Learning to Discover Abstractions for LLM Reasoning

Yuxiao Qu [*1] Anikait Singh [*2] Yoonho Lee [*2] Amrith Setlur [1]
Ruslan Salakhutdinov [1] Chelsea Finn [2] Aviral Kumar [1]

Abstract

Effective reasoning often requires going beyond pattern matching or memorization of solutions to identify and implement "algorithmic procedures" that can be used to deduce answers to hard problems. These algorithmic procedures consist of reusable primitives, intermediate results, or procedures that themselves can be applied across many problems. While current methods of RL post-training on long chains of thought ultimately desire to uncover this kind of algorithmic behavior, their sensitivity to benchmarks and the brittle and locally optimal nature of strategies learned by these systems suggest that this is far from a fulfilled promise. To instantiate this, we introduce ***reasoning abstractions***: concise natural language descriptions of procedural and factual knowledge that guide the model toward successful reasoning strategies. We train models to be capable of proposing several useful abstractions given a problem, followed by RL training that incentivizes building a solution while using the information provided by these abstractions. This results in a two-agent cooperative RL training paradigm, RL through Abstraction Discovery (***RLAD***), that jointly trains an abstraction generator and an abstraction-conditioned solution generator. This bi-level setup effectively enables structured exploration, decouples learning signals pertaining to abstraction proposal and solution generation, and improves generalization to harder problems, analogous to what we would expect from hierarchical RL. Empirically, ***RLAD*** improves performance on challenging math benchmarks.

1 Introduction

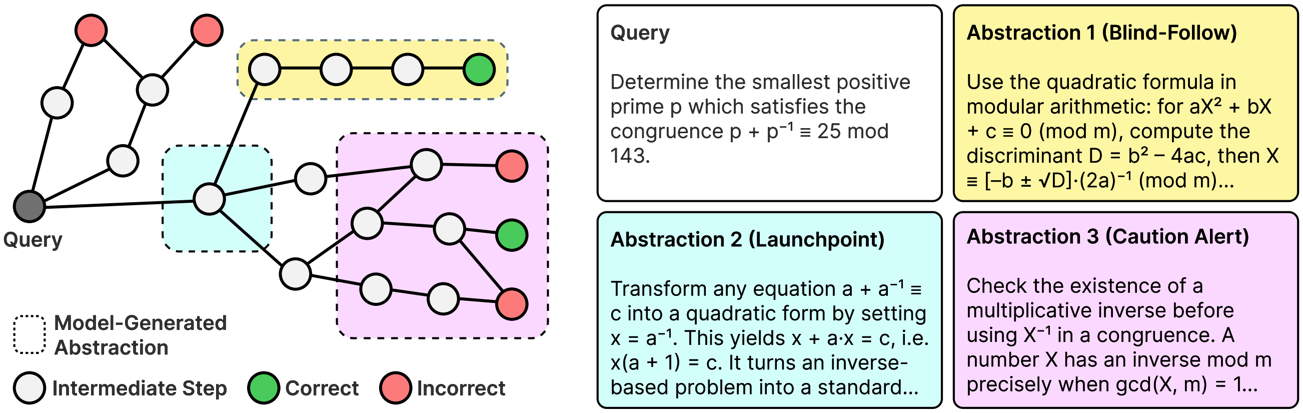
Modern machinery for solving reasoning tasks with large language models (LLMs) relies on incentivizing the use of longer chains of thought via reinforcement learning (RL). This training approach largely incentivizes "depth": subsequent training iterations increase response length by incorporating new operations that usually verify or build on top of the existing line of reasoning (Anonymous Author(s), 2025). In many hard problems, it is instead more desirable to optimize for "breadth": explore a diverse array of solution strategies, rather than committing to a seemingly optimal set of reasoning strategies right away (Yu et al., 2025; Yue et al., 2025). Optimizing for breadth is important: even when models optimized for depth succeed on some problems, they fail on structurally similar ones that require slightly different strategies, revealing brittle reasoning and poor generalization (Shi et al., 2023; Ma et al., 2024; Mirzadeh et al., 2024; Li et al., 2024; Petrov et al., 2025).

How can we help models discover a breadth of reasoning strategies for a given problem? Abstractly, the most natural approach is to train models to hypothesize new solutions to difficult problems and then attempt to utilize these strategies in the solution. We can do this by making models capable of discovering ***reasoning abstractions***: compressed representations of shared procedures that underlie multiple candidate solutions. For example, in math reasoning problems, such abstractions might correspond to useful intermediate lemmas or even some intermediate steps that do not succeed but illustrate what not to do. When presented in context, these abstractions function akin to hints on an exam, enabling LLMs to solve harder problems by building on the insights appearing in the abstraction, rather than from scratch. That is, when conditioned on abstractions, an LLM should learn to implement useful algorithmic procedures via RL that can utilize and compose the procedural information in the context as best as possible and apply it to the problem at hand. This naturally boosts the diversity of solution strategies and behaviors that a model learns to utilize when encountering an unseen problem, in contrast to committing to a narrow set of approaches like existing models. In RL terminology, abstractions serve as high-level subgoals, skills, or priors – any of them depending upon context – guiding the low-level solution-generating policy.

In this work, we imbue LLMs with the capability of proposing and utilizing reasoning abstractions for reasoning problems. Concretely, we build reasoning models that, first, given an input problem, propose one or more reasoning

[*] Equal contribution [1] Carnegie Mellon University [2] Stanford University. Correspondence to: Yuxiao Qu <yuxiaoq@andrew.cmu.edu>.

The second AI for MATH Workshop at the 42nd International Conference on Machine Learning, Vancouver, Canada. Copyright 2025 by the author(s).



Query

Determine the smallest positive prime p which satisfies the congruence $p + p^{-1} \equiv 25 \bmod 143$.

Abstraction 1 (Blind-Follow)

Use the quadratic formula in modular arithmetic: for $aX^2 + bX + c \equiv 0 \pmod{m}$, compute the discriminant $D = b^2 - 4ac$, then $X \equiv [-b \pm \sqrt{D}] \cdot (2a)^{-1} \pmod{m}$...

Abstraction 2 (Launchpoint)

Transform any equation $a + a^{-1} \equiv c$ into a quadratic form by setting $x = a^{-1}$. This yields $x + a \cdot x = c$, i.e. $x(a + 1) = c$. It turns an inverse-based problem into a standard...

Abstraction 3 (Caution Alert)

Check the existence of a multiplicative inverse before using X^{-1} in a congruence. A number X has an inverse mod m precisely when $\gcd(X, m) = 1$...

Figure 1. ***Reasoning abstractions illustrated in the solution-space graph for a problem.*** We represent the problem as a node (labeled "query") and various traces (both correct and incorrect) attempting to solve the problem as a graph. In this illustration, *reasoning abstractions* describe useful high-level structure in this space of all traces, such as **(1)** an abstract idea that can induce a predictable sequence of successful states (blind follow), **(2)** an initial step that informs the approach to take (launchpoint), or (3) a common critical error to avoid (caution alert). Note that reasoning abstractions encode a mix of procedural and factual knowledge that may be helpful.

abstractions, expressed in natural language. Subsequently, they generate a solution that utilizes the information and principles prescribed by these abstractions. To achieve this, we jointly train two LLMs via RL: **(1)** an abstraction generator, and **(2)** an abstraction-conditioned solution generator. The abstraction generator is rewarded for the improvement in the accuracy of the solution generator, stemming from the abstractions it provides. The solution generator is rewarded to maximize accuracy in solving a problem while utilizing the abstraction. To obtain a good initialization for RL training, we warmstart both models by running supervised fine-tuning (SFT) on data generated from a stronger models. For the abstraction generator, we collect multiple candidate solutions on a dataset of problems and prompt a stronger LLM to generate diverse abstractions. For the solution generator, we generate solutions from an LLM, conditioning on the abstraction. We call this approach RL through Abstraction Discovery (***RLAD***).

The main contribution of this paper is the notion of *reasoning abstractions*, how they can be obtained and amplified via RL training, and an illustration of how they can be used to improve reasoning performance. Concretely, we build an approach to imbue LLMs with the capability of proposing abstractions, and evaluate the model on a variety of math-reasoning benchmarks, AIME 2025 (Mathematical Association of America, 2025), DeepScaleR Hard (Anonymous Author(s), 2025), and AMC 2023. We find an average 44% improvement over state-of-the-art long chain-of-thought RL approaches (i.e., DAPO (Yu et al., 2025)) on AIME 2025, and show an effective benefit from generating diverse abstractions over brute-force solution sampling.

