DARS: DYNAMIC ALIGNMENT REWARD STEERING FOR ROBUST ADAPTIVE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive reasoning models allocate inference-time computation to tasks of varying difficulty, enabling an efficiency–performance trade-off. Yet two key flaws exist: under-specified rewards that fail to penalize misaligned reasoning budgets, and an "elastic" failure mode in which models tend to revert to non-adaptive reasoning. To this end, we propose ***DARS*** (**D**ynamic **A**lignment **R**eward **S**teering), a novel Reinforcement Learning Finetuning (RLF) framework that combines two synergistic strategies—Minimal Sufficient Reasoning (MSR)-guided optimization and Fisher regularization—to improve the robustness of adaptive reasoning fine-tuning. Concretely, *MSR-guided optimization* begins with an oracle-calibrated difficulty estimator that assigns each instance to a difficulty tier. Building on this, *MSR-driven reward* explicitly penalizes both short "arrogant" reasoning and prolonged overthinking while promoting diverse reasoning paths. To stabilize RLF and mitigate elastic reversion, we further incorporate a *Fisher-informed regularizer* that constrains updates to high-importance parameters according to the Fisher information matrix. Experiments demonstrate that ***DARS*** prevents catastrophic collapse and achieves robust adaptivity across diverse benchmarks. On standard complex reasoning tasks like MATH, it substantially improves the accuracy–efficiency trade-off (+0.6% accuracy with 3.5% fewer tokens on 14B), and on extremely challenging benchmarks like AIME, DARS maintains a much narrower token fluctuation range under adversarial attacks (e.g., reducing the token changes span by over 30% on 14B compared to strong baselines), and effectively resists severe mode reversion.

1 INTRODUCTION

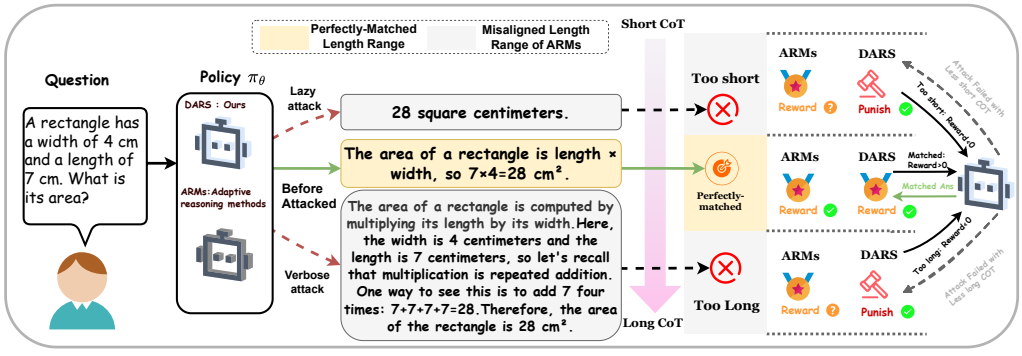

Figure 1: **Illustration of Length–Difficulty Misalignment in Adaptive Reasoning.** Using a simple geometric query, we contrast the resilience of DARS against standard adaptive models under adversarial pressure. While existing methods drift toward task-agnostic, uniform reasoning by accepting misaligned traces, DARS remains anchored to intrinsic complexity. By rewarding only a narrow, difficulty-matched band and penalizing both lazy and verbose responses, DARS ensures reasoning depth scales strictly with task difficulty.

Large reasoning models (LRMs) such as OpenAI o1 Jaech et al. (2026) and DeepSeek-R1 (Guo et al., 2025) achieve strong performance by scaling test-time compute through long Chain-of-Thought (CoT)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

