	0	3	0	1	2	3

Table 13. **Final performance of models trained without random 50% of patterns:** All-correct@32 (All@), Average Accuracy (Avg), and Pass@32 (Pass@) of models. Models can generalize to most unseen patterns. 3B model is able to generalize to more diverse patterns than 1.5B, but has a lower All-correct@32.

	n=4					n=5		
	All@	Avg	Pass@	Absent (%)	Incorrect (%)	All@	Avg	Pass@
1.5B	0.47	0.68	0.82	4	4	0.06	0.25	0.52
3B	0.35	0.69	0.88	1	2	-	-	-

C.5. Choice of RL Algorithm

Here, we investigate the effect of the RL algorithm. We replace the GRPO training with PPO (Schulman et al., 2017). Due to higher computational constraints, we only train with 1 seed.

1.5B does not escape the formatting-only shortcut as in Appendix C.1. 3B and 7B improve training accuracy faster with respect to the same number of gradient updates. However, training entropy decreases much faster during PPO training, which translates to a lower Pass@32 but a higher All-correct@32 (Table 14). Compared to GRPO, PPO tends to push the model behavior on each pattern to one extreme — there are simultaneously more 1) `Absent` patterns and 2) patterns with precision of 100%.

Table 14. **Effect of RL algorithm on final performance.** All-correct@32 (All@), Average Accuracy (Avg), and Pass@32 (Pass@) of models. Using PPO instead of GRPO improves All-correct@32 at the cost of Pass@32.

	GRPO			PPO		
	All@	Avg	Pass@	All@	Avg	Pass@
3B	0.61	0.84	0.95	0.65	0.81	0.90
7B	0.62	0.83	0.94	0.76	0.85	0.93

D. Results on a Different Model Family

We additionally conduct some ablations studies on the `Llama-3.2` models (1B and 3B) (Llama3.2 license) and a `Llama-3.1` model (8B) (Llama3.1 license). We take the chat template from the `Instruct` variant but discard most of the system prompt, since the original version is too verbose and is out-of-distribution for the base model. See Appendix E for the exact chat template. We apply the same set of training hyperparmaters.

D.1. Formatting-only Shortcut

Across all model sizes (even the 8B model), `Llama-3.2` and `Llama-3.1` models also find a similar formatting-only shortcut as outlined in Appendix C.1. `Llama-3.1-8B` is able to escape the shortcut without a manual intervention ($G = 4$ rollouts throughout training), but the `Llama-3.2` models are unable to escape the shortcut. By applying the group size switching strategy ($G = 16$ for `Llama-3.2-1B` and $G = 8$ for `Llama-3.2-3B`), the models are able to escape the shortcut on 2 out of 3 runs each. This shows that the mitigation strategy for the shortcut is applicable to another model family, but the strength of its effect may depend on the performance of the base model.

D.2. `Llama` Models Do Not Properly Learn

Even after escaping the formatting-only shortcut, the `Llama` models do not properly learn to use chain-of-thought to improve its answers. Instead, the models generally guess one expression and pad it with repeated answers or gibberish text.

D.3. After SFT on Basic Formatting, RL Shows the Same Trend

We speculate that the `Llama` base models may require a little more training on the correct formatting (including the `<think>` and `<answer>` tags). We SFT the `Llama-3.2-3B` model on examples from out dataset to teach the formatting.[10] RL training on top of this warmed-up checkpoint shows a similar learning dynamics as in the main part of the paper (Figure 14).

D.4. Effect of Standard Deviation Normalization

We train with standard deviation normalization and reproduce its effect on the response length as outlined in Appendix C.2 and report the result in Figure 13.

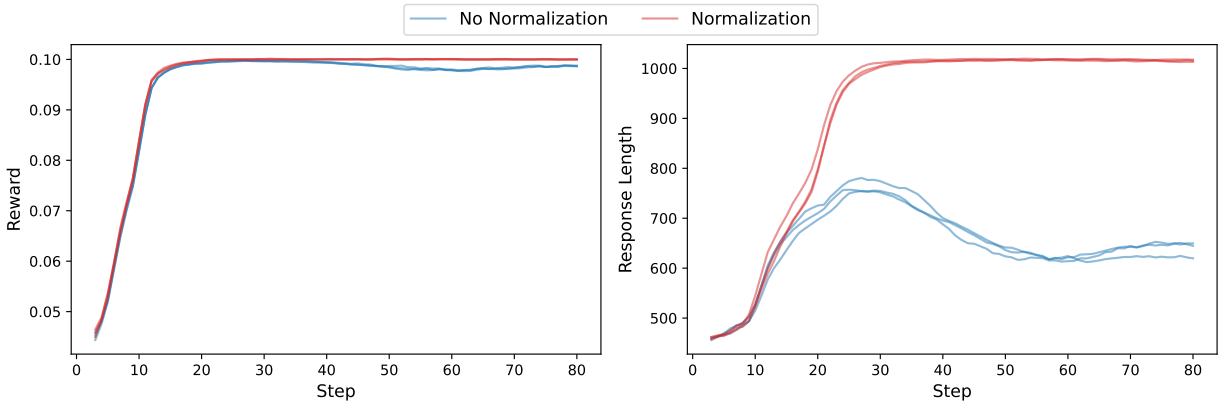

Figure 13. **Effect of standard deviation normalization on response length:** We reproduce the results from Figure 12.

[10]We apply the random CoT template and the hyperparameters in Appendix B.2. We manually stop training after 50 gradient updates (before the learning rate reaches max value) when the training loss decreases below 1.

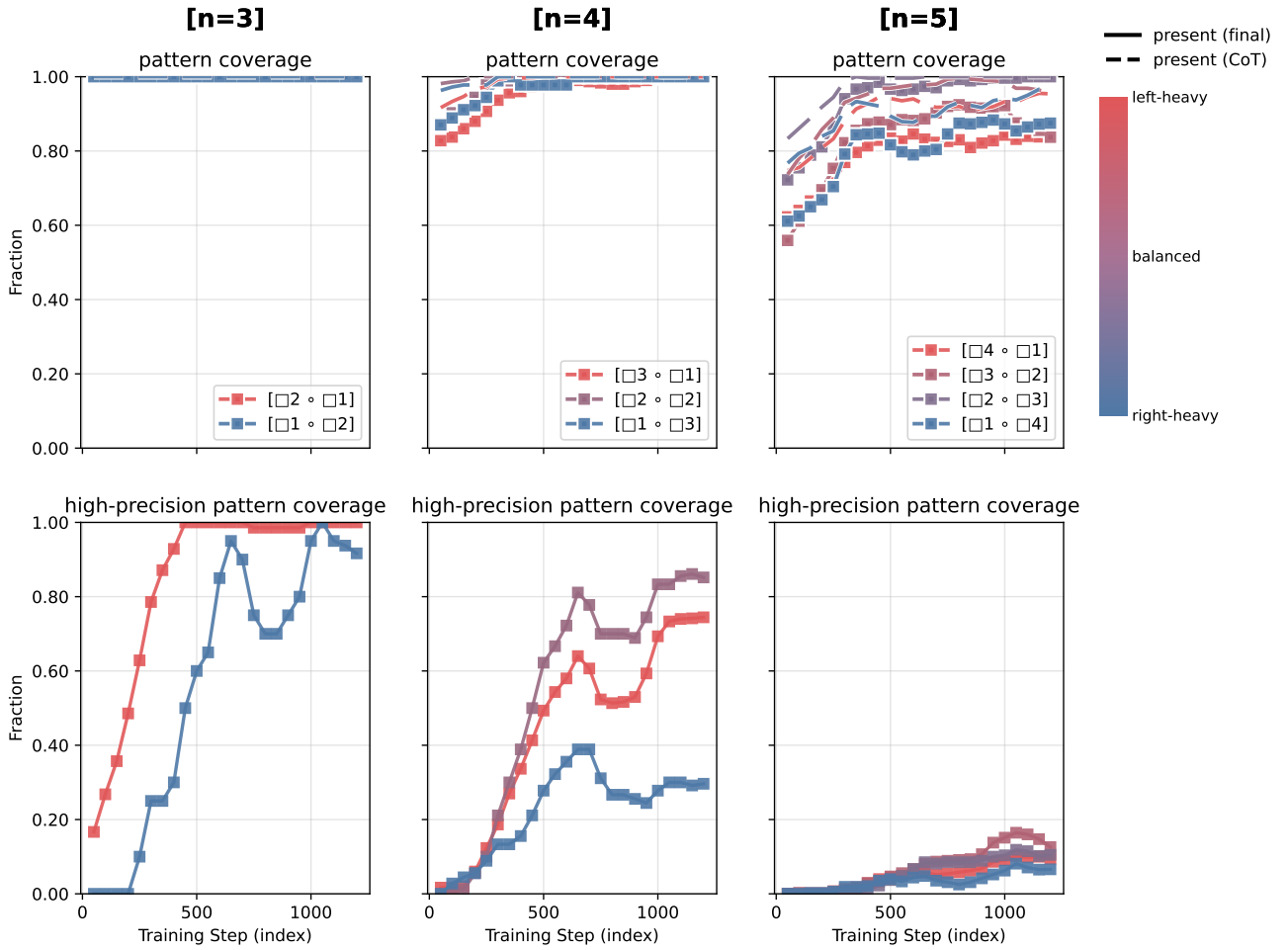

Figure 14. **Learning dynamics at a glance (Llama-3.2-3B after lightweight SFT)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced < left-heavy < right-heavy.

E. Prompts Used For Training

Prompt Template for RL

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.
User: Using the numbers {numbers}, create an equation that equals {target}. You can use basic arithmetic operations (+, -, *, /) and each number can only be used once. Show your work in <think> </think> tags. And return the final answer in <answer> </answer> tags, for example <answer> (1 + 2) / 3 </answer>.
Assistant: Let me solve this step by step.
<think>

Prompt Template for SFT

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.
User: Using the numbers {numbers}, create an equation that equals {target}. You can use basic arithmetic operations (+, -, *, /) and each number can only be used once. Show your work in <think> </think> tags. And return the final answer in <answer> </answer> tags, for example <answer> (1 + 2) / 3 </answer>.
Assistant: Let me solve this step by step.
<think>Using the numbers {numbers}, create an equation that equals {target}.{CoT text}</think>
<answer>{Solution text}</answer>

Example CoT Text for SFT

8 / 8 + 3 / 3 = 1 (Incorrect). 3 + 8 + 8 - 3 = 16 (Incorrect). 3 * 3 / 8 - 8 = -6.87 (Incorrect). 8 / (3 - 8 / 3) = 24 (Correct).