2 Related Work

Scaling test-time compute and exploration. Recent work highlights the promise of scaling test-time compute in different ways. One approach involves parallel sampling: sampling multiple reasoning rollouts and then selecting a winner via a scoring rule (Uesato et al., 2022; Wang et al., 2023; Charniak and Johnson, 2005; Feng et al., 2024; Snell et al., 2024; Yao et al., 2023a; Hao et al., 2023; Snell et al., 2024). A complementary line of work iteratively edits a single trace, attempting to implement some sort of a sequential search within a single solution trace (Madaan et al., 2023; Qu et al., 2024; Qu et al., 2024; Kumar et al., 2024). As such, the sequential approach performs a bit worse on harder problems (Snell et al., 2024; Qu et al., 2025), where it often gets trapped in strategies that seem optimal but aren't actually (Pan et al., 2025). Yet it still performs better than parallel search on easier and medium difficulty problems (Snell et al., 2024). Our approach of proposing and leveraging abstractions enables a kind of a hybrid between sequential sampling and parallel sampling, guided by the proposed abstractions. This should address failure modes of current methods. Prior work has also utilized hand-designed scaffolds to integrate multi-step evaluations of intermediate hypotheses into reasoning (Yao et al., 2023b; Ho et al., 2023; Hao et al., 2023; Li et al., 2023). In contrast, we do not rely on pre-defined interfaces but learn to *automatically* propose useful abstractions.

Using prior knowledge for LLM reasoning. Several threads of work converge on the idea that *textual artifacts* such as examples, plans, or prompts, can serve as reusable knowledge that steers LLM behavior. Existing retrieval-augmented generation (RAG) pipelines assume a static corpus, typically of human-written text, and focus on improving retrieval heuristics (Lewis et al., 2020; Borgeaud et al., 2022; Trivedi et al., 2022; Verma et al., 2024; Anonymous, 2025; Li et al., 2025). Many works use LLMs to learn or refine prompts, either in an input-agnostic fashion (Zhou et al., 2022; Yang et al., 2023; Pryzant et al., 2023; Fernando et al., 2023) or through input-specific edits based on feedback (Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2023; Yuksekgonul et al., 2025; Lin et al., 2025). Other related work explores the use of synthetic demonstrations (Zelikman et al., 2022b), scratchpads (Nye et al., 2021), and memory-augmented agents (Schäfer et al., 2020) to encode prior problem-solving knowledge. Two recent works demonstrate that LLMs can accumulate and reuse their own experience across tasks (Zhao et al., 2024; Suzgun et al., 2025). While one can view our reasoning abstractions as a form of prior procedural and factual knowledge produced before the model's solution attempt, this knowledge is **(a)** input-dependent and **(c)** is not acquired from an external source at deployment, but rather is "proposed" by the model itself. Imbuing models with this capability requires a two-player cooperative RL training procedure that we develop. To our knowledge, such procedures have not been used for generating textual artifacts of any type, let alone the abstractions we consider.

3 Preliminaries and Notation

We study reasoning with LLMs, where the LLM is provided access to a problem $\mathbf{x}$, and generates a stream of tokens $\widetilde{\mathbf{y}}$ that ends in an estimate of the answer. We assume access to a rule-based ground-truth 0/1 reward $\mathrm{Acc}_{\mathbf{x}}(\widetilde{\mathbf{y}}, \mathbf{y}^\star) \in \{0, 1\}$ that measures correctness of the produced answer $\widetilde{\mathbf{y}}$, against the ground-truth solution $\mathbf{y}^\star$ for a question $\mathbf{x}$. For training, we are given a dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, \mathbf{y}_i^\star)\}_{i=1}^N$ of problems $\mathbf{x}_i$ and solutions $\mathbf{y}_i^\star$ that end with the correct answer. Our goal is to train the LLM $\pi(\cdot|\mathbf{x})$ such that it achieves high rewards on a test distribution of problems $\mathcal{P}_{\text{test}}$.

We primarily evaluate models in terms of their average accuracy under $\mathcal{P}_{\text{test}}$. We also measure the pass@k metric, where for problem $\mathbf{x}$, we sample k solutions $\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_k \sim \pi(\cdot|\mathbf{x})$, and consider the problem to be solved if any of these k traces is correct. This metric couples accuracy with diversity, i.e., it attains the largest value when the model effectively finds diverse, good responses. To reduce variance in estimating pass@k, we sample $n \geq k$ samples per problem and use the unbiased estimator introduced in OpenAI Codex (Chen et al., 2021): $1 - \binom{n-c}{k} / \binom{n}{k}$, where $c \leq n$ is the number of correct samples.

4 Reasoning Abstractions and Why They Are Useful

Solving reasoning problems often requires composing both *procedural* knowledge (e.g., how to apply a root-finding algorithm) and *factual* knowledge (e.g., relevant lemmas or intermediate results). Current approaches typically train reasoning models to elicit such knowledge entirely through reinforcement learning (RL) with long chains of thought. However, this is often ineffective as RL often tends to optimize for "depth": producing longer traces where each subsequent segment extends the last (e.g., verifying prior calculations), rather than "breadth", which involves exploring diverse solution strategies. In this section, we introduce ***reasoning abstractions***, that provide a mechanism for explicitly encoding a range of procedural and factual concepts useful in solving a problem.

Intuition. We instantiate reasoning abstractions as concise textual descriptions of core insights that are useful for solving a problem. We show some examples of abstractions in Figure 1, in the domain of math reasoning. Here, these abstractions can correspond to useful techniques (e.g., "launchpoint" in Figure 1), a useful lemma or heuristic principle (e.g., "blind-follow" in Figure 1), and cautionary examples that demonstrate common pitfalls encountered when solving a problem (e.g., "caution alert" in Figure 1). These abstractions distill complex reasoning patterns and potential approaches into useful nuggets, allowing models to generalize across structurally similar problems.

Conceptual understanding. With this intuitive notion in place, we now consider a more conceptual definition. We can view abstractions as a compressed representation of the reasoning procedures embedded within longer chains of thought. Consider the space of possible reasoning traces for a given problem, which can be visualized as a graph structure where nodes represent intermediate states encountered when solving a question (see Figure 1). Good abstractions identify useful substructures within this larger reasoning graph. For example, an abstraction can capture if a set of strategies lead to a similar outcome or another set of tactics leads to an error being consistently made.

Concretely, let us denote the LLM policy that produces a solution conditioned on the problem $\mathbf{x}$ as $\pi_\theta^{\text{sol}}(\cdot|\mathbf{x})$. A good abstraction $\mathbf{z}$ is a sequence of tokens that provides some useful procedural and factual information to improve model performance:

$$\mathbb{E}_{\widetilde{\mathbf{y}} \sim \pi_\theta^{\text{sol}}(\cdot|\mathbf{x},\mathbf{z})} \left[\mathrm{Acc}(\widetilde{\mathbf{y}}, \mathbf{y}^*)\right] > \mathbb{E}_{\widetilde{\mathbf{y}} \sim \pi_\theta^{\text{sol}}(\cdot|\mathbf{x})} \left[\mathrm{Acc}(\widetilde{\mathbf{y}}, \mathbf{y}^*)\right]. \tag{1}$$

How can we generate good reasoning abstractions? Do good reasoning abstractions exist? We now attempt to understand whether good reasoning abstractions exist and how one might discover them. Perhaps the most natural way to obtain an initial set of reasoning abstractions is to collect a diverse set of traces attempting to solve a problem and then summarize useful concepts appearing in these traces, mimicking the illustration in Figure 1. To evaluate the existence and utility of reasoning abstractions (before developing our method to train LLMs to discover useful

	Breast Cancer Detection	Tweet Hate Speech Detection	Corporate Lobbying Relevance	Bank Note Authentication
Abstraction	Classify breast masses as malignant or benign using BI-RADS, shape, and margin criteria. ...If BI-RADS ≥ 5, then malignant. ...if BI-RADS = 4 AND shape = irregular AND margin = ill-defined, then malignant...	...If the text contains explicit derogatory slurs (e.g., b****, c***, s****, h**, r********), classify as hate speech. ...if the text degrades or dehumanizes a protected group (nationality, race, religion...	...If the bill addresses regulation, labeling, pricing, reimbursement, R&D funding, or licensing of the company's core products or services, label "Yes." ...Else if the bill alters taxes, credits, bonds, infrastructure...	1. If variance > 4, predict Fake. 2. Else if variance < −3, predict Original. 3. Else if skewness > 5, predict Fake. 4. Else if entropy > 0 and skewness > 1, predict Fake.
GPT-4o-mini	45%	60%	88%	45%
GPT-4o-mini + Abs	**90%**	**90%**	**94%**	**100%**

Figure 2. ***Examples of good reasoning abstractions in non-math domains***. Adding the abstraction to the prompt of GPT-4o-mini consistently improves performance on unseen instances.