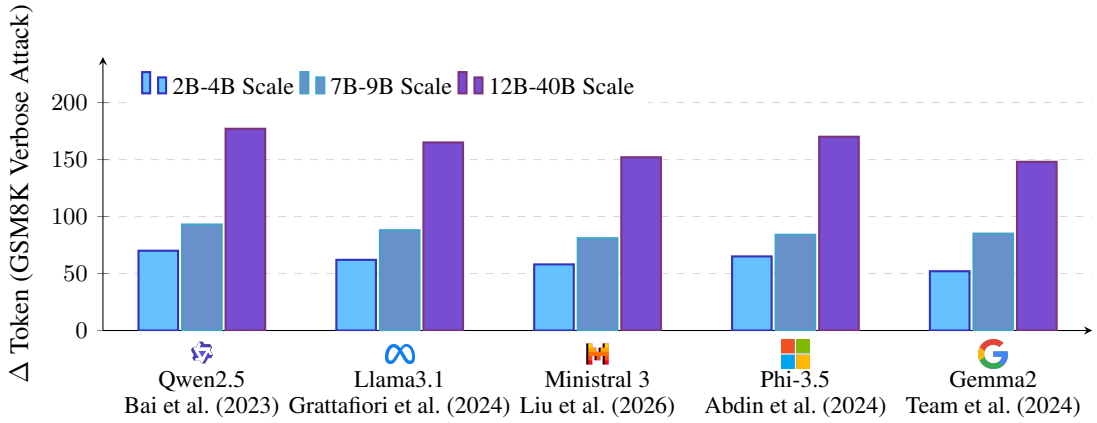

Figure 2: **Impact of Model Scale on Reasoning Stability.** Average Δ token usage on GSM8K under Verbose attacks, *i.e.*, extra tokens under attack vs. each model's GSM8K run without attacks. Instability grows with model scale.

reasoning (Wei et al., 2022; Zhu et al., 2025). However, applying long CoT uniformly is inefficient: even simple queries can elicit unnecessarily detailed reasoning traces, wasting computation and latency budget (Li et al., 2025). This motivates adaptive reasoning, where models allocate inference-time computation in proportion to task difficulty, usually by a routing mechanism that estimates query complexity and switches between direct answering and long-CoT regimes, enabling an explicit efficiency–performance trade-off (Hao et al., 2025a; Chen et al., 2025a).

Despite these advantages, we find adaptive reasoning to be surprisingly brittle. To probe this vulnerability, we propose three adversarial attacks—*Verbose*, *Lazy*, and *Random* (§4.1)—that deliberately decouple reasoning length from problem difficulty. Our findings reveal that even minimal fine-tuning on such misaligned data can trigger a complete failure of the learned routing policy: after only a few updates, the model reverts to a task-agnostic reasoning pattern, such as skipping essential steps on complex problems, even while headline accuracy remains deceptive stable (§4.4 and Fig. 5). We term this phenomenon *reasoning mode collapse*.

We argue that this fragility arises from two intertwined limitations in current RLF frameworks. First, prevailing rewards are often ***under-specified***, emphasizing final correctness while only weakly penalizing reasoning budgets that are misaligned with difficulty. This creates a broad region of reward indifference within the difficulty–length space (as shown in Fig. 1). Second, adaptive reasoning suffers from an ***elastic failure mode***, where models tend to drift back toward their base pretraining behaviors—a regression that intensifies as model scale increases. Consequently, routing becomes increasingly non-adaptive after minimal updates and degrades further under adversarial pressure. Fig. 2 quantifies this trend, demonstrating that attack-induced excess computation (Δ Token) increases sharply with model scale. Together, these factors render difficulty-aware routing highly unstable, particularly in larger frontier models.

To address these limitations, we propose **D**ynamic **A**lignment **R**eward **S**teering (***DARS***), an RL framework that enhances the robustness of adaptive reasoning by jointly aligning task incentives and stabilizing model updates. ***DARS*** integrates two complementary strategies. First, we employ *Minimal Sufficient Reasoning (MSR)*-guided optimization, which utilizes an oracle-calibrated difficulty estimator to assign per-instance complexity and anchor an appropriate reasoning budget (§ 3.2). Building on this calibration, we introduce an *MSR-driven reward* with a time-aware curriculum that explicitly penalizes both "arrogant" under-reasoning on hard problems and "verbose" over-thinking on simple ones, while simultaneously encouraging diverse reasoning trajectories (§ 3.3). Second, to mitigate elastic reversion during fine-tuning, we incorporate a Fisher-informed regularizer. By constraining updates to high-importance parameters—as identified by the Fisher Information Matrix—this regularizer preserves stable difficulty-aware routing during subsequent training (§ 3.4).

In summary, our main contributions are:

- We identify a fundamental robustness gap in existing adaptive reasoning finetuning, formalize how reward under-specification and elastic failure destabilize adaptive reasoning finetuning.