Example Solution Text for SFT

8 / (3 - 8 / 3)

Chat Template for `Llama` Models

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

{User prompt}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

F. Alternate Ways to Canonicalize Patterns

A central claim of this work is that the compositional structure of a problem solution (specifically balanced vs. unbalanced) dictates its difficulty for the model. This analysis, however, relies on our specific method for mapping syntactically diverse expressions to unique canonical operators. A valid concern is whether our findings are an artifact of this mapping. For instance, our normalization rules (Appendix A.1) resolve ambiguities by systematically preferring left-heavy structures (e.g., mapping both $9 \times (3 + 2)$ and $(3 + 2) \times 9$ to the canonical form $(A + B) \times C$). This choice could interact with an intrinsic bias of the model towards left-to-right sequential generation, thereby confounding the analysis of structural difficulty. To isolate the effect of structure from our analytical choices, we conduct two ablations to test the robustness of our conclusions.

Reversed Canonicalization Preference First, we investigate whether our findings hold under an alternative normalization scheme. We invert the rule that resolves ambiguity based on subtree weight, now enforcing a preference for *right-heavy* structures. For example, under this new canonicalization scheme, both $9 \times (3 + 2)$ and $(3 + 2) \times 9$ to the canonical form $A \times (B + C)$. We then redo the analysis. The results, shown in Figure 15, indicate that our central observations are unaffected by this change. We observe the same hierarchy of difficulty as reported in Section 4.3: balanced patterns are mastered most reliably, while the structures that were originally classified as right-heavy (and additionally structures now included in the right-heavy grouping) remain the most challenging. This suggests that the difficulty arises from the procedural requirement to commit to an operator before a complex subroutine is generated, rather than an arbitrary choice of a left- or right-associative canonical representation.

Analysis Without Canonicalization Second, we remove the canonicalization process entirely and analyze the raw expressions generated by the model. In this setting, a single problem may have multiple correct solutions corresponding to different tree shapes (e,g, $9 \times (3 + 2)$ would map to $A \times (B + C)$ and $(3 + 2) \times 9$ would map to $(A + B) \times C$. We group all valid expressions generated across the evaluation set by their unnormalized tree shapes (left-heavy, balanced, or right-heavy), and measure the per-shape pattern coverage and high-precision pattern coverage. As shown in Figure 16, the structural difficulty persists. However, a model's intrinsic bias toward generating solutions in a particular syntactic form can artificially deflate the measured performance of any structural group that includes alternative, algebraically equivalent forms. For example, if a model only ever generates structures like $(A + B) + C$, but not $A + (B + C)$, then the coverage for $[1] \circ [2]$ goes down even though the model can correctly compose two additions.

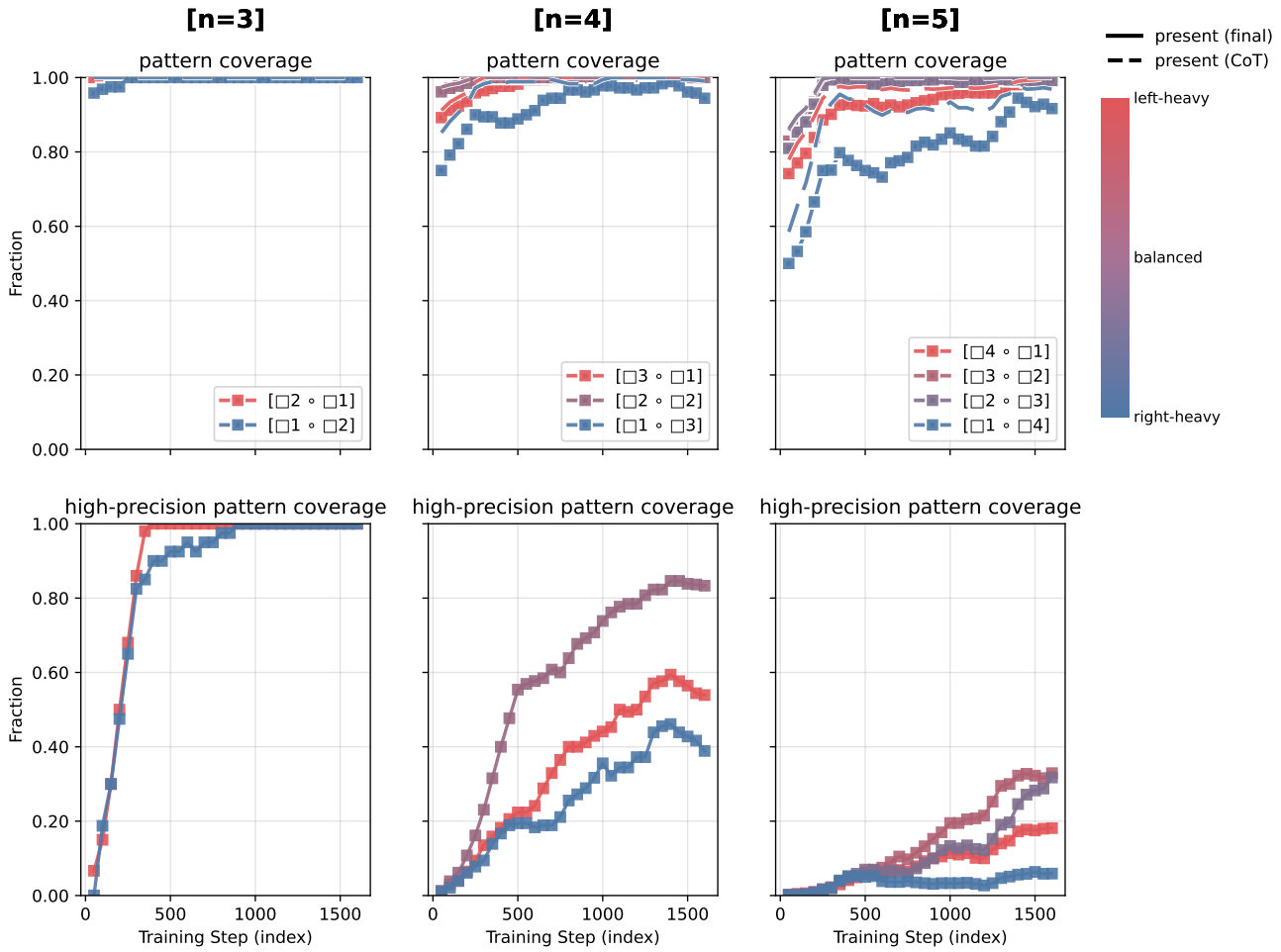

Figure 15. **Learning dynamics at a glance (Qwen2.5-1.5B)** with *reversed* canonicalization. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced < left-heavy < right-heavy.

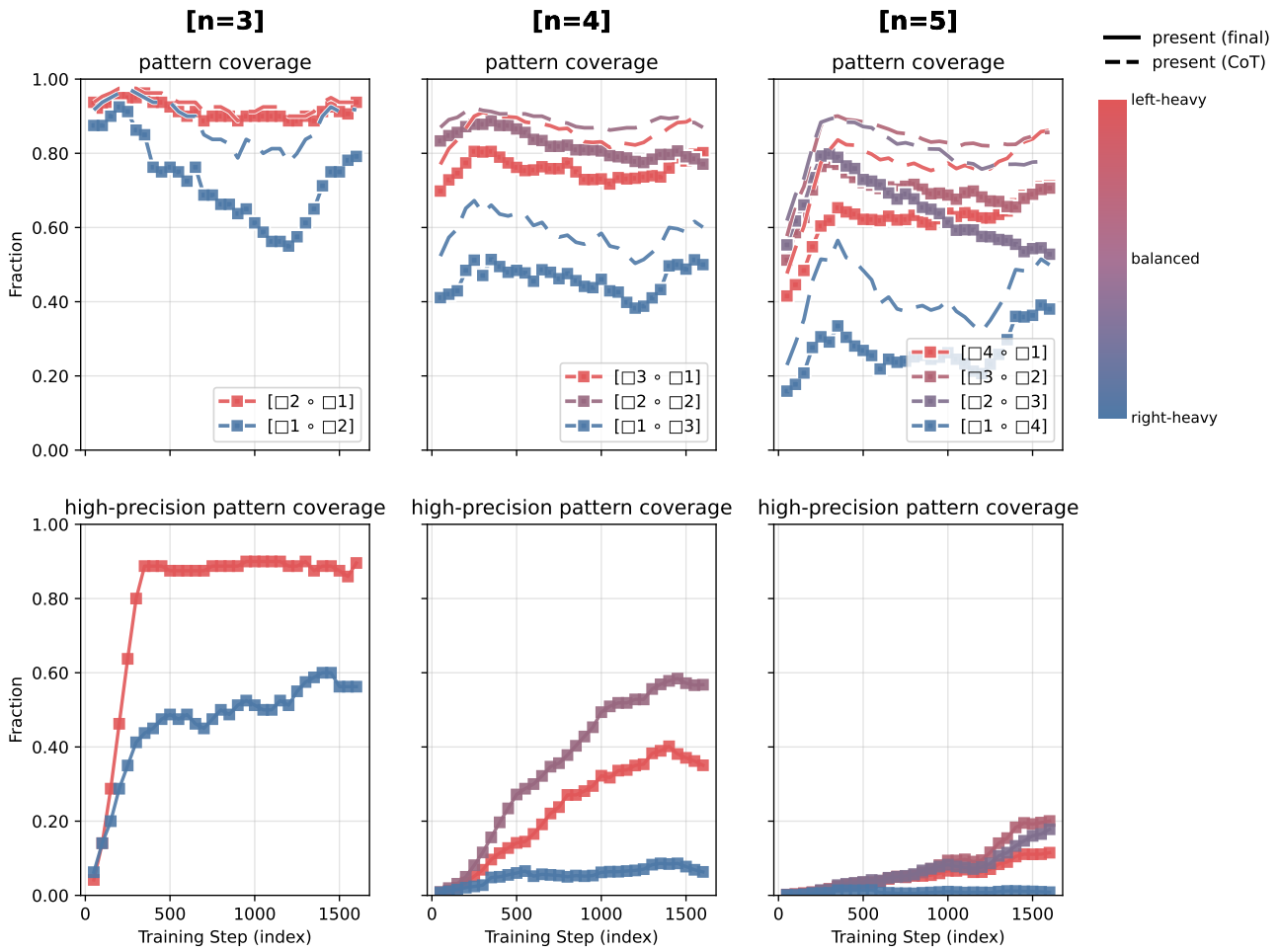

Figure 16. **Learning dynamics at a glance (Qwen2.5-1.5B)** *without* canonicalization. While coverage decreases throughout training (models converge on one of multiple equivalent expressions), the same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy.