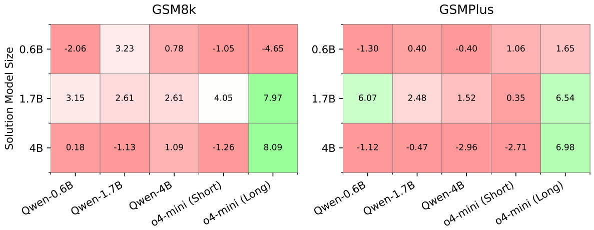


Figure 3. ***Many factors matter for consistent benefits from abstractions***: solver capability, abstraction length, and abstraction generator. We show relative accuracy change (%) for GSM8k and GSMPlus datasets. Rows indicate solution model size, and columns indicate the source of the abstraction.

reasoning abstractions), we instantiate this idea by prompting a model to generate solutions for a given problem and prompting a stronger model to deduce useful patterns from the responses of the first model. Concretely, we utilize the Qwen3 series of models to produce solutions and a stronger reasoning model, o4-mini, to generate abstractions. While this approach is not perfect, and it is not meant to be our final approach, it still enables us to validate the feasibility of the concept of reasoning abstractions. To ensure that the abstractions do not "leak" content of the solution, we verify that post-hoc prompting the solver with only the abstraction and no question yields zero accuracy.

Results and observations. After generating abstractions as above, we measure their quality by evaluating Equation 1, i.e., by checking if conditioning the problem solver on a set of abstractions improves its accuracy. Results in Figure 3 show that conditioning a problem solver on generated abstractions improves accuracy when three conditions hold simultaneously: (i) the abstraction is not too short (e.g., not just a few words that are not informative) and generated by a strong generator (`o4-mini (High)`) and (ii) the solution generator has sufficient capability (`Qwen3-1.7B` or `Qwen3-4B`) of interpreting and utilizing the generated abstraction. These results confirm that good abstractions (satisfying Eq. 1) exist for math problems, but neither the ability to generate them nor the ability to leverage them in solutions arises naturally. In Section 5, we will describe our method for explicitly training models to propose and use such abstractions effectively.

Good abstractions exist in many domains. We also find that this procedure can be used to identify an initial set of useful reasoning abstractions on many problem domains, including healthcare, human behavior, legal reasoning, and web security. Of course, the proportion of abstraction devoted to procedural knowledge and factual knowledge is different in these domains compared to math reasoning. Nonetheless, we find that using reasoning abstractions improves performance by 30% on average over 37 tasks from RAFT (Alex et al., 2021), CLUES (Menon et al., 2022), and LegalBench (Guha et al., 2023). We show four representative abstractions in Figure 2 and full results in Table 3 in the appendix.

> **Takeaways: Abstractions improve performance**
>
> Reasoning abstractions summarize procedural and factual knowledge that is useful for learning to solve problems via diverse strategies. Prompting abstractions generated by merely prompting models already improves performance by 30% on average for reasoning.

5 Learning to Discover Reasoning Abstractions

Having defined the notion of reasoning abstractions and shown that they can improve performance when adhered to for tackling reasoning problems, we now wish to develop an approach that can allow us to imbue and improve an LLM's ability to propose and utilize abstractions. Doing so requires training an abstraction generator: an LLM, $\mathbf{z} \sim \pi_\theta^{\text{abs}}(\cdot|\mathbf{x})$ that proposes candidate abstractions $\mathbf{z}$ given problem $\mathbf{x}$, and an abstraction-conditioned solution generator, $\mathbf{y} \sim \pi_\theta^{\text{sol}}(\cdot|\mathbf{x}, \mathbf{z})$, that produces a solution $\mathbf{y}$ given $\mathbf{x}$ and abstraction $\mathbf{z}$. Note that $\mathbf{z}$ is parameterized as a variable-length string of tokens and might consist of one or more pieces of information or procedures. While our approach applies to the case when π_θ^{abs} produces more than one abstraction, we abuse notation and subsume more than one abstraction into one to avoid notational clutter. In this section, we describe RL through Abstraction Discovery (***RLAD***), our method for training these models via RL.

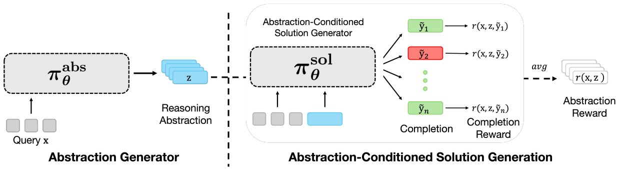


Figure 4. **RLAD** ***training paradigm.*** We train an abstraction generator, π_θ^{abs}, that proposes some reasoning abstractions conditioned on the question $\mathbf{x}$, denoted as $\mathbf{z}$. Then, the solution generator, π_θ^{sol}, is trained to produce a response, $\tilde{\mathbf{y}}$ conditioned on the generated abstraction $\mathbf{z}$. The reward used for training π_θ^{abs} corresponds to the average success rate of the solution generator conditioned on the proposed abstraction.

5.1 Training π_θ^{abs} and π_θ^{sol} via RL

The core principle behind our approach is that an abstraction $\mathbf{z}$ is successful at a given problem $\mathbf{x}$ if it can maximally help $\pi_\theta^{\text{sol}}(\cdot|\mathbf{x}, \mathbf{z})$ find correct responses to question $\mathbf{x}$, without actually leaking the answer itself. To convert this into an RL objective, we design a reward function that rewards an abstraction $\mathbf{z}$ with the expected success of solutions generated by π_θ^{sol} conditioned on $\mathbf{z}$:

$$r_{\pi_\theta^{\text{sol}}}(\mathbf{x}, \mathbf{z}) := \mathbb{E}_{\tilde{\mathbf{y}} \sim \pi_\theta^{\text{sol}}(\cdot|\mathbf{x}, \mathbf{z})} \left[\text{Acc}_{\mathbf{x}}(\tilde{\mathbf{y}}, \mathbf{y}^*) \right], \tag{2}$$

where $\mathbf{y}^*$ is the groundtruth answer and $\text{Acc}_{\mathbf{x}}(\cdot, \cdot)$ denotes the 0/1 accuracy on problem $\mathbf{x}$. To train π_θ^{sol}, one can then adopt the fairly straightforward approach of maximizing 0/1 binary outcome reward, now conditioned on a given abstraction $\mathbf{z}$ sampled previously from π_θ^{abs}, akin to recent results RL (DeepSeek-AI et al., 2025). Formally, we set the reward for a solution as: $r(\mathbf{x}, \mathbf{z}, \tilde{\mathbf{y}}) := \text{Acc}_{\mathbf{x}}(\tilde{\mathbf{y}}, \mathbf{y}^*)$. With these reward functions in place, perhaps the most natural approach then would be to train π_θ^{abs} to maximize $r_{\pi_\theta^{\text{sol}}}$ for a fixed π_θ^{sol} on a dataset of prompts $\mathcal{D}_{\pi_\theta^{\text{abs}}}$, while also iteratively training π_θ^{sol} to maximize the reward function r on modified prompts generated by concatenating a set of sampled abstraction $\mathbf{z}$ on a dataset of problems, $\mathcal{D}_{\pi_\theta^{\text{sol}}}$. This maximization could be done via on-policy RL methods like GRPO (Shao et al., 2024) or (batched) offline RL methods like DPO (Rafailov et al., 2023) and STaR (Zelikman et al., 2022a).

Challenges with naïve reward design. While the approach so far is extremely simple, it presents some challenges. In particular, the reward functions defined above can result in spurious, undesirable solutions in a rather nuanced manner: **(1)** if π_θ^{abs} learns to solve problem $\mathbf{x}$ in its entirety, it will still be rewarded highly by $r_{\pi_\theta^{\text{sol}}}$ but is not a desirable abstraction; **(2)** if π_θ^{sol} is too weak or too strong, such that it is either able to always solve the problem $\mathbf{x}$ or never solves it, then $r_{\pi_\theta^{\text{sol}}}$ will not provide a meaningful signal to update π_θ^{abs}; and **(3)** similar to the above failure modes, training π_θ^{sol} via on-policy RL may result in it ignoring the abstraction $\mathbf{z}$ altogether no matter how useful it is. Abstractly, all of these challenges correspond to a "signal obfuscation" problem, where an imbalance in the strength of π_θ^{abs} and π_θ^{sol} may drown out the learning signal for the other.