- We then introduce ***DARS***, a principled training framework that restores and preserves difficulty-aware reasoning with *MSR*-driven with a time-aware curriculum and *Fisher-informed regularizer*.
- Extensive experiments on Qwen2.5-3B/7B/14B show that DARS improves accuracy to 3.0% under clean conditions and maintains reasoning length consistency with less than 10% variation under adversarial attacks, demonstrating robust and efficient adaptive reasoning.

2 Related Work

Adaptive Reasoning Methods Chain-of-Thought (CoT) reasoning Wei et al. (2022); Yao et al. (2025); Prasad et al. (2024) empowers Large Language Models to tackle sophisticated problem-solving via structured intermediate steps. Recent evidence suggests that scaling test-time computation can effectively augment model capabilities, often yielding gains that rival or surpass those from traditional parameter scaling Snell et al. (2024); Bi et al. (2024); Yang et al. (2025). However, the uniform application of expansive reasoning chains to every query incurs substantial computational waste. To mitigate this, current research has converged on dynamic resource allocation strategies Xu et al. (2025); Yang et al. (2025), primarily through *routing-based architectures* that toggle reasoning modes based on instance difficulty Huang et al. (2025a); Xiangqi et al. (2025); Fang et al. (2025), or *self-adaptive policies* that internalize early-exit mechanisms directly within the model's generation logic Wu et al. (2025a); Jiajie et al. (2025); Junjie et al. (2025). While these systems typically leverage preference-based Reinforcement Learning (RL) objectives such as DPO Rafailov et al. (2023); Chen et al. (2025b), they remain susceptible to reward hacking and goal misgeneralization stemming from underspecified reward signals Jinyan and Claire (2025); Huang et al. (2025b).

Model Fragility and Robustness Existing work underscores the inherent brittleness of aligned models, noting that safety guardrails can be compromised by benign fine-tuning Wan et al. (2023); Hao et al. (2023); Luo et al. (2024); Jin and Lee (2025) or readily overwritten by minimal updates due to the effects of alignment elasticity Ji et al. (2025); Wu et al. (2024); Tan et al. (2025). Furthermore, robustness often degrades under adversarial prompts, instruction perturbations, or distribution shifts Zhu et al. (2023); Hao et al. (2025b); Kun et al. (2025). This fragility is echoed in continual-learning studies on catastrophic forgetting in LLMs Tian et al. (2023); Zike et al. (2025), which have inspired mitigations such as parameter-efficient constraints Wang et al. (2024); Xiao et al. (2023)and robust alignment techniques Pravin et al. (2024); Luo et al. (2024) to preserve desirable behaviors. Diverging from prior work focused on safety or general knowledge, we investigate the robustness of *difficulty-aware routing*—an efficiency-critical behavior prone to "reasoning mode collapse" during fine-tuning. In this paper, we propose *DARS* to stabilize this routing policy by combining Fisher-information-based parameter preservation with an explicit difficulty–length-aware reward mechanism, ensuring resilient compute-allocation logic against distribution shifts.

3 Method

3.1 Overview

DARS forms a robust framework for dynamic reasoning (Fig. 3). Following Supervised Fine-Tuning (SFT), we compute the Fisher Information Matrix (FIM) to estimate parameter importance before train via Group Relative Policy Optimization (GRPO) Shao et al. (2024), with three mechanisms: oracle-calibrated difficulty stratification, MSR-guided reward shaping, and Fisher-informed regularization.

Problem Formulation We consider a dataset $\mathcal{D} = \{(q, a^*)\}$, where q is the question and a^* is the ground-truth answer. A reasoning model is parameterized as a policy π_θ that generates a reasoning trace r and a final answer a. To regulate inference effort according to task complexity, we employ a stronger oracle model M_{oracle} to assign each query q to a difficulty tier $k(q) \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$. For each tier k, we estimate a reference reasoning length L_{ref}^k from successful oracle reasoning traces, providing a difficulty-conditioned reference for reasoning effort. Our objective jointly optimizes answer correctness, reasoning-format diversity, and difficulty-aligned efficiency:

$$R_{\text{total}} = R_{\text{correct}} + R_{\text{div}} + R_{\text{MSR}}, \tag{1}$$

where $R_{\text{correct}}(a) = \mathbb{1}[a = a^*]$ is a binary correctness reward. The MSR term uses a correctness-dependent length target, L_{GT} for correct responses and $L_{\text{ref}}^{k(q)}$ otherwise, penalizing both excessive verbosity (*inefficiency*) and premature termination (*arrogance*); details are provided in Sec. 3.3.