G. Additional Plots

In Figures 17 to 22, we present the learning dynamics plots for the `Qwen2.5-3B/7B` models and a different training run for the `Qwen2.5-1.5B` model (equivalent to Figure 4).

In Figure 25, we present the learning dynamics plots for the `Qwen2.5-1.5B` model trained on $n \in \{2, 3, 4\}$ (equivalent to Figures 2 and 4). For each of 4 patterns for $n = 2$, we generated 1000 examples for training and 10 examples for testing in the same way as in Appendix A.2.3.

In Figure 26, we present the learning dynamics plots for the `Qwen2.5-1.5B` model (equivalent to Figure 4) when the high-precision threshold is changed to 70% or 90%.

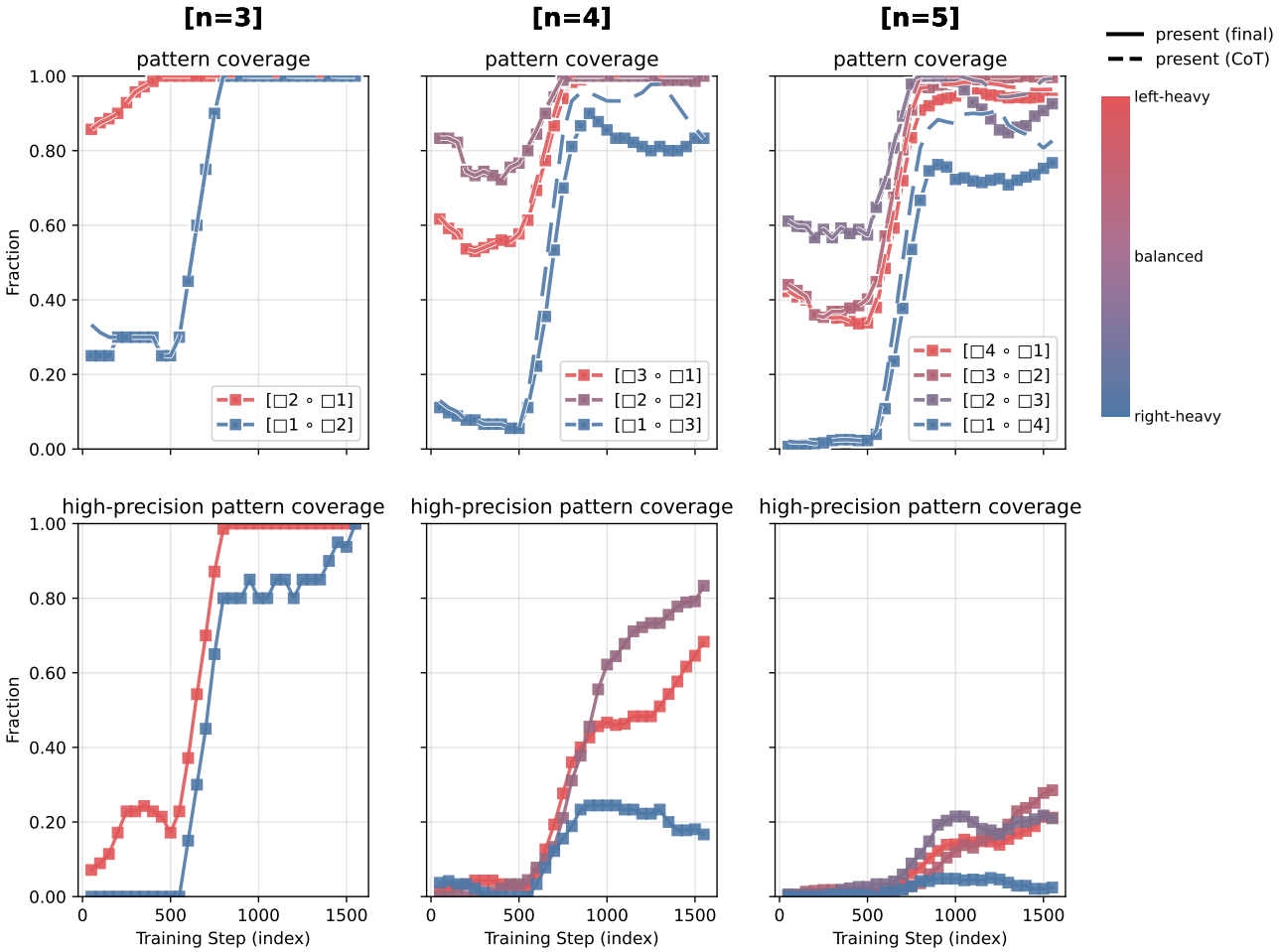

Figure 17. **Learning dynamics at a glance (Qwen2.5-1.5B, another seed)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy.

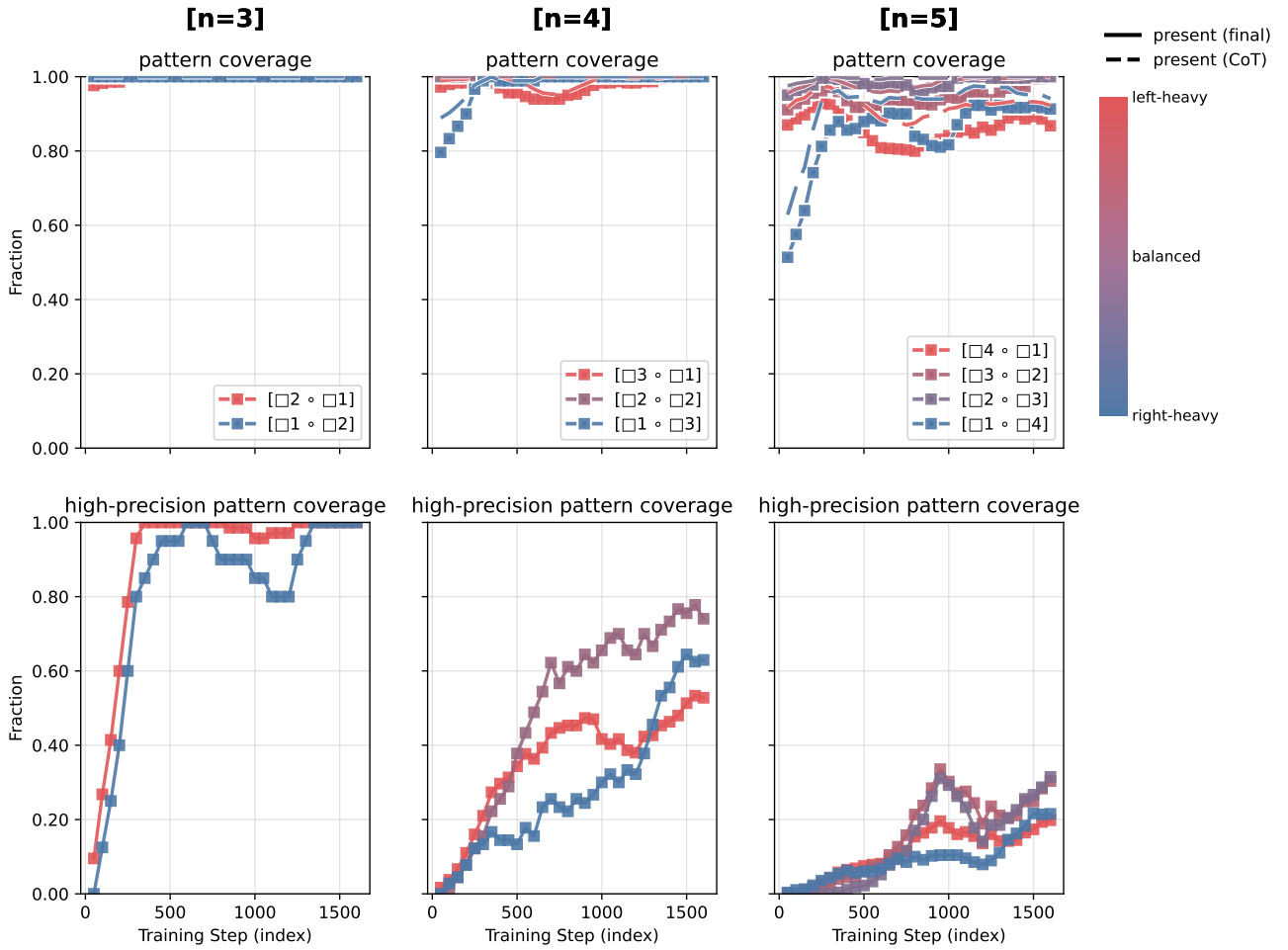

Figure 18. **Learning dynamics at a glance (Qwen2.5-3B)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy.