Modifying reward design. To address these signal obfuscation challenges, we make a slight but consequential changes to the training process. In particular, we train π_θ^{sol} on a mixture of prompts $\mathbf{x}$ augmented by abstractions $\mathbf{z}$ and prompts $\mathbf{x}$ without any abstractions at all. In this process, while we utilize $\text{Acc}_{\mathbf{x}}$ as discussed above on a given response, we simply zero out rewards for any trace generated on $\mathbf{x}$ without abstractions. When utilizing KL-constrained RL, e.g., GRPO (Shao et al., 2024), π_θ^{sol} is now trained to closely mimic the distribution of responses as the reference LLM on questions $\mathbf{x}$ but must attempt to find ways to optimize reward on the same question $\mathbf{x}$ when augmented with an abstraction. This can be accomplished only when π_θ^{sol} learns to utilize the provided abstraction carefully, hence addressing one of the challenges above. Second, we ensure that $\mathbf{z} \sim \pi_\theta^{\text{abs}}(\cdot|\mathbf{x})$ itself does not contain the answer to the question $\mathbf{x}$, which means that $\text{Acc}(\mathbf{z}, \mathbf{y}^*)$ is penalized to be small. Finally, we utilize separate partitions of the training dataset to train π_θ^{abs} and π_θ^{sol} to avoid overfitting on subsets of data. We present detailed ablations of these design choices in Appendix B.3. Formally, the updated versions of these reward functions are shown as:

$$r(\mathbf{x}, \mathbf{z}, \tilde{\mathbf{y}}) := \begin{cases} 0, & \text{if } \mathbf{z} = \emptyset \\ \text{Acc}_{\mathbf{x}}(\tilde{\mathbf{y}}, \mathbf{y}^*), & \text{otherwise} \end{cases} \tag{3}$$

$$r_{\pi_\theta^{\text{sol}}}(\mathbf{x}, \mathbf{z}) := \mathbb{E}_{\tilde{\mathbf{y}} \sim \pi_\theta^{\text{sol}}(\cdot|\mathbf{x}, \mathbf{z})} [\text{Acc}_{\mathbf{x}}(\tilde{\mathbf{y}}, \mathbf{y}^*)]. \tag{4}$$

5.2 Warmstarting π_θ^{sol} and π_θ^{abs} from Good Initializations

While the above approach prescribes a recipe for RL training of π_θ^{abs} and π_θ^{sol}, any such recipe critically relies on the ability of the initialization to be able to generate somewhat meaningful abstractions and meaningful solutions conditioned on the abstraction input, respectively, right from the beginning of RL training. How can we ensure that our model initializations have this capability? Inspired from the approach of running an initial phase of SFT to imbue into the model the basic structure of a long chain-of-thought before running RL (DeepSeek-AI et al., 2025; Qu et al., 2025), we run an initial phase of SFT to imbue into π_θ^{abs} and π_θ^{sol} the basic capabilities of producing abstractions and attempting to follow abstractions respectively, even if the resulting models are not very good. For this initial warmstart phase, we follow the protocol from Section 4 and construct a corpus $\{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)\}_{i=1}^M$ by prompting strong models. For each training problem-solution pair $(\mathbf{x}, \mathbf{y}^*)$, in our training set, we first generate a abstraction $\mathbf{z}$ using an instruction-tuned model, discarding any that leak $\mathbf{y}^*$. We then sample a solution trace $\mathbf{y}$ conditioned on $(\mathbf{x}, \mathbf{z})$. As mentioned in Section 5.1, we partition this corpus into non-overlapping splits for π_θ^{sol} and π_θ^{abs} to avoid overfitting.

5.3 Practical Approach and Algorithm Details

For warmstarting the abstraction generator, we utilize abstractions generated by `o4-mini`. We then use a weaker solution generator (`GPT 4.1-mini`) to check the efficacy of each abstraction when conditioned on by comparing the success rate of the solution generator with and without a abstraction. We filter abstractions that don't result in an increase in solution generation performance to form our seed set of abstractions. Then, we run SFT for 5 epochs on the seed dataset to obtain an initial abstraction generator. For solution generation, we utilize Qwen 3 1.7b (Qwen Team, 2025), a 1.7B reasoning model distilled from Qwen 3-32B.

After SFT, we employ ***RLAD*** to further fine-tune the abstraction generator and abstraction-conditioned solution generator via RL. For the abstraction generator, we opt to use "batched" offline RL instantiation of our approach via RFT (Yuan et al., 2023) and RPO (Pang et al., 2024), since reward computation by rolling out the solution generator the on the fly was infeasible in our RL infrastructure and compute. To train the solution generator, we utilize the DAPO approach (Yu et al., 2025), and include token-level policy loss normalization and asymmetric clipping, and prompt difficulty/length curriculum. Building upon implementation of concurrent work (Anonymous Author(s), 2025), we employ a two stage curriculum where we partition the DeepScaleR (Luo et al., 2025) mixture by success rate of the base model into three sets: (1) easy, (2) medium, and (3) hard, where we fine-tune first on easy problems with an 8K token budget and then on medium problems. We utilize the hard split as a held out, evaluation subset, which we denote as `DeepScaleR [Hard]`. We outline hyperparameters and details in Appendix A.1 and provide a pseudocode in Algorithm 1.

Takeaways: ***RLAD*** method design

RLAD jointly optimizes the abstraction generator π_θ^{abs} and solution generator π_θ^{sol} with RL, in an alternating manner, using reward functions in Equation 3. These reward functions incentivize π_θ^{sol} to utilize abstractions and incentivize π_θ^{abs} to never "leak" the response to the input problem. We obtain the initialization for π_θ^{abs} by distilling traces from a stronger teacher model (`o4-mini`).

6 Experimental Evaluation

The goal of our experiments is to evaluate the efficacy of ***RLAD*** in improving the reasoning capabilities of LLMs through abstraction-guided solution generation. Specifically, we aim to answer the following research questions: **(1)** Does ***RLAD*** improve pass@1 accuracy across several mathematical reasoning benchmarks compared to direct solution generation? **(2)** How does ***RLAD*** scale as more abstractions and solutions are generated? **(3)** What makes the generated abstractions useful, how faithfully are they followed, and how do they guide and improve solution generation? To this end, we compare ***RLAD*** with strong base models on three representative mathematical reasoning datasets: AMC 2023, AIME 2025, and DeepScaleR Hard (Luo et al., 2025), which itself is a subset of hard problems from the OmniMATH mixture on which DeepSeek-R1 distilled Qwen-32B model attains an accuracy of $\leq 10\%$. We also conduct several ablations to better understand the abstractions produced by ***RLAD***.

6.1 Main Performance Results for *RLAD*

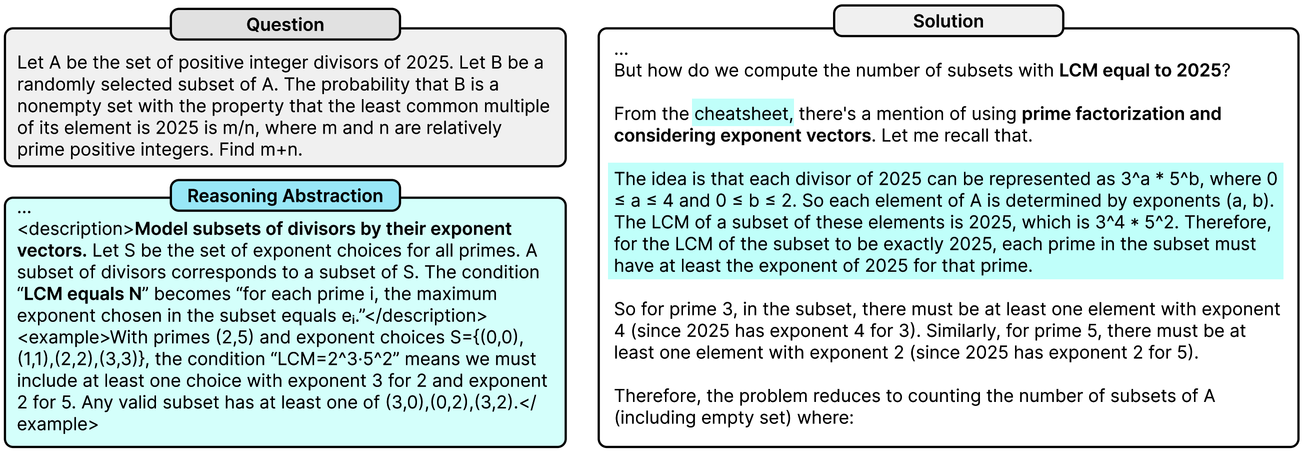
We evaluate ***RLAD*** in three settings: (1) **w/o abs**,without abstractions; (2) **w/ abs (avg)**, average performance over generations conditioned on 4 proposed abstractions per problem; and (3) **w/ abs (best)**: using the best-performing abstraction (in a set of 4 proposed abstractions per problem).