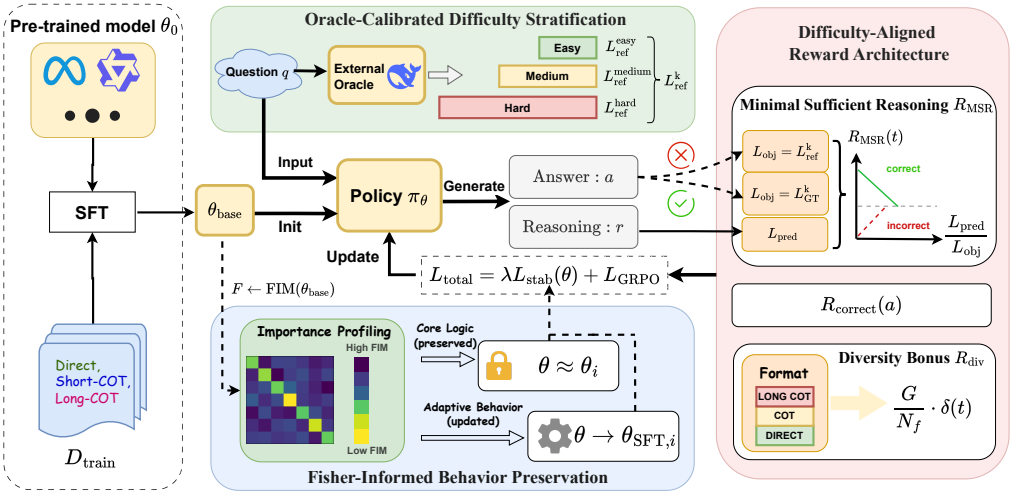

Figure 3: **Overview of Dynamic Alignment Reward Steering (DARS).** SFT turns a pre-trained model into a base policy θ_{base} with a Fisher information profile, and GRPO with an external Oracle then jointly optimizes correctness and MSR-based efficiency (*Modules 1–2*), while Fisher-informed regularization stabilizes high-importance parameters to preserve core reasoning behavior (*Module 3*).

3.2 Oracle-Calibrated Difficulty Stratification

A key challenge in adaptive reasoning is defining an objective notion of *sufficient* reasoning effort. Policy-derived difficulty estimates may be miscalibrated, leading to complexity underestimation. We therefore introduce *oracle-calibrated stratification*, which provides an external standard for difficulty-budget alignment.

External Difficulty Partitioning We obtain external difficulty estimates to avoid bias from training dynamics. Concretely, a high-capacity oracle model (e.g., GPT-4o) assesses the *reasoning complexity* of each query q and assigns it to a discrete tier $k \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$. This yields a fixed partition of the training set $\mathcal{D} = \mathcal{D}_{\text{easy}} \cup \mathcal{D}_{\text{medium}} \cup \mathcal{D}_{\text{hard}}$, serving as an immutable reference throughout RLF. By anchoring difficulty labels to an external evaluator, we eliminate difficulty drift, where the student model's evolving behavior shifts the effective boundary between tiers.

Reference Length Anchoring Tier labels alone are insufficient for precise reward shaping; we therefore introduce a quantitative surrogate for the *minimum computation likely required* for a given difficulty level. For each tier k, we establish a *Reference Length* L_{ref}^{k} as principled efficiency target. Let r_{oracle}^{*} denote a oracle generated reasoning trace (conditioned on query q) that attains a correct final answer. The reference length for tier k is defined as the expected length of such successful oracle traces under this tier:

$$L_{\text{ref}}^{k} = \frac{1}{|\mathcal{D}_k|} \sum_{q \in \mathcal{D}_k} \text{len}(r_{\text{oracle}}^{*} | q). \tag{2}$$

This anchor L_{ref}^{k} distinguishes efficient reasoning from under-computation. Concretely, for instances in tier k, reasoning substantially shorter than the anchor (*e.g.*, $|r| < \tau \cdot L_{\text{ref}}^{k}$) is penalized as *difficulty-misaligned*, ensuring that token reductions do not compromise semantic rigor.

3.3 MSR-Guided RL with Curriculum

Standard RL fine-tuning may cause premature convergence to inflexible reasoning patterns. We therefore formulate the optimization as a dynamic interplay between correctness, efficiency, and diversity. We introduce *Minimal Sufficient Reasoning (MSR)*, which orchestrates a staged learning process: prioritizing correctness in early training before progressively enforcing reasoning efficiency.