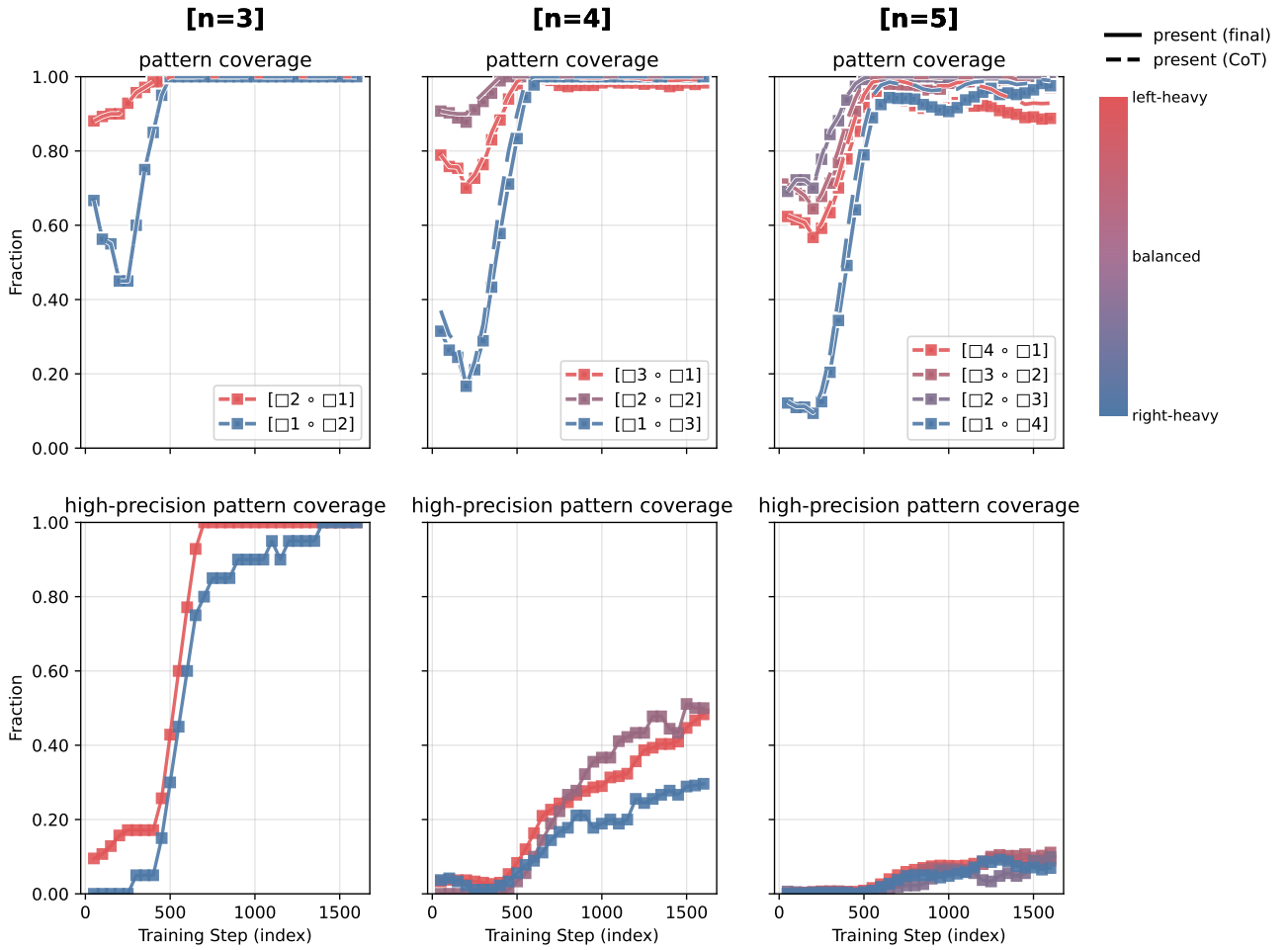

Figure 19. **Learning dynamics at a glance (Qwen2.5-3B, another seed)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy.

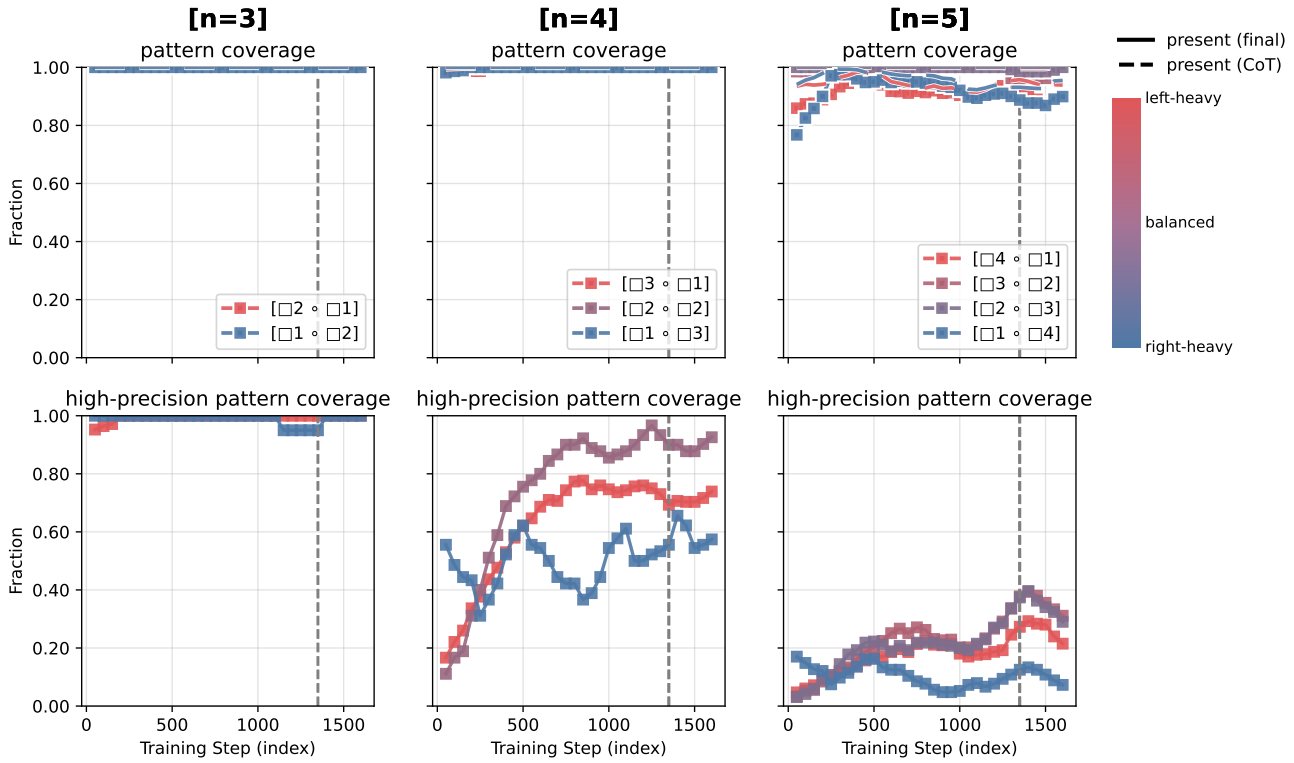

Figure 20. **Learning dynamics at a glance (Qwen2.5-7B)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced < left-heavy < right-heavy. We take the **best** checkpoint at 1350 steps (grey dotted), after which training reward starts declining.

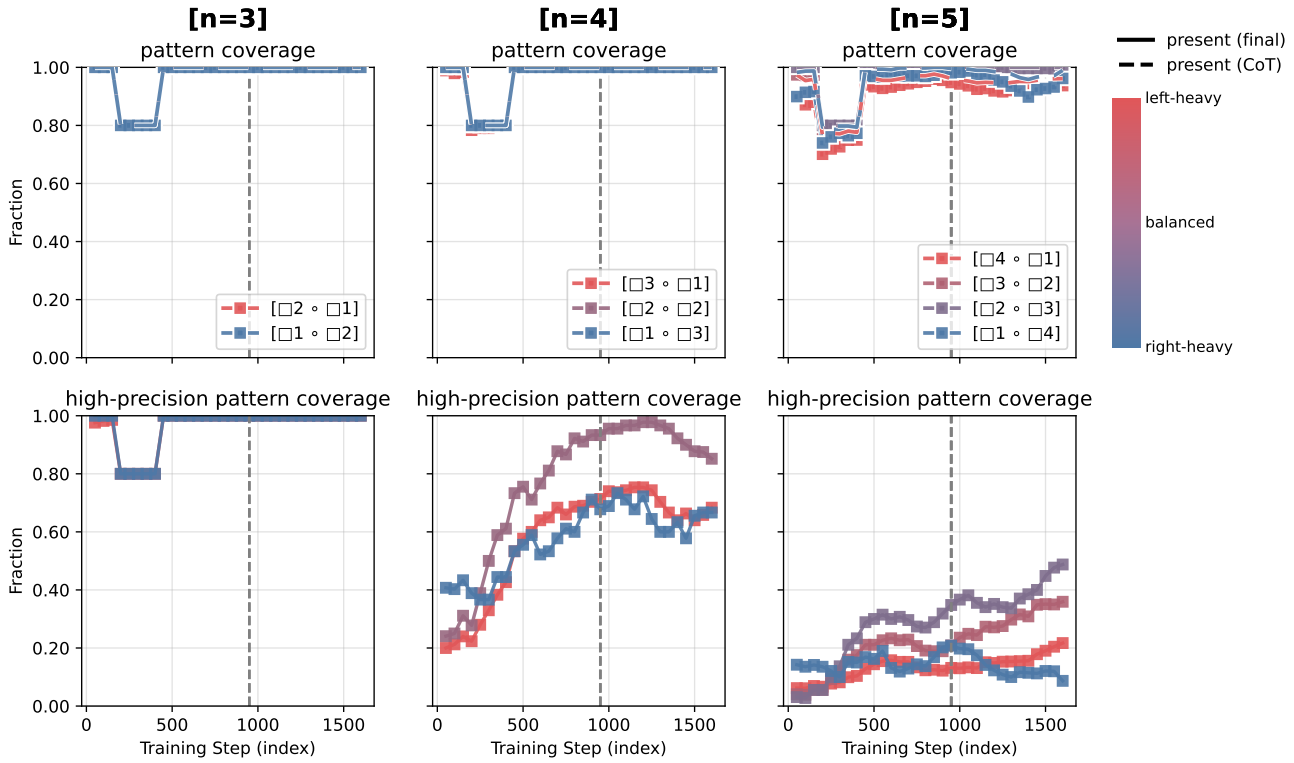

Figure 21. **Learning dynamics at a glance (Qwen2.5-7B, another seed)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy. We take the **best** checkpoint at 950 steps (grey dotted), after which training reward starts declining.