Observe that ***RLAD*** consistently outperforms the base model and variant fine-tuned with RL on the same prompts via DAPO (Yu et al., 2025), but without any abstractions, across all settings and benchmarks (Table 1). This highlights that ***RLAD*** can propose and leverage abstractions to improve its reasoning performance. We also note that these performance gains are not limited to abstraction-conditioned inference: even in the **w/o abs** setting, where no abstraction is provided during inference, ***RLAD*** improves over the prior methods, when trained with abstractions via ***RLAD***. This suggests that exposure to diverse abstractions during training enhances the model's general reasoning ability. We observe similar trends on additional benchmarks, including AIME 2024 and HMMT 2025 (see Appendix B.2), where ***RLAD*** improves in the w/o abs setting.

In Appendix C, we also measure the performance of ***RLAD*** when different budgets are allowed for reasoning – while Table 1 measures performance at a budget of 32K tokens, we also measure performance at 8K and 16K budgets and find ***RLAD*** to be more effective compared to the comparisons.

6.2 Understanding Properties of *RLAD*

Compute tradeoffs between abstraction and solution generation. We now study how to allocate compute between generating diverse abstractions and sampling solutions conditioned on them to attain maximal performance within a given budget on the total sampling allowed. This corresponds to a "compute-optimal strategy" (Snell et al., 2024) for partitioning compute between abstraction and solution generation. If the model typically fails by making small local errors in its computations, then additional concise abstractions may not help it as much as simply trying again. In contrast, if the model tends to pursue a seemingly plausible but incorrect approach and is unable to easily recover or switch approaches, then conditioning on diverse abstractions can help by offering alternative high-level approaches toward the correct answer. In other words, when the model



Question

Let A be the set of positive integer divisors of 2025. Let B be a randomly selected subset of A. The probability that B is a nonempty set with the property that the least common multiple of its element is 2025 is m/n, where m and n are relatively prime positive integers. Find m+n.

Reasoning Abstraction

...
<description>**Model subsets of divisors by their exponent vectors.** Let S be the set of exponent choices for all primes. A subset of divisors corresponds to a subset of S. The condition "**LCM equals N**" becomes "for each prime i, the maximum exponent chosen in the subset equals e_i."</description>
<example>With primes (2,5) and exponent choices S={(0,0),(1,1),(2,2),(3,3)}, the condition "LCM=2^3·5^2" means we must include at least one choice with exponent 3 for 2 and exponent 2 for 5. Any valid subset has at least one of (3,0),(0,2),(3,2).</example>

Solution

...
But how do we compute the number of subsets with **LCM equal to 2025**?

From the cheatsheet, there's a mention of using **prime factorization and considering exponent vectors.** Let me recall that.

The idea is that each divisor of 2025 can be represented as 3^a * 5^b, where 0 ≤ a ≤ 4 and 0 ≤ b ≤ 2. So each element of A is determined by exponents (a, b). The LCM of a subset of these elements is 2025, which is 3^4 * 5^2. Therefore, for the LCM of the subset to be exactly 2025, each prime in the subset must have at least the exponent of 2025 for that prime.

So for prime 3, in the subset, there must be at least one element with exponent 4 (since 2025 has exponent 4 for 3). Similarly, for prime 5, there must be at least one element with exponent 2 (since 2025 has exponent 2 for 5).

Therefore, the problem reduces to counting the number of subsets of A (including empty set) where:

Figure 5. ***Example of a reasoning abstraction.*** Here, we provide an example of a reasoning abstraction for a given problem. In the solution, we see (**in blue**) references to the abstraction and keywords from the abstraction being utilized in the thinking trace of the reasoning model.

Approach	**AIME 2025**			**DeepScaleR [Hard]**			**AMC 2023**		
	w/o abs	w/ abs (avg)	w/ abs (best)	w/o abs	w/ abs (avg)	w/ abs (best)	w/o abs	w/ abs (avg)	w/ abs (best)
Qwen-3-1.7B	33.75	36.25	40.00	20.21	22.14	32.50	86.41	78.01	84.53
+ DAPO	37.92	34.90	39.79	21.67	21.88	33.54	86.41	81.99	88.44
+ ***RLAD***	38.04	42.45	48.33	23.54	24.84	35.54	87.25	88.35	91.72

Table 1. **Pass@1 accuracy across three math reasoning benchmarks.** ***RLAD*** achieves consistent gains in both abstraction-conditioned and w/o abstraction settings.

has a tendency to explore "depth" over "breadth" of solution strategies, abstractions can help improve performance. With this intuition, we hypothesize that when the compute budget permits only a limited number of samples, allocating more compute into sampling multiple solutions will enable the model to succeed at least once. In order words, sampling multiple solutions for the same abstraction will result in a higher pass@k performance. However, once pass@k for a single abstraction begins to saturate, performance gains are more likely to come from scaling diversity of abstractions, which enables the model to explore qualitatively different regions of the solution space.

To visualize this tradeoff, we plot *iso-compute* scaling curves under a fixed compute budget, where multiple abstractions are generated and multiple solutions are sampled per abstraction. Specifically, we denote the number of abstractions as m and the number of solutions sampled per abstraction as k. To better isolate the effect of abstraction diversity, we introduce a normalization offset k_0, which accounts for performance gains that do not stem from new strategies, but arise from local modifications in the solution and the model's own stochasticity (e.g., small edits that do not require new abstractions). Figure 6 shows multiple *iso-compute* frontiers, one for each total compute budget. Each curve corresponds to a fixed total number of abstraction-conditioned samples, with compute defined as $m \times (k - k_0)$, where m is the number of abstractions, k is the number of solutions per abstraction, and k_0 offsets for solutions. This formulation captures the number of "meaningful" samples that go beyond the model's local neighborhood. The x-axis plots the ratio between abstractions and adjusted solutions, $m/(k - k_0)$ We observe in Figure 6 that across $k_0 \in \{0, 2, 4, 6, 8\}$, shifting compute toward abstractions consistently yields greater performance improvements than allocating the same additional compute to solution refinements. ***This supports the conclusion that*** *once local errors in the chain-of-thought have been addressed, it is more effective to increase the breadth of the search through abstraction conditioning rather than to continuing to scale up sampling alone.*

Understanding behavior of the abstraction-conditioned solution generator. A desirable property of the solution generator is the ability to follow proposed abstractions. To study this, we prompt a strong reasoning model `o4-mini` to classify whether a particular solution trace produced by a trained solution generator closely adheres to a given abstraction. We ask for a binary decision on each pair of hint and solution, and measure the adherence rate across 200 pairs. In Figure 7 (right), we report adherence rates under four conditions: `Abstraction` (solution generated with the intended abstraction), `No Abstraction` (solution with only question), `Retrieval` (a semantically similar past solution is retrieved), and `Unrelated Abstraction` (solution conditioned on an abstraction from a different prob-

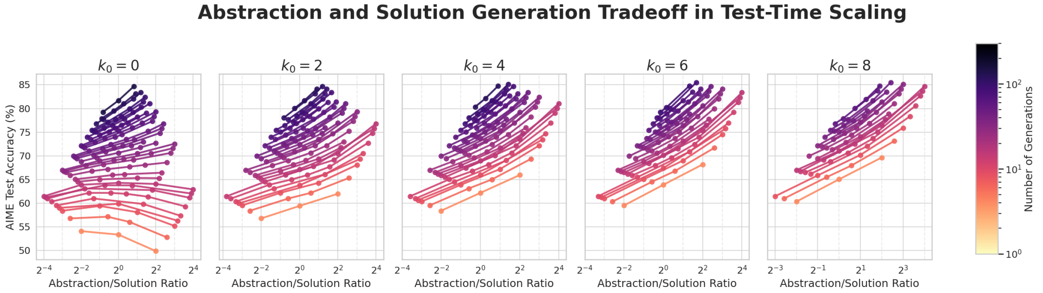


Figure 6. ***Tradeoff of abstraction and solution generation on AIME 2025.*** As the compute budget increases, we find better performance efficiency when allocating our budget to abstraction generation rather than solution generation, for all values of normalization offset k_0.