The Curriculum-MSR Objective (R_{MSR}) We formalize the efficiency reward as a time-varying function that differentially regulates reasoning length based on answer correctness: the model is encouraged to produce shorter reasoning when the prediction is correct, while being penalized for overconfident short answers when incorrect, thus encouraging longer, more cautious reasoning traces

in failure cases. The reward at training step t is defined as:

$$R_{\text{MSR}}(t) = \sigma(t) \cdot \begin{cases} \alpha(1-\rho), & a = a^*; \\ -\beta(1-\frac{\rho}{\tau}), & a \neq a^*, \rho < \tau; \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\rho = L_{\text{pred}}/L_{\text{obj}}$ denotes the normalized reasoning length with respect to a reference target. The curriculum factor $\sigma(t)$ controls activation of this reward during training. The definition of the reference target L_{obj} is provided next.

Context-Aware Reference Standards (L_{obj}) Unlike static length penalties, DARS employs a dual-track standard for L_{obj} that adapts dynamically based on outcome correctness. Specifically, for correct responses ($a = a^*$), we set $L_{\text{obj}} = L_{\text{GT}}$ (Ground Truth Length), treating the human reference as an efficiency ceiling to incentivize the model to achieve correctness with greater conciseness than the reference trace. Conversely, for incorrect responses ($a \neq a^*$), the objective shifts to a *minimum effort baseline* by setting $L_{\text{obj}} = L^k_{\text{ref}}$ (the Oracle-derived anchor for difficulty tier k). In this regime, the penalty term $-\beta$ (triggered when $\rho < \tau$) explicitly targets *arrogance*, discouraging speculative guesses that fall significantly below the necessary complexity threshold ($\tau = 0.4$, see Appendix C).

Temporal Curriculum Modulation ($\sigma(t)$) To prevent premature optimization of length at the expense of accuracy, we introduce a warm-up factor $\sigma(t) = (t/T)^\gamma$. This term suppresses the magnitude of R_{MSR} early in training, ensuring the policy first stabilizes on reasoning correctness before the gradient signal shifts toward optimizing brevity.

Diversity Bonus (R_{div}) To encourage diversity in reasoning formats, we introduce a diversity bonus that promotes balanced usage of different output structures, including {Direct, CoT, Long-CoT}. This term incentivizes the model to explore and maintain multiple formats during training. Formally, for an output belonging to format f at training step t, the diversity reward is defined as:

$$R_{\text{div}}(f, t) = R_{\text{correct}} \cdot \frac{G}{N_f} \cdot \delta(t), \tag{4}$$

where $R_{\text{correct}} = \mathbb{1}[a = a^*]$ ensures that the diversity bonus is only applied to correctly answered samples, G denotes the group size, i.e., the number of outputs sampled by the model for each question in the traditional GRPO procedure, N_f is the number of samples in the group using format f, and $\delta(t) = \frac{1}{2}\left(1 + \cos\frac{\pi t}{T}\right)$ is a cosine decay schedule that assigns higher weight to format diversity in early training and gradually reduces its influence over time.

The final reward combines correctness, format diversity, and reasoning-length alignment:

$$R^{(t)}_{\text{total}} = R_{\text{correct}} + R_{\text{div}}(f, t) + R_{\text{MSR}}(t). \tag{5}$$

3.4 Fisher-Informed Parameter Stabilization

While the MSR objective incentivizes difficulty-aligned reasoning, the resulting policy remains vulnerable to **elasticity**: the tendency of fine-tuned models to revert back to original behavior. Without additional stabilization, this snap-back effect can erode the RLF process, particularly under the stress of adversarial updates or distribution shifts.

The Mechanics of Collapse We conceptualize the RLF process as a tug-of-war between two opposing forces: the reward-driven gradient $\|\nabla_\theta \mathcal{L}_{\text{RL}}\|$ pushing the model toward the new adaptive strategy, and the elastic stiffness k_{elastic} pulling parameters back to their original manifold. Scaling laws suggest this stiffness grows with model size N and pre-training data volume $|\mathcal{D}_{\text{pre}}|$, theoretically modeled as $k_{\text{elastic}} \propto N^\alpha \cdot |\mathcal{D}_{\text{pre}}|^\beta$. Consequently, successful adaptation requires the update signal to overcome this inertia. We define the *Collapse Ratio* η to quantify this risk:

$$\eta \propto \frac{\|\nabla_\theta \mathcal{L}_{\text{RL}}\|}{k_{\text{elastic}}} \approx \frac{\|\nabla_\theta \mathcal{L}_{\text{RL}}\|}{N^\alpha \cdot |\mathcal{D}_{\text{pre}}|^\beta}. \tag{6}$$

If $\eta \to 0$, the model experiences *Elastic Rebound*, collapsing back to difficulty-agnostic (lazy or verbose) patterns.