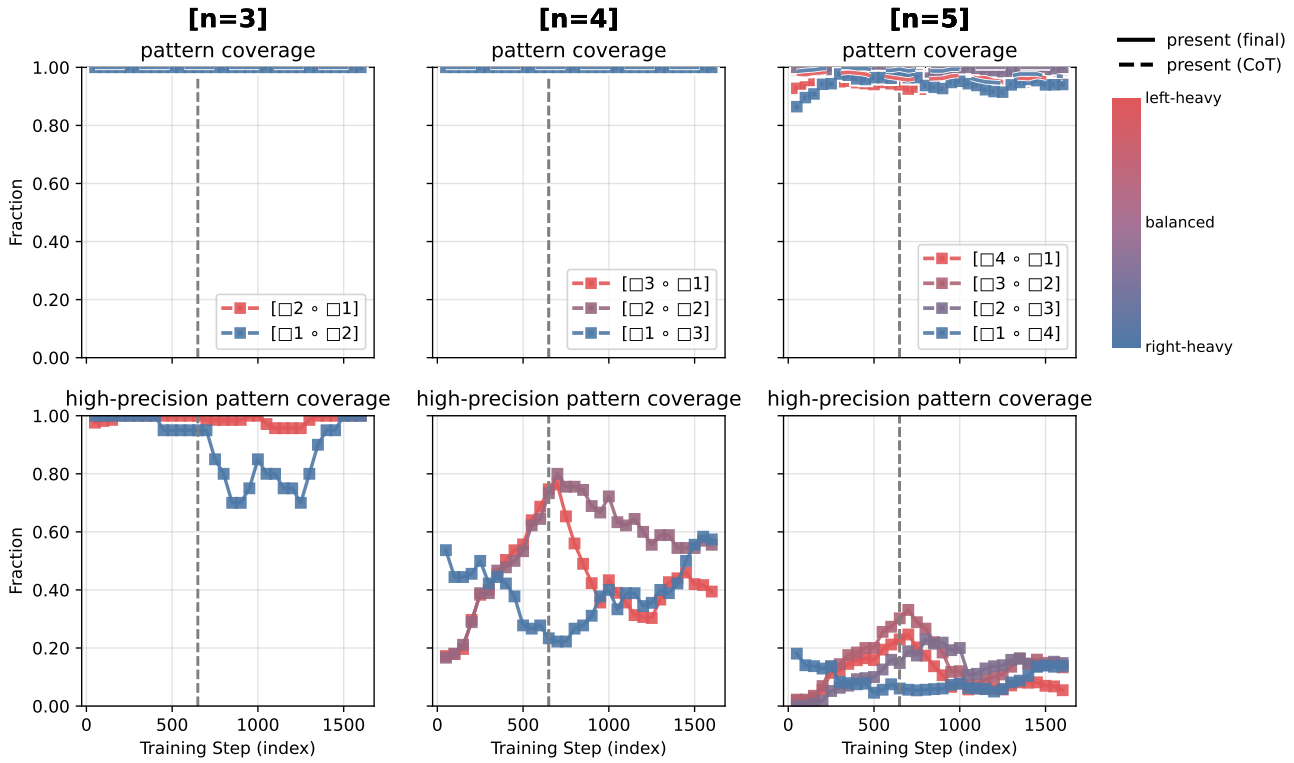

Figure 22. **Learning dynamics at a glance (Qwen2.5-7B, another seed)**. The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy. We take the **best** checkpoint at 650 steps (grey dotted), after which training reward starts declining.

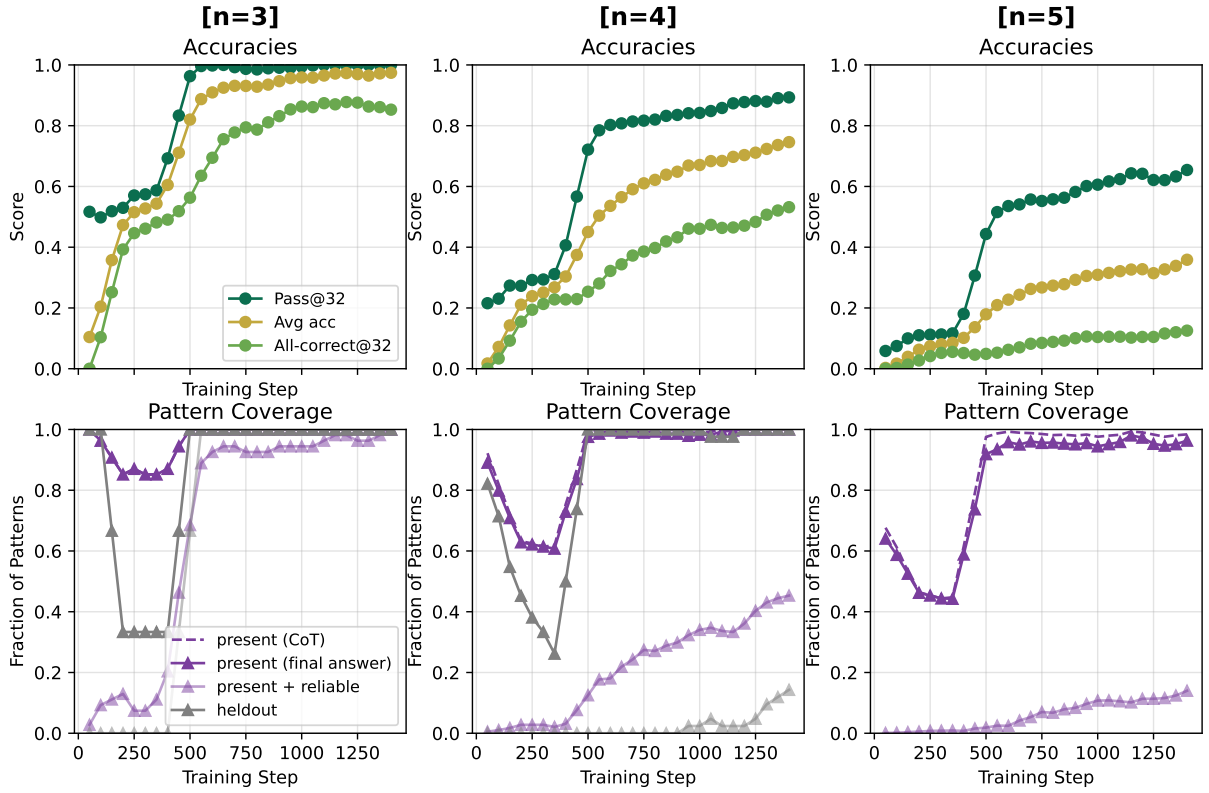

Figure 23. **Generalizes to entire families of held-out compositional patterns (Qwen2.5-1.5B, another seed).** The same conclusion from Figure 3: the model can reuse learned substructures to assemble novel operator-tree shapes it has never been explicitly trained on.

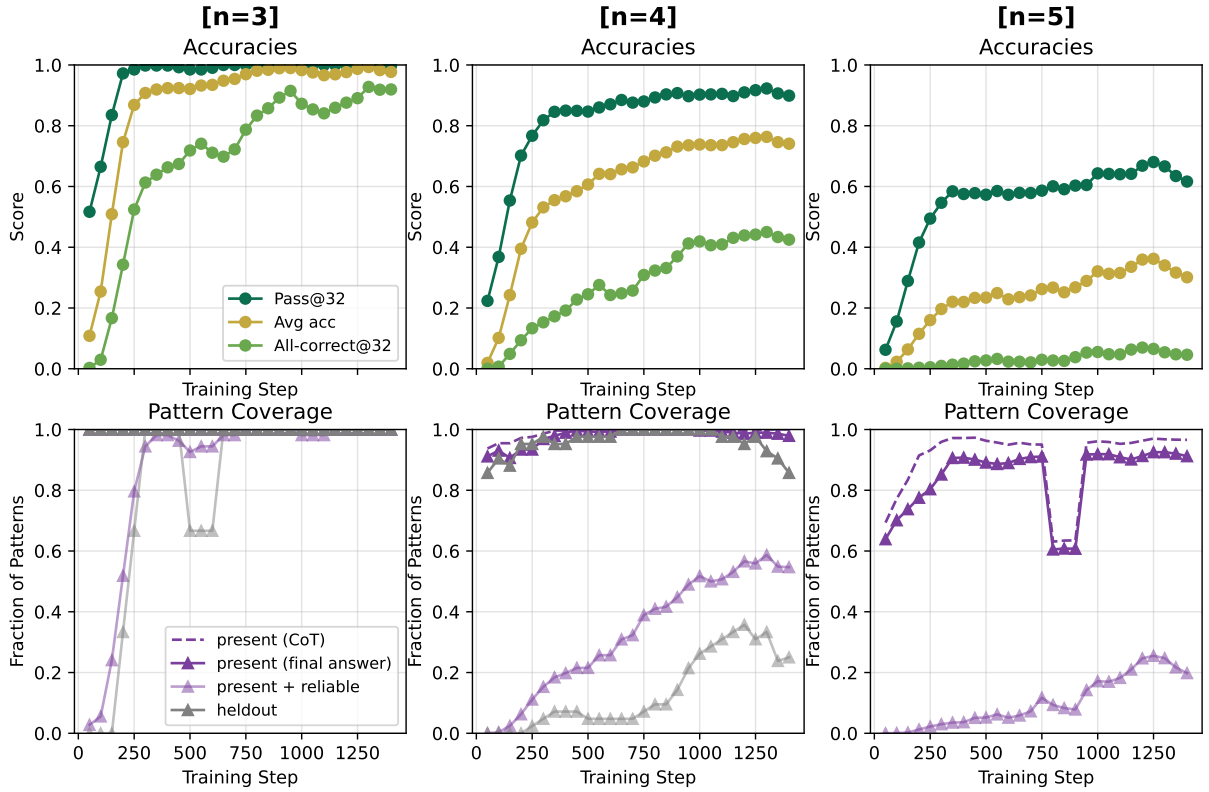

Figure 24. **Generalizes to entire families of held-out compositional patterns (Qwen2.5-1.5B, another seed).** The same conclusion from Figure 3: the model can reuse learned substructures to assemble novel operator-tree shapes it has never been explicitly trained on.