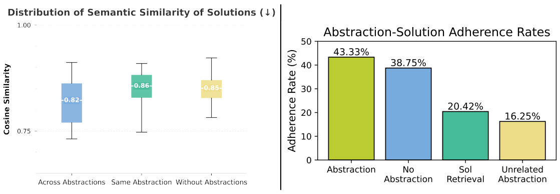


Figure 7. **Abstraction-conditioned solution generation analysis.** ***RLAD*** produces solutions with **(left)** greater semantic diversity across different abstractions and **(right)** higher abstraction adherence than baselines.

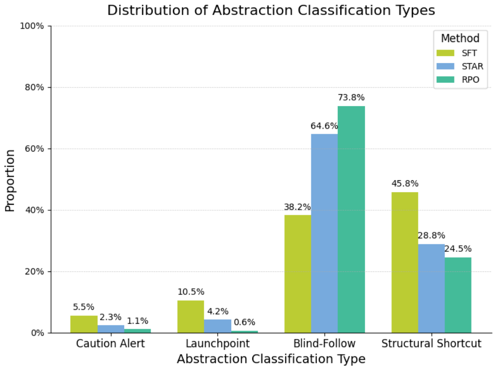


Figure 8. **Abstraction Categorization.** ***RLAD*** produces a diverse characterization of abstractions (characterize `o4-mini`).

lem). We find that the `Abstraction` condition achieves the highest adherence rate, outperforming all control variants on average. Intuitively, this means that the trained solution generator is detected to be more likely to follow the strategy or guidance of a given abstraction. Additionally, we measure the semantic similarity of solutions generated without abstraction conditioning, conditioned on the same abstraction, and across abstractions. Here we find across abstractions, the semantic similarity of solutions is lower, indicating abstractions allow for higher solution diversity.

Categorizing abstractions. As outlined in Appendix C, we classify each model-generated abstraction into four mutually-exclusive categories: (1) `Caution Alert` that warns the solver to avoid a specific approach; (2) `Productive Launchpoint` that suggests strategic framings or problem reformulations that open high-potential solution paths; (3) `Blind-Follow Trajectory` that prescribes repeatable, step-by-step procedures executable without further insight; and (4) `Structural Shortcut` that leverages abstract insights or invariants to collapse multiple reasoning steps into a single leap. In Figure 8, we show that after training via ***RLAD***, the distribution over these categories shifts, with a notable increase in blind-follow abstractions, which a stronger reasoning model classifies as an effective reasoning path to a successful solution as seen in Appendix C.

> **Takeaways: Experimental Results**
>
> ***RLAD*** outperforms RL fine-tuning approaches that do not propose or leverage abstractions on math reasoning. Jointly scaling the number of abstractions and solution samples enables continued performance gains even when scaling solutions alone begins to saturate.

7 Discussion, Conclusion

We introduce reasoning abstractions: concise representations of procedural and factual knowledge that are expressed in natural language, as a means to broaden the reasoning strategies of LLMs. Our method, ***RLAD***, instantiates a two-player training framework that trains an abstraction generator and an abstraction-conditioned solution generator. ***RLAD*** yields consistent improvements across several mathematical reasoning benchmarks, outperforming existing methods for training LLMs to reason. Moreover, we show that allocating compute toward generating diverse abstractions, rather than increasing solution sampling alone, leads to greater performance gains. This highlights abstractions as a complementary axis for scaling test-time compute: while longer length chains-of-thoughts and parallel solution sampling provide some existing ways to scale compute, using abstractions provides us with an orthogonal axis to improve performance. While we showed abstractions can be helpful, we limited our evaluation to mathematical reasoning tasks, leaving open-ended reasoning unexplored. Finally, ***RLAD*** incurs additional computational overhead, and training a single model that can both generate abstractions and solutions is open for future work.

References

N. Alex, E. Lifland, L. Tunstall, A. Thakur, P. Maham, C. J. Riedel, E. Hine, C. Ashurst, P. Sedille, A. Carlier, et al. Raft: A real-world few-shot text classification benchmark. *arXiv preprint arXiv:2109.14076*, 2021.

Anonymous. Optimizing inference-time reasoning in LLMs via retrieval-augmented reflection, 2025. URL https://openreview.net/forum?id=ElYRG3pJcv.

Anonymous Author(s). e3: Learning to Explore Enables Extrapolation of Test-Time Compute for LLMs. Submitted to the 39th Conference on Neural Information Processing Systems (NeurIPS 2025), 2025.

S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 173–180, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219862. URL https://doi.org/10.3115/1219840.1219862.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

X. Feng, Z. Wan, M. Wen, S. M. McAleer, Y. Wen, W. Zhang, and J. Wang. Alphazero-like tree-search can guide large language model decoding and training, 2024. URL https://arxiv.org/abs/2309.17179.

C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36: 44123–44279, 2023.

S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. Reasoning with language model is planning with world model, 2023. URL https://arxiv.org/abs/2305.14992.

N. Ho, L. Schmid, and S.-Y. Yun. Large language models are reasoning teachers, 2023. URL https://arxiv.org/abs/2212.10071.

A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, L. M. Zhang, K. McKinney, D. Shrivastava, C. Paduraru, G. Tucker, D. Precup, F. Behbahani, and A. Faust. Training language models to self-correct via reinforcement learning, 2024. URL https://arxiv.org/abs/2409.12917.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. Camel: Communicative agents for ”mind” exploration of large language model society, 2023. URL https://arxiv.org/abs/2303.17760.

Q. Li, L. Cui, X. Zhao, L. Kong, and W. Bi. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint arXiv:2402.19255*, 2024.

X. Li, G. Dong, J. Jin, Y. Zhang, Y. Zhou, Y. Zhu, P. Zhang, and Z. Dou. Search-o1: Agentic search-enhanced large reasoning models, 2025. URL https://arxiv.org/abs/2501.05366.

K. Lin, C. Snell, Y. Wang, C. Packer, S. Wooders, I. Stoica, and J. E. Gonzalez. Sleep-time compute: Beyond inference scaling at test-time. *arXiv preprint arXiv:2504.13171*, 2025.

M. Luo, S. Tan, J. Wong, X. Shi, W. Tang, M. Roongta, C. Cai, J. Luo, T. Zhang, E. Li, R. A. Popa, and I. Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca3030 2025. Notion Blog.

J. Ma, D. Dai, L. Sha, and Z. Sui. Large language models are unconscious of unreasonability in math problems. *arXiv preprint arXiv:2403.19346*, 2024.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. Prasad Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv e-prints*, art. arXiv:2303.17651, Mar. 2023. doi: 10.48550/arXiv.2303.17651.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594, 2023.

Mathematical Association of America. 2025 American Invitational Mathematics Examination (AIME) I: Problems and Solutions, Feb. 2025. URL https://artofproblemsolving.com/wiki/index.php/2025_AIME_I. Art of Problem Solving Wiki entry.

R. R. Menon, S. Ghosh, and S. Srivastava. Clues a benchmark for learning classifiers using natural language explanations. 2022.

I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena. Show your work: Scratchpads for intermediate computation with language models, 2021. URL https://arxiv.org/abs/2112.00114.

J. Pan, X. Li, L. Lian, C. Snell, Y. Zhou, A. Yala, T. Darrell, K. Keutzer, and A. Suhr. Learning adaptive parallel reasoning with language models. *arXiv preprint arXiv:2504.15466*, 2025.

R. Y. Pang, W. Yuan, K. Cho, H. He, S. Sukhbaatar, and J. Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

I. Petrov, J. Dekoninck, L. Baltadzhiev, M. Drencheva, K. Minchev, M. Balunović, N. Jovanović, and M. Vechev. Proof or bluff? evaluating llms on 2025 usa math olympiad. *arXiv preprint arXiv:2503.21934*, 2025.

R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng. Automatic prompt optimization with” gradient descent” and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

Y. Qu, T. Zhang, N. Garg, and A. Kumar. Recursive Introspection: Teaching Language Model Agents How to Self-Improve. *arXiv e-prints*, art. arXiv:2407.18219, July 2024. doi: 10.48550/arXiv.2407.18219.