Fisher-Informed Regularization To mitigate elastic rebound without hindering the acquisition of new routing capabilities, we adopt a *selective* stabilization strategy based on the Fisher Information

Matrix (FIM). Instead of imposing a uniform constraint over all trainable parameters, FIM provides a selective stabilization mechanism by identifying trainable parameters that are important for preserving the post-SFT *Adaptive Behavior*. After SFT, we estimate the diagonal FIM F on a representative calibration set $\mathcal{D}_{\text{cal}}$:

$$F_i \approx \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{(q,r)\in\mathcal{D}_{\text{cal}}} \left(\partial_{\theta_i} \log \pi_{\theta_{\text{base}}}(r \mid q)\right)^2 , \tag{7}$$

where i indexes the individual model parameters, θ are the current trainable parameters, θ_{base} are the frozen reference parameters, and $\pi_{\theta_{\text{SFT}}}$ denotes the post-SFT policy. Intuitively, a large F_i indicates that small perturbations to θ_i would substantially affect the likelihood of the reference model's outputs, suggesting that the parameter encodes high-utility representations that should be protected.

Inspired by Elastic Weight Consolidation (EWC) Kirkpatrick et al. (2017), we introduce a Fisher-weighted stabilization term to preserve adaptive reasoning behaviors during RLF. Unlike EWC, which addresses catastrophic forgetting in continual learning, our formulation constrains policy drift by penalizing deviations of Fisher-important parameters from the SFT reference:

$$\mathcal{L}_{\text{stab}}(\theta) = \frac{1}{2} \sum_i F_i(\theta_i - \theta_{\text{SFT},i})^2. \tag{8}$$

The overall optimization objective combines the GRPO objective with the stabilization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \lambda \mathcal{L}_{\text{stab}}(\theta), \tag{9}$$

where λ controls the trade-off between policy adaptation and behavior preservation.

4 Experiments

4.1 Experimental Details

Datasets We construct a composite training corpus $\mathcal{D}_{\text{train}}$ of approximately 25,000 samples from *MATH* Hendrycks et al. (2021), *GSM8K* Cobbe et al. (2021), and *CommonsenseQA* Talmor et al. (2019), spanning diverse challenges from basic commonsense to complex mathematical derivation. To support adaptive reasoning, the corpus contains three reasoning granularities—***Direct Answer*** (`<DIRECT_ANSWER>`), ***Standard CoT*** (`<COT>`), and ***Long CoT*** (`<LONG_COT>`)—with each training instance represented in one of these formats. This multi-granular mixture exposes the model to reasoning trajectories of different depths, encouraging it to condition reasoning effort on task complexity rather than adopting a fixed verbosity preference. For evaluation, we utilize *MATH-Test*, *GSM8K-Test*, *MMLU-STEM*, and *AIME* (90 problems from AIME 2024–2026). MATH-Test and GSM8K-Test provide held-out in-domain evaluation, while MMLU-STEM and AIME assess generalization to broader and more challenging reasoning tasks.

Adversarial Evaluation Scenarios To evaluate robustness against length–difficulty mismatch during continued fine-tuning, we construct three adversarial datasets that intentionally decouple problem difficulty from response length (see Appendix B for construction protocols and statistics): (1) ***Verbose Attack*** pairs simple queries with excessively long reasoning traces to induce over-reasoning; (2) ***Lazy Attack*** pairs difficult problems with truncated short answers to induce under-reasoning; (3) ***Random Attack*** uses out-of-distribution creative-writing queries with variable-length outputs and no required reasoning structure, testing whether reasoning modes are spuriously activated when logical reasoning is unnecessary.