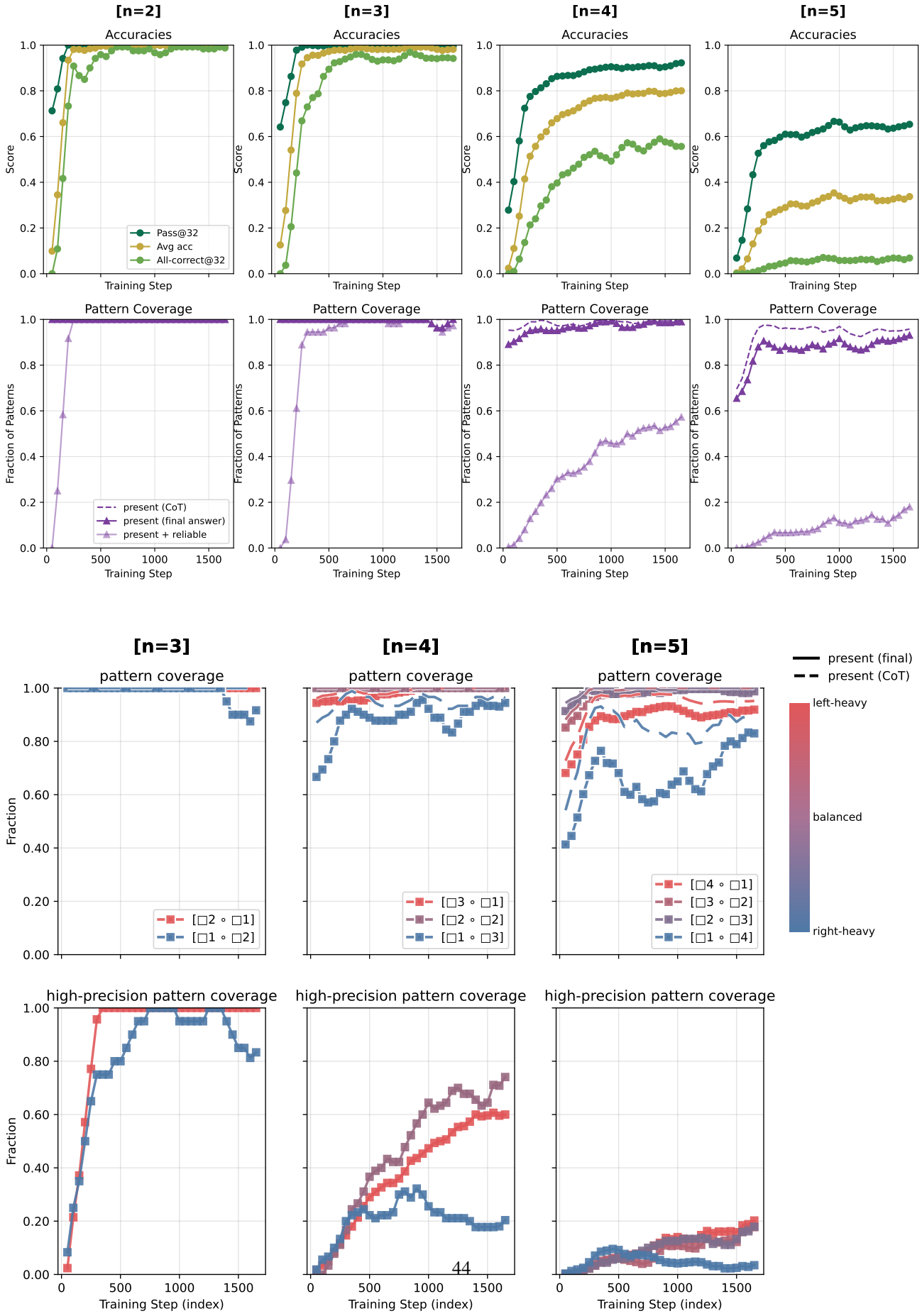

Figure 25. **Learning dynamics at a glance (Qwen2.5-1.5B)** trained on $n \in \{2, 3, 4\}$. The same conclusions from Figures 2 and 4 hold: models exhibit length generalization and master patterns progressively following a hierarchy of difficulty by subtree structure: balanced <

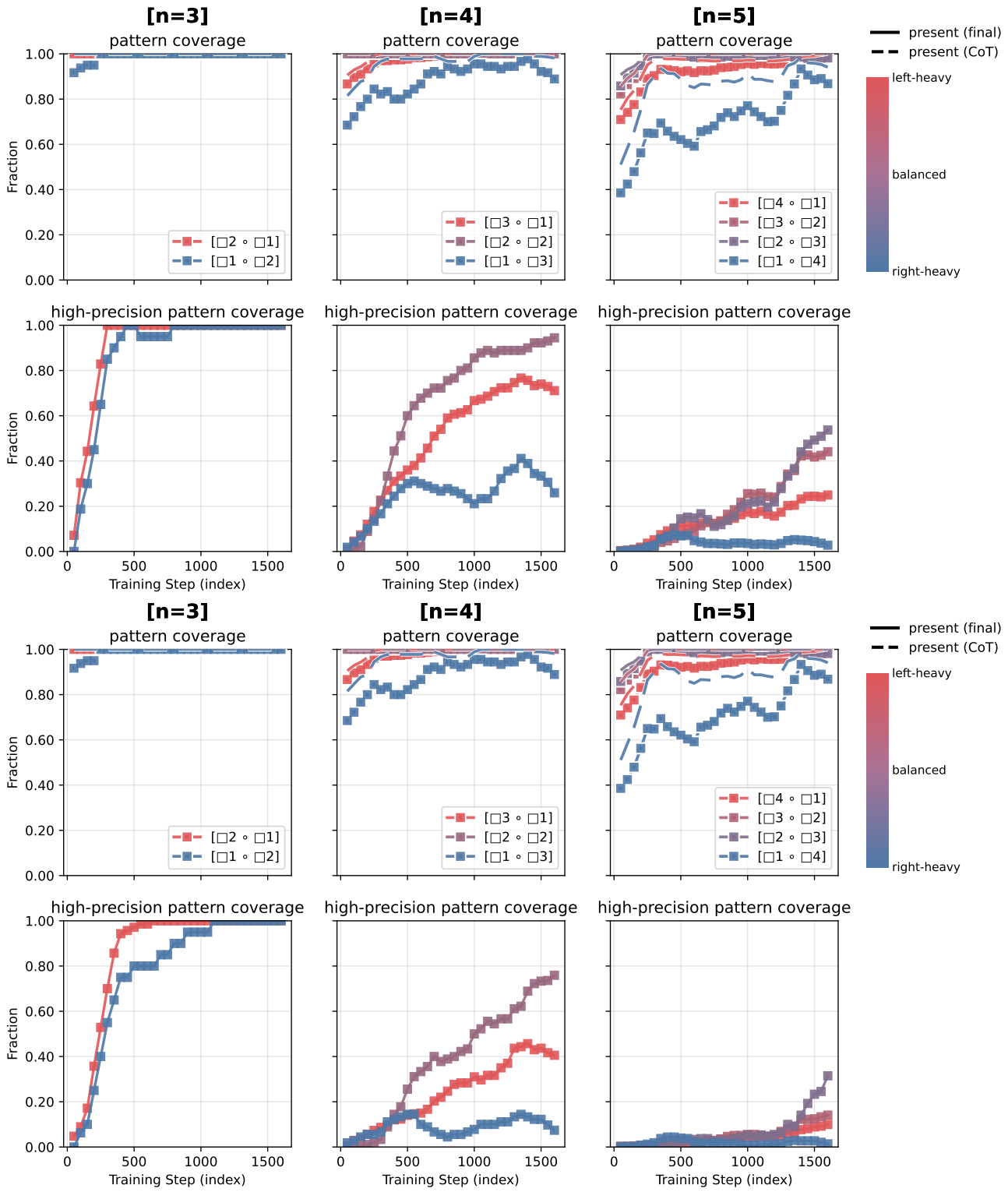

Figure 26. **Learning dynamics at a glance (Qwen2.5-1.5B)** but high-precision defined as precision $\geq 70\%$ (top); $\geq 90\%$ (bottom). The same conclusion from Figure 4 holds about the hierarchy of difficulty by subtree structure: balanced $<$ left-heavy $<$ right-heavy.

H. List of Patterns

Table 15. Expressions and their Canonical Forms (Patterns) (Page 1 of 4)