Y. Qu, T. Zhang, N. Garg, and A. Kumar. Recursive introspection: Teaching language model agents how to self-improve. *arXiv preprint arXiv:2407.18219*, 2024.

Y. Qu, M. Y. Yang, A. Setlur, L. Tunstall, E. E. Beeching, R. Salakhutdinov, and A. Kumar. Optimizing test-time compute via meta reinforcement fine-tuning. *arXiv preprint arXiv:2503.07572*, 2025.

Qwen Team. Qwen3 technical report. Technical report, Qwen, May 2025. URL https://huggingface.co/Qwen. Available at: https://huggingface.co/Qwen, https://modelscope.cn/organization/qwen, https://github.com/QwenLM/Qwen3.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

M. B. Schäfer, F. Ohme, and A. H. Nitz. Detection of gravitational-wave signals from binary neutron star mergers using machine learning. *Physical Review D*, 102(6), Sept. 2020. ISSN 2470-0029. doi: 10.1103/physrevd.102.063015. URL http://dx.doi.org/10.1103/PhysRevD.102.063015.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv e-prints*, art. arXiv:2408.03314, Aug. 2024. doi: 10.48550/arXiv.2408.03314.

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

M. Suzgun, M. Yuksekgonul, F. Bianchi, D. Jurafsky, and J. Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*, 2025.

H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022.

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275.

P. Verma, S. P. Midigeshi, G. Sinha, A. Solin, N. Natarajan, and A. Sharma. Plan×RAG: Planning-guided Retrieval Augmented Generation. *arXiv e-prints*, art. arXiv:2410.20753, Oct. 2024. doi: 10.48550/arXiv.2410.20753.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023a. URL https://arxiv.org/abs/2305.10601.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models, 2023b. URL https://arxiv.org/abs/2210.03629.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Z. Yuan, H. Yuan, C. Li, G. Dong, C. Tan, and C. Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, S. Song, and G. Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

M. Yuksekgonul, F. Bianchi, J. Boen, S. Liu, P. Lu, Z. Huang, C. Guestrin, and J. Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.

E. Zelikman, Y. Wu, J. Mu, and N. Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022a.

E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022b. URL https://arxiv.org/abs/2203.14465.

A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, and G. Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642, 2024.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.

Appendices

A Experimental Details

A.1 Pseudocode for *RLAD*

Algorithm 1 Joint RL Training of π_θ^{abs} and π_θ^{sol}

```
Require: Policies $\pi_\theta^{\text{abs}}(\mathbf{z} \mid \mathbf{x})$, $\pi_\theta^{\text{sol}}(\tilde{\mathbf{y}} \mid \mathbf{x}, \mathbf{z})$ Datasets $\mathcal{D}_{\pi_\theta^{\text{abs}}},\ \mathcal{D}_{\pi_\theta^{\text{sol}}}$; rates $\alpha_{\pi_\theta^{\text{abs}}}, \alpha_{\pi_\theta^{\text{sol}}}$; batch sizes $N, M$; epochs $E$
Initialize $\pi_\theta^{\text{abs}}, \pi_\theta^{\text{sol}}$
for $e = 1$ to $E$ do                                                    ▷ Update abstraction policy
    for $\{\mathbf{x}_i\}_{i=1}^{N} \sim \mathcal{D}_{\pi_\theta^{\text{abs}}}$ do
        $\mathbf{z}_i \sim \pi_\theta^{\text{abs}}(\cdot|\mathbf{x}_i)$
        $r_i \leftarrow r_{\pi_\theta^{\text{sol}}}(\mathbf{x}_i, \mathbf{z}_i)$
        $\pi_\theta^{\text{abs}} \leftarrow \pi_\theta^{\text{abs}} - \alpha_{\pi_\theta^{\text{abs}}} \nabla_{\pi_\theta^{\text{abs}}} \mathcal{L}_{\text{STAR/RPO}}(\pi_\theta^{\text{abs}}; \mathbf{x}_i, \mathbf{z}_i, r_i)$
    end for                                                              ▷ Update solution policy
    for $\{\mathbf{x}_j\}_{j=1}^{M} \sim \mathcal{D}_{\pi_\theta^{\text{sol}}}$ do
        $\mathbf{z}_j \sim \pi_\theta^{\text{abs}}(\cdot|\mathbf{x}_j), \quad \tilde{\mathbf{y}}_j \sim \pi_\theta^{\text{sol}}(\cdot|\mathbf{x}_j, \mathbf{z}_j)$
        $r_j \leftarrow r(\mathbf{x}_j, \mathbf{z}_j, \tilde{\mathbf{y}}_j)$
        $\pi_\theta^{\text{sol}} \leftarrow \pi_\theta^{\text{sol}} - \alpha_{\pi_\theta^{\text{sol}}} \nabla_{\pi_\theta^{\text{sol}}} \mathcal{L}_{\text{GRPO}}(\pi_\theta^{\text{sol}}; \mathbf{x}_j, \mathbf{z}_j, \tilde{\mathbf{y}}_j, r_j)$
    end for
end for
```

A.2 Hyperparameters

Hyperparameter	Value
algorithm	DaPO (Yu et al., 2025)
training steps	100
epochs	10
train batch size	128
max prompt length	3072
max response length	16384
max extrapolation length	32768
learning rate	1e-6
PPO mini batch size	64
PPO micro batch size	64
clip ratio (low / high)	0.2 / 0.5
entropy coefficient	0.001
KL loss coefficient	0.001
KL loss type	low_var_kl
sampling temperature (train / val)	0.6 / 0.6
samples per prompt (train / val)	16 / 8
max batched tokens	32768

Table 2. Key training hyperparameters used in ***RLAD***.

B Additional Experimental Results

B.1 Abstraction on Diverse Text Classification

Dataset	Zero-shot	Best Abstraction	Average Abstraction
UCI Dry Bean	0.00	0.65	0.51
Wikipedia Proteinogenic Acid	0.22	0.78	0.58
UCI Student Performance	0.25	0.45	0.28
UCI Website Phishing	0.25	0.25	0.22
UCI Teaching Assistant Evaluation	0.25	0.45	0.33
UCI Contraceptive Method Choice	0.30	0.60	0.43
UCI Vertebral Column	0.30	0.75	0.64
UCI Shill Bidding	0.30	1.00	0.95
Kaggle Job Change	0.30	0.85	0.83
UCI Caesarian Section	0.38	0.75	0.64
Wikipedia Coin Face Value	0.40	1.00	0.88
UCI Wine	0.40	0.95	0.85
UCI Tic-Tac-Toe Endgame	0.40	0.80	0.42
Kaggle Campus Placement	0.40	0.85	0.72
Wikipedia Driving Championship Points	0.40	1.00	0.74
UCI Mammographic Mass	0.45	0.90	0.82
UCI Banknote Authentication	0.45	1.00	0.78
Kaggle Engineering Placement	0.50	0.85	0.79
RAFT One Stop English	0.50	0.40	0.36
LegalBench Function of Decision Section	0.54	0.72	0.61
Kaggle Entrepreneur Competency	0.55	0.65	0.58
UCI Indian Liver Patient	0.55	0.80	0.68
LegalBench International Citizenship Questions	0.56	0.74	0.63
LegalBench Abercrombie	0.56	0.80	0.67
Wikipedia Color Luminance	0.60	1.00	1.00
RAFT Twitter Hate Speech	0.60	0.90	0.76
Wikipedia Award Nomination Result	0.64	1.00	0.76
UCI Car Evaluation	0.65	0.75	0.64
Kaggle Water Potability	0.65	0.50	0.38
Kaggle Travel Insurance	0.65	0.70	0.59
UCI Internet Firewall	0.70	1.00	0.97
RAFT ADE Corpus	0.70	1.00	0.89
UCI Somerville Happiness Survey	0.70	0.80	0.68
UCI Mushroom	0.75	1.00	0.95
UCI Occupancy Detection	0.80	1.00	0.92
Kaggle Stroke Prediction	0.85	0.90	0.90
LegalBench Corporate Lobbying	0.88	0.94	0.88
Average	0.50	0.80	0.68

Table 3. Evaluation of abstractions on diverse collection of 37 domains. We sampled 10 abstractions by prompting `o4-mini`, and measure test set accuracy while prompting `GPT-4o-mini` with each abstraction. We report both the average performance of the 10 abstractions and the best abstraction. **We find that the average and best abstractions outperform standard prompting by 18.0% and 30.0% on average, respectively.**

B.2 *RLAD*'s w/ abs performance on AIME 2024 and HMMT 2025

In this section, we evaluate the performance of the base model (Qwen-3-1.7B), GRPO-enhanced model, and our proposed method ***RLAD*** on two math reasoning benchmarks: AIME 2024 and HMMT 2025. As shown in Table 4, our method achieves the best performance across both datasets.