Expression	Pattern	Expression	Pattern
$(A+B)+C$	$A+B+C$	$A+(B+C)$	$A+B+C$
$(A+B)-C$	$A+B-C$	$A+(B-C)$	$A+B-C$
$(A+B)\times C$	$(A+B)\times C$	$A+(B\times C)$	$A\times B+C$
$(A+B)/C$	$(A+B)/C$	$A+(B/C)$	$A/B+C$
$(A-B)+C$	$A+B-C$	$A-(B+C)$	$A-B-C$
$(A-B)-C$	$A-B-C$	$A-(B-C)$	$A+B-C$
$(A-B)\times C$	$(A-B)\times C$	$A-(B\times C)$	$A-B\times C$
$(A-B)/C$	$(A-B)/C$	$A-(B/C)$	$A-B/C$
$(A\times B)+C$	$A\times B+C$	$A\times(B+C)$	$(A+B)\times C$
$(A\times B)-C$	$A\times B-C$	$A\times(B-C)$	$(A-B)\times C$
$(A\times B)\times C$	$A\times B\times C$	$A\times(B\times C)$	$A\times B\times C$
$(A\times B)/C$	$A\times B/C$	$A\times(B/C)$	$A\times B/C$
$(A/B)+C$	$A/B+C$	$A/(B+C)$	$A/(B+C)$
$(A/B)-C$	$A/B-C$	$A/(B-C)$	$A/(B-C)$
$(A/B)\times C$	$A\times B/C$	$A/(B\times C)$	$A/B/C$
$(A/B)/C$	$A/B/C$	$A/(B/C)$	$A\times B/C$
$A+(B+(C+D))$	$A+B+C+D$	$A+(B+(C-D))$	$A+B+C-D$
$A+(B+(C\times D))$	$A\times B+C+D$	$A+(B+(C/D))$	$A/B+C+D$
$A+(B-(C+D))$	$A+B-C-D$	$A+(B-(C-D))$	$A+B+C-D$
$A+(B-(C\times D))$	$A+B-C\times D$	$A+(B-(C/D))$	$A+B-C/D$
$A+(B\times(C+D))$	$(A+B)\times C+D$	$A+(B\times(C-D))$	$(A-B)\times C+D$
$A+(B\times(C\times D))$	$A\times B\times C+D$	$A+(B\times(C/D))$	$A\times B/C+D$
$A+(B/(C+D))$	$A/(B+C)+D$	$A+(B/(C-D))$	$A/(B-C)+D$
$A+(B/(C\times D))$	$A/B/C+D$	$A+(B/(C/D))$	$A\times B/C+D$
$A-(B+(C+D))$	$A-B-C-D$	$A-(B+(C-D))$	$A+B-C-D$
$A-(B+(C\times D))$	$A-B\times C-D$	$A-(B+(C/D))$	$A-B/C-D$
$A-(B-(C+D))$	$A+B+C-D$	$A-(B-(C-D))$	$A+B-C-D$
$A-(B-(C\times D))$	$A\times B+C-D$	$A-(B-(C/D))$	$A/B+C-D$
$A-(B\times(C+D))$	$A-(B+C)\times D$	$A-(B\times(C-D))$	$A-(B-C)\times D$
$A-(B\times(C\times D))$	$A-B\times C\times D$	$A-(B\times(C/D))$	$A-B\times C/D$
$A-(B/(C+D))$	$A-B/(C+D)$	$A-(B/(C-D))$	$A-B/(C-D)$
$A-(B/(C\times D))$	$A-B/C/D$	$A-(B/(C/D))$	$A-B\times C/D$
$A\times(B+(C+D))$	$(A+B+C)\times D$	$A\times(B+(C-D))$	$(A+B-C)\times D$
$A\times(B+(C\times D))$	$(A\times B+C)\times D$	$A\times(B+(C/D))$	$(A/B+C)\times D$
$A\times(B-(C+D))$	$(A-B-C)\times D$	$A\times(B-(C-D))$	$(A+B-C)\times D$
$A\times(B-(C\times D))$	$(A-B\times C)\times D$	$A\times(B-(C/D))$	$(A-B/C)\times D$
$A\times(B\times(C+D))$	$(A+B)\times C\times D$	$A\times(B\times(C-D))$	$(A-B)\times C\times D$
$A\times(B\times(C\times D))$	$A\times B\times C\times D$	$A\times(B\times(C/D))$	$A\times B\times C/D$
$A\times(B/(C+D))$	$A\times B/(C+D)$	$A\times(B/(C-D))$	$A\times B/(C-D)$
$A\times(B/(C\times D))$	$A\times B/C/D$	$A\times(B/(C/D))$	$A\times B\times C/D$
$A/(B+(C+D))$	$A/(B+C+D)$	$A/(B+(C-D))$	$A/(B+C-D)$
$A/(B+(C\times D))$	$A/(B\times C+D)$	$A/(B+(C/D))$	$A/(B/C+D)$
$A/(B-(C+D))$	$A/(B-C-D)$	$A/(B-(C-D))$	$A/(B+C-D)$
$A/(B-(C\times D))$	$A/(B-C\times D)$	$A/(B-(C/D))$	$A/(B-C/D)$
$A/(B\times(C+D))$	$A/(B+C)/D$	$A/(B\times(C-D))$	$A/(B-C)/D$
$A/(B\times(C\times D))$	$A/B/C/D$	$A/(B\times(C/D))$	$A\times B/C/D$
$A/(B/(C+D))$	$(A+B)\times C/D$	$A/(B/(C-D))$	$(A-B)\times C/D$
$A/(B/(C\times D))$	$A\times B\times C/D$	$A/(B/(C/D))$	$A\times B/C/D$
$A+((B+C)+D)$	$A+B+C+D$	$A+((B-C)+D)$	$A+B+C-D$
$A+((B\times C)+D)$	$A\times B+C+D$	$A+((B/C)+D)$	$A/B+C+D$

Table 16. Expressions and their Canonical Forms (Patterns) (Page 2 of 4)

Expression	Pattern	Expression	Pattern
$A+((B+C)-D)$	$A+B+C-D$	$A+((B-C)-D)$	$A+B-C-D$
$A+((B\times C)-D)$	$A\times B+C-D$	$A+((B/C)-D)$	$A/B+C-D$
$A+((B+C)\times D)$	$(A+B)\times C+D$	$A+((B-C)\times D)$	$(A-B)\times C+D$
$A+((B\times C)\times D)$	$A\times B\times C+D$	$A+((B/C)\times D)$	$A\times B/C+D$
$A+((B+C)/D)$	$(A+B)/C+D$	$A+((B-C)/D)$	$(A-B)/C+D$
$A+((B\times C)/D)$	$A\times B/C+D$	$A+((B/C)/D)$	$A/B/C+D$
$A-((B+C)+D)$	$A-B-C-D$	$A-((B-C)+D)$	$A+B-C-D$
$A-((B\times C)+D)$	$A-B\times C-D$	$A-((B/C)+D)$	$A-B/C-D$
$A-((B+C)-D)$	$A+B-C-D$	$A-((B-C)-D)$	$A+B+C-D$
$A-((B\times C)-D)$	$A+B-C\times D$	$A-((B/C)-D)$	$A+B-C/D$
$A-((B+C)\times D)$	$A-(B+C)\times D$	$A-((B-C)\times D)$	$A-(B-C)\times D$
$A-((B\times C)\times D)$	$A-B\times C\times D$	$A-((B/C)\times D)$	$A-B\times C/D$
$A-((B+C)/D)$	$A-(B+C)/D$	$A-((B-C)/D)$	$A-(B-C)/D$
$A-((B\times C)/D)$	$A-B\times C/D$	$A-((B/C)/D)$	$A-B/C/D$
$A\times((B+C)+D)$	$(A+B+C)\times D$	$A\times((B-C)+D)$	$(A+B-C)\times D$
$A\times((B\times C)+D)$	$(A\times B+C)\times D$	$A\times((B/C)+D)$	$(A/B+C)\times D$
$A\times((B+C)-D)$	$(A+B-C)\times D$	$A\times((B-C)-D)$	$(A-B-C)\times D$
$A\times((B\times C)-D)$	$(A\times B-C)\times D$	$A\times((B/C)-D)$	$(A/B-C)\times D$
$A\times((B+C)\times D)$	$(A+B)\times C\times D$	$A\times((B-C)\times D)$	$(A-B)\times C\times D$
$A\times((B\times C)\times D)$	$A\times B\times C\times D$	$A\times((B/C)\times D)$	$A\times B\times C/D$
$A\times((B+C)/D)$	$(A+B)\times C/D$	$A\times((B-C)/D)$	$(A-B)\times C/D$
$A\times((B\times C)/D)$	$A\times B\times C/D$	$A\times((B/C)/D)$	$A\times B/C/D$
$A/((B+C)+D)$	$A/(B+C+D)$	$A/((B-C)+D)$	$A/(B+C-D)$
$A/((B\times C)+D)$	$A/(B\times C+D)$	$A/((B/C)+D)$	$A/(B/C+D)$
$A/((B+C)-D)$	$A/(B+C-D)$	$A/((B-C)-D)$	$A/(B-C-D)$
$A/((B\times C)-D)$	$A/(B\times C-D)$	$A/((B/C)-D)$	$A/(B/C-D)$
$A/((B+C)\times D)$	$A/(B+C)/D$	$A/((B-C)\times D)$	$A/(B-C)/D$
$A/((B\times C)\times D)$	$A/B/C/D$	$A/((B/C)\times D)$	$A\times B/C/D$
$A/((B+C)/D)$	$A\times B/(C+D)$	$A/((B-C)/D)$	$A\times B/(C-D)$
$A/((B\times C)/D)$	$A\times B/C/D$	$A/((B/C)/D)$	$A\times B\times C/D$
$(A+B)+(C+D)$	$A+B+C+D$	$(A+B)+(C-D)$	$A+B+C-D$
$(A+B)+(C\times D)$	$A\times B+C+D$	$(A+B)+(C/D)$	$A/B+C+D$
$(A-B)+(C+D)$	$A+B+C-D$	$(A-B)+(C-D)$	$A+B-C-D$
$(A-B)+(C\times D)$	$A\times B+C-D$	$(A-B)+(C/D)$	$A/B+C-D$
$(A\times B)+(C+D)$	$A\times B+C+D$	$(A\times B)+(C-D)$	$A\times B+C-D$
$(A\times B)+(C\times D)$	$A\times B+C\times D$	$(A\times B)+(C/D)$	$A\times B+C/D$
$(A/B)+(C+D)$	$A/B+C+D$	$(A/B)+(C-D)$	$A/B+C-D$
$(A/B)+(C\times D)$	$A\times B+C/D$	$(A/B)+(C/D)$	$A/B+C/D$
$(A+B)-(C+D)$	$A+B-C-D$	$(A+B)-(C-D)$	$A+B+C-D$
$(A+B)-(C\times D)$	$A+B-C\times D$	$(A+B)-(C/D)$	$A+B-C/D$
$(A-B)-(C+D)$	$A-B-C-D$	$(A-B)-(C-D)$	$A+B-C-D$
$(A-B)-(C\times D)$	$A-B\times C-D$	$(A-B)-(C/D)$	$A-B/C-D$
$(A\times B)-(C+D)$	$A\times B-C-D$	$(A\times B)-(C-D)$	$A\times B+C-D$
$(A\times B)-(C\times D)$	$A\times B-C\times D$	$(A\times B)-(C/D)$	$A\times B-C/D$
$(A/B)-(C+D)$	$A/B-C-D$	$(A/B)-(C-D)$	$A/B+C-D$
$(A/B)-(C\times D)$	$A/B-C\times D$	$(A/B)-(C/D)$	$A/B-C/D$
$(A+B)\times(C+D)$	$(A+B)\times(C+D)$	$(A+B)\times(C-D)$	$(A+B)\times(C-D)$
$(A+B)\times(C\times D)$	$(A+B)\times C\times D$	$(A+B)\times(C/D)$	$(A+B)\times C/D$
$(A-B)\times(C+D)$	$(A+B)\times(C-D)$	$(A-B)\times(C-D)$	$(A-B)\times(C-D)$
$(A-B)\times(C\times D)$	$(A-B)\times C\times D$	$(A-B)\times(C/D)$	$(A-B)\times C/D$

Table 17. Expressions and their Canonical Forms (Patterns) (Page 3 of 4)

Expression	Pattern	Expression	Pattern
$(A \times B) \times (C + D)$	$(A + B) \times C \times D$	$(A \times B) \times (C - D)$	$(A - B) \times C \times D$
$(A \times B) \times (C \times D)$	$A \times B \times C \times D$	$(A \times B) \times (C/D)$	$A \times B \times C/D$
$(A/B) \times (C + D)$	$(A + B) \times C/D$	$(A/B) \times (C - D)$	$(A - B) \times C/D$
$(A/B) \times (C \times D)$	$A \times B \times C/D$	$(A/B) \times (C/D)$	$A \times B/C/D$
$(A + B)/(C + D)$	$(A + B)/(C + D)$	$(A + B)/(C - D)$	$(A + B)/(C - D)$
$(A + B)/(C \times D)$	$(A + B)/C/D$	$(A + B)/(C/D)$	$(A + B) \times C/D$
$(A - B)/(C + D)$	$(A - B)/(C + D)$	$(A - B)/(C - D)$	$(A - B)/(C - D)$
$(A - B)/(C \times D)$	$(A - B)/C/D$	$(A - B)/(C/D)$	$(A - B) \times C/D$
$(A \times B)/(C + D)$	$A \times B/(C + D)$	$(A \times B)/(C - D)$	$A \times B/(C - D)$
$(A \times B)/(C \times D)$	$A \times B/C/D$	$(A \times B)/(C/D)$	$A \times B \times C/D$
$(A/B)/(C + D)$	$A/(B + C)/D$	$(A/B)/(C - D)$	$A/(B - C)/D$
$(A/B)/(C \times D)$	$A/B/C/D$	$(A/B)/(C/D)$	$A \times B/C/D$
$(A + (B + C)) + D$	$A + B + C + D$	$(A + (B - C)) + D$	$A + B + C - D$
$(A + (B \times C)) + D$	$A \times B + C + D$	$(A + (B/C)) + D$	$A/B + C + D$
$(A - (B + C)) + D$	$A + B - C - D$	$(A - (B - C)) + D$	$A + B + C - D$
$(A - (B \times C)) + D$	$A + B - C \times D$	$(A - (B/C)) + D$	$A + B - C/D$
$(A \times (B + C)) + D$	$(A + B) \times C + D$	$(A \times (B - C)) + D$	$(A - B) \times C + D$
$(A \times (B \times C)) + D$	$A \times B \times C + D$	$(A \times (B/C)) + D$	$A \times B/C + D$
$(A/(B + C)) + D$	$A/(B + C) + D$	$(A/(B - C)) + D$	$A/(B - C) + D$
$(A/(B \times C)) + D$	$A/B/C + D$	$(A/(B/C)) + D$	$A \times B/C + D$
$(A + (B + C)) - D$	$A + B + C - D$	$(A + (B - C)) - D$	$A + B - C - D$
$(A + (B \times C)) - D$	$A \times B + C - D$	$(A + (B/C)) - D$	$A/B + C - D$
$(A - (B + C)) - D$	$A - B - C - D$	$(A - (B - C)) - D$	$A + B - C - D$
$(A - (B \times C)) - D$	$A - B \times C - D$	$(A - (B/C)) - D$	$A - B/C - D$
$(A \times (B + C)) - D$	$(A + B) \times C - D$	$(A \times (B - C)) - D$	$(A - B) \times C - D$
$(A \times (B \times C)) - D$	$A \times B \times C - D$	$(A \times (B/C)) - D$	$A \times B/C - D$
$(A/(B + C)) - D$	$A/(B + C) - D$	$(A/(B - C)) - D$	$A/(B - C) - D$
$(A/(B \times C)) - D$	$A/B/C - D$	$(A/(B/C)) - D$	$A \times B/C - D$
$(A + (B + C)) \times D$	$(A + B + C) \times D$	$(A + (B - C)) \times D$	$(A + B - C) \times D$
$(A + (B \times C)) \times D$	$(A \times B + C) \times D$	$(A + (B/C)) \times D$	$(A/B + C) \times D$
$(A - (B + C)) \times D$	$(A - B - C) \times D$	$(A - (B - C)) \times D$	$(A + B - C) \times D$
$(A - (B \times C)) \times D$	$(A - B \times C) \times D$	$(A - (B/C)) \times D$	$(A - B/C) \times D$
$(A \times (B + C)) \times D$	$(A + B) \times C \times D$	$(A \times (B - C)) \times D$	$(A - B) \times C \times D$
$(A \times (B \times C)) \times D$	$A \times B \times C \times D$	$(A \times (B/C)) \times D$	$A \times B \times C/D$
$(A/(B + C)) \times D$	$A \times B/(C + D)$	$(A/(B - C)) \times D$	$A \times B/(C - D)$
$(A/(B \times C)) \times D$	$A \times B/C/D$	$(A/(B/C)) \times D$	$A \times B \times C/D$
$(A + (B + C))/D$	$(A + B + C)/D$	$(A + (B - C))/D$	$(A + B - C)/D$
$(A + (B \times C))/D$	$(A \times B + C)/D$	$(A + (B/C))/D$	$(A/B + C)/D$
$(A - (B + C))/D$	$(A - B - C)/D$	$(A - (B - C))/D$	$(A + B - C)/D$
$(A - (B \times C))/D$	$(A - B \times C)/D$	$(A - (B/C))/D$	$(A - B/C)/D$
$(A \times (B + C))/D$	$(A + B) \times C/D$	$(A \times (B - C))/D$	$(A - B) \times C/D$
$(A \times (B \times C))/D$	$A \times B \times C/D$	$(A \times (B/C))/D$	$A \times B/C/D$
$(A/(B + C))/D$	$A/(B + C)/D$	$(A/(B - C))/D$	$A/(B - C)/D$
$(A/(B \times C))/D$	$A/B/C/D$	$(A/(B/C))/D$	$A \times B/C/D$
$((A + B) + C) + D$	$A + B + C + D$	$((A - B) + C) + D$	$A + B + C - D$
$((A \times B) + C) + D$	$A \times B + C + D$	$((A/B) + C) + D$	$A/B + C + D$
$((A + B) - C) + D$	$A + B + C - D$	$((A - B) - C) + D$	$A + B - C - D$
$((A \times B) - C) + D$	$A \times B + C - D$	$((A/B) - C) + D$	$A/B + C - D$
$((A + B) \times C) + D$	$(A + B) \times C + D$	$((A - B) \times C) + D$	$(A - B) \times C + D$
$((A \times B) \times C) + D$	$A \times B \times C + D$	$((A/B) \times C) + D$	$A \times B/C + D$

Table 18. Expressions and their Canonical Forms (Patterns) (Page 4 of 4)

Expression	Pattern	Expression	Pattern
$((A+B)/C)+D$	$(A+B)/C+D$	$((A-B)/C)+D$	$(A-B)/C+D$
$((A\times B)/C)+D$	$A\times B/C+D$	$((A/B)/C)+D$	$A/B/C+D$
$((A+B)+C)-D$	$A+B+C-D$	$((A-B)+C)-D$	$A+B-C-D$
$((A\times B)+C)-D$	$A\times B+C-D$	$((A/B)+C)-D$	$A/B+C-D$
$((A+B)-C)-D$	$A+B-C-D$	$((A-B)-C)-D$	$A-B-C-D$
$((A\times B)-C)-D$	$A\times B-C-D$	$((A/B)-C)-D$	$A/B-C-D$
$((A+B)\times C)-D$	$(A+B)\times C-D$	$((A-B)\times C)-D$	$(A-B)\times C-D$
$((A\times B)\times C)-D$	$A\times B\times C-D$	$((A/B)\times C)-D$	$A\times B/C-D$
$((A+B)/C)-D$	$(A+B)/C-D$	$((A-B)/C)-D$	$(A-B)/C-D$
$((A\times B)/C)-D$	$A\times B/C-D$	$((A/B)/C)-D$	$A/B/C-D$
$((A+B)+C)\times D$	$(A+B+C)\times D$	$((A-B)+C)\times D$	$(A+B-C)\times D$
$((A\times B)+C)\times D$	$(A\times B+C)\times D$	$((A/B)+C)\times D$	$(A/B+C)\times D$
$((A+B)-C)\times D$	$(A+B-C)\times D$	$((A-B)-C)\times D$	$(A-B-C)\times D$
$((A\times B)-C)\times D$	$(A\times B-C)\times D$	$((A/B)-C)\times D$	$(A/B-C)\times D$
$((A+B)\times C)\times D$	$(A+B)\times C\times D$	$((A-B)\times C)\times D$	$(A-B)\times C\times D$
$((A\times B)\times C)\times D$	$A\times B\times C\times D$	$((A/B)\times C)\times D$	$A\times B\times C/D$
$((A+B)/C)\times D$	$(A+B)\times C/D$	$((A-B)/C)\times D$	$(A-B)\times C/D$
$((A\times B)/C)\times D$	$A\times B\times C/D$	$((A/B)/C)\times D$	$A\times B/C/D$
$((A+B)+C)/D$	$(A+B+C)/D$	$((A-B)+C)/D$	$(A+B-C)/D$
$((A\times B)+C)/D$	$(A\times B+C)/D$	$((A/B)+C)/D$	$(A/B+C)/D$
$((A+B)-C)/D$	$(A+B-C)/D$	$((A-B)-C)/D$	$(A-B-C)/D$
$((A\times B)-C)/D$	$(A\times B-C)/D$	$((A/B)-C)/D$	$(A/B-C)/D$
$((A+B)\times C)/D$	$(A+B)\times C/D$	$((A-B)\times C)/D$	$(A-B)\times C/D$
$((A\times B)\times C)/D$	$A\times B\times C/D$	$((A/B)\times C)/D$	$A\times B/C/D$
$((A+B)/C)/D$	$(A+B)/C/D$	$((A-B)/C)/D$	$(A-B)/C/D$
$((A\times B)/C)/D$	$A\times B/C/D$	$((A/B)/C)/D$	$A/B/C/D$

I. Acknowledging LLM Usage

At various stages of the project, we received the aid of LLMs. Here, we report all such use cases.

- We received aid on writing the code to automate the process of extracting and canonicalizing an expression. The list of $n = 4$ patterns were initially manually mapped and were used as test cases for the LLM generated code. All code was manually verified.
- We received aid on writing the code to draw plots from existing data that were generated from human-written scripts. All code was manually verified.
- We used an LLM for copy-editing tasks, such as rephrasing sentences to improve clarity, conciseness, and flow, as well as for correcting grammatical errors.