It is important to note that ***RLAD*** is trained using access to abstractions, yet it also generalizes better even when evaluated without abstraction. This suggests that ***RLAD*** does not merely overfit to the abstraction format but instead learns to effectively leverage high-level procedural guidance, leading to better generalization on challenging reasoning benchmarks.

Approach	**AIME 2024**	**HMMT 2025**
Qwen-3-1.7B	48.54	22.50
+ GRPO	44.17	23.13
+ ***RLAD***	51.46	23.75

Table 4. ***RLAD*'s w/ abs performance on AIME 2024 and HMMT 2025.**

B.3 Design Choice Ablations

In this section, we run run some ablation experiments to better understand the contributions of individual components of ***RLAD*** in attaining good performance. In particular, we are interested in understanding the role of **(a)** inclusion of prompts that are not annotated with an abstraction, **(b)** reward masking on these prompts if they are included, and **(c)** training via a curriculum approach, following the protocol in Anonymous Author(s) (2025).

We present our results in Table 5. The first experiment we run focuses on understanding how important it is to include a small fraction of prompts with no abstractions in training of $\pi_\theta^{\mathrm{sol}}$. Observe in

Curriculum training refers to a staged training process where the model first learns from simpler problems and gradually transitions to harder examples. We borrow this idea from concurrent work Anonymous Author(s) (2025) (which we also attach in the supplementary material) as it showed that this approach led to better performance without any abstractions, for just direct math problem-solving. In contrast, non-curriculum training mixes problems of all difficulties throughout training. As shown in the table, when training with abstractions as well, curriculum training improves both average and best-case abstraction-conditioned performance (0.41 and 0.48 vs. 0.38 and 0.43).

Second, we explore whether **including no-abstraction prompts** during training helps the solution-generator pay attention to the abstractions. We find that including these abstractions minorly improves the average performance from 0.37 to 0.38, in isolation when curriculum is not utilized.

Lastly, we study the effect of masking the problem-solving reward on no-abstraction prompts. We apply **reward masking** to prevent updates that might cause the solution-generator to ignore abstractions altogether. Specifically, we zero out the advantage (i.e., no policy reward) for completions from no-abstraction prompts, while retaining the KL penalty to maintain regularization. This design discourages the model from over-optimizing on no-abstraction examples, which could otherwise lead it to bypass abstractions entirely, a shortcut that may yield improved performance on the training set but hinders generalization to test problems when abstractions are provided. Empirically, we find reward masking is helpful.

Approach	**Design Choice**			**AIME 2025**	
	curriculum training	including no-abstraction prompt	reward masking	w/ abs (avg)	w/ abs (best)
variant 1	✗	✓	✗	36.51	42.29
variant 2	✗	✗	-	37.08	42.50
variant 3	✗	✓	✓	37.50	43.33
RLAD	✓	✓	✓	**42.45**	**48.33**

Table 5. **Ablation of Design Choices in *RLAD*.** We isolate the effects of curriculum training, no-abstraction inclusion, and reward masking. The full method achieves the strongest performance under abstraction-conditioned evaluation.

C Qualitative Examples of Math Reasoning Abstractions

C.1 Prompt for Abstraction Classification

We prompt `GPT-4o-mini` with the following prompt template to classify each abstraction into one of four categories.

Post-hoc abstraction classifier prompt

```
You are a abstraction classifier.  You will be given a problem-solving heuristic
or abstraction used for mathematical reasoning.  Your task is to classify it into
exactly one of the following mutually exclusive categories, based on the primary
cognitive function the heuristic serves.

(A) Caution alert:  any abstraction that warns the reader to double-check a specific
aspect of their solution or to not take a specific approach to the problem.
(B) Productive launchpoint:  an early move or framing that opens up high-potential
trajectories.  Examples include clever reformulations or symmetries.
(C) Blind-follow trajectory:  a description of a repeatable, sequential path that
can be reliably followed to solve the problem.  Examples include plug-and-play
formulas that can be followed blindly, without insight.  Do not choose this is
further reasoning is required to solve the problem.
(D) Structural shortcut:  a conceptual move that collapses multiple graph paths into
a single jump via insight or abstraction.  This can include introducing invariants.
(E) Other:  a abstraction that does not fit into the above categories.
Give a 1-2 sentence explanation for your classification, and end your answer with
exactly one of:  (A), (B), (C), (D), or (E).

---
abstraction:
{abstraction}
```

C.2 Example for Each Abstraction Category

Below, we show examples of abstractions classified into the four categories above.

Examples of (A) Caution alert

```
<description>Always record forbidden values from denominators before and after
manipulation.  After solving the polynomial, discard any roots that make a
denominator zero or that do not satisfy the original equation, to avoid extraneous
solutions.</description>
<example>In the equation (x+2)/(2x{1) = x{3, 2x{1 cannot be zero (so x is not ½).
If solving yields x=½ or any root that makes any denominator zero, reject it.  Then
verify the accepted roots in the original equation.</example>

<description>Keep units consistent when moving between area and length or
summing lengths.  After extracting a length from an area (via square root),
ensure subsequent arithmetic stays in the same unit to avoid scaling errors.
</description>
<example>If a square's area is 10000 cm², its side is sqrt(10000) = 100 cm.  To
express in meters, convert 100 cm to 1 m.  All later distances computed with that
side length must be in meters to remain consistent.</example>
```

Examples of (B) Productive launchpoint

<description>Translate comparative statements into algebraic equations using the chosen variables. Phrases like \twice as many" or \one less than" correspond to multiplication or addition/subtraction expressions. This step captures the core relationship in a solvable form.</description>
<example>If the problem states \Group A has twice as many as Group B," write the equation x = 2y. For \Group B has three fewer than Group C," you would write y = z − 3.</example>

<description>Select one variable as a parameter (often setting it to 1 or keeping it symbolic) to express all other variables in terms of it. This reduces the number of independent symbols and streamlines substitutions.</description>
<example>Given p/q = 3 and r/q = 2, choose q as the base variable. Write p = 3q and r = 2q, so all expressions involving p and r can be handled through q alone.</example>

Examples of (C) Blind-follow trajectory

<description>Logarithms offer a streamlined way to compute floor-based digit counts: for y>0, the number of integer digits is floor(log10 y) + 1. Use this to handle arbitrary exponents without juggling large powers explicitly.</description>
<example>To count digits of y = x^7, compute d = floor(7 * log10 x) + 1. If x=2.5, then d = floor(7 * log10(2.5))+1 = 2+1 = 3 digits.</example>

<description>The mean of a set equals its total sum divided by its number of elements. Use this to move between sums and averages when counts or totals are known. It works because \average" is defined as that ratio.</description>
<example>Suppose a subset has k items with mean m. Then its total sum is S = k·m. Conversely, if you know the sum S and the count k, the mean is m = S/k. For instance, if 5 items average to 10, their total is 5×10 = 50, and if you later learn the total is 60 for 6 items, the new mean becomes 60/6 = 10.</example>

Examples of (D) Structural shortcut

<description>When the same distance appears in multiple geometric roles (e.g., as radius to a vertex and to a tangen t point), express it in different algebraic forms and equate them. Solving the resulting equation produces the unknown variable, which then gives the desired length.</description>
<example>If r is both the distance from O to a vertex (r = sqrt[x^2 + $(L/2)^2$]) and the distance from O to the tangent point (r = f(x)), set sqrt[x^2 + $(L/2)^2$] = f(x). Solving this equation for x and back-substituting determines r explicitly, closing the geometric problem with an algebraic solution.</example>

<description>Use the perimeter constraint a+b+c=P to eliminate one variable, e.g. set c=P−a−b, reducing the problem to two degrees of freedom. This simplification turns the three-variable Heron expression into a function of a and b alone, facilitating analysis or enumeration.</description>
<example>For a target perimeter P=10, one writes c=10−a−b. Substituting into Heron's formula yields A(a,b)=sqrt[5 * (5−a) * (5−b) * (a+b−5)], which is now a two-variable function to study instead of three.</example>