Metrics To comprehensively assess our model, we utilize four key metrics targeting reasoning performance, computational efficiency, policy stability, and adaptivity. We evaluate ***Accuracy (Acc)*** through exact answer matching, measure computational cost using ***Token Usage***, and quantify policy shifts induced by fine-tuning with ***Policy KL Divergence***. Furthermore, to quantify dynamic compute allocation, we introduce the ***Adaptivity Gain (AG)*** metric, which measures the alignment between estimated problem difficulty and reasoning length. AG is defined as the Spearman rank correlation:

$$\text{AG} = \rho_{\text{Spearman}}\big(d(q),\, \text{len}(r)\big), \tag{10}$$

where $d(q)$ represents the estimated difficulty score of query q (e.g., from dataset stratification or baseline error rates), and $\text{len}(r)$ is the number of generated reasoning tokens. A higher AG value indicates stronger difficulty–length coupling, where harder queries tend to receive longer reasoning

Table 1: **Main results across Qwen2.5 scales under clean and adversarial conditions.** We report Accuracy (Acc), average token usage (Tok), and token change (ΔTok). DARS generally preserves accuracy while reducing attack-induced token deviations across model scales.

Method	Qwen2.5-3B								Qwen2.5-7B								Qwen2.5-14B							
	Clean		Verbose		Lazy		Random		Clean		Verbose		Lazy		Random		Clean		Verbose		Lazy		Random	
	Acc	Tok	Acc	ΔTok	Acc	ΔTok	Acc	ΔTok	Acc	Tok	Acc	ΔTok	Acc	ΔTok	Acc	ΔTok	Acc	Tok	Acc	ΔTok	Acc	ΔTok	Acc	ΔTok
MATH-Test																								
Coldstart-SFT	61.4	420	62.2	+400	61.7	-270	63.1	+266	68.2	436	66.3	+571	62.3	-134	69.8	-88	75.4	499	80.3	+333	76.5	-190	74.9	+61
ARM	66.7	677	68.1	**+10**	60.9	-72	65.3	+7	70.3	570	70.0	**+34**	63.2	**-8**	72.1	+11	78.9	456	80.9	+90	77.3	-108	80.1	+28
AdaCtrl	68.9	668	69.5	+92	56.2	-70	62.3	-23	68.7	529	63.8	+154	63.9	-52	70.3	+31	74.8	448	75.1	+235	69.4	-27	73.6	+121
DARS	**69.5**	640	**70.2**	+11	**63.4**	**-4**	**66.0**	**+3**	**71.0**	560	**72.1**	+42	**66.4**	-12	**73.0**	**-4**	**79.5**	440	**81.5**	**+14**	**77.4**	**-14**	**81.0**	**-8**
GSM8K																								
Coldstart-SFT	84.5	353	90.5	+70	89.2	-42	90.1	+11	89.1	365	87.3	+93	91.3	-92	88.4	**-4**	92.4	385	91.5	+177	93.6	-81	93.8	-10
ARM	88.5	372	89.2	+46	91.6	-4	91.2	+8	91.2	323	94.5	+48	95.3	-26	92.9	+22	91.6	327	97.8	+216	94.4	**-7**	96.5	+158
AdaCtrl	87.6	366	92.3	+39	85.4	-9	83.7	+5	92.5	310	96.2	+49	**96.8**	-9	94.6	+16	95.6	324	97.5	+89	95.3	-35	95.1	+43
DARS	**89.0**	360	**93.0**	**+11**	**92.0**	**-2**	**92.0**	**+3**	**93.0**	305	**97.0**	**+16**	95.2	**-5**	**95.0**	-5	**96.0**	310	**98.4**	**+8**	**96.0**	-11	**96.9**	**-4**
MMLU-STEM																								
Coldstart-SFT	67.0	140	66.4	+210	65.1	-60	66.7	+15	71.8	156	71.0	+204	69.4	-76	71.4	+14	74.5	170	73.8	+210	72.4	-85	74.1	+15
ARM	67.2	92	64.0	+318	62.8	-11	66.0	+21	72.1	98	68.8	+322	67.4	-13	70.6	+19	74.9	105	71.4	+335	70.1	-17	73.2	+24
AdaCtrl	67.6	90	64.4	+305	63.2	-17	66.3	+18	72.5	94	69.2	+311	67.9	-11	70.9	+17	75.2	102	71.8	+323	70.5	-16	73.5	+20
DARS	**68.0**	85	**66.9**	**+185**	**65.9**	**+10**	**67.5**	**+10**	**72.9**	89	**71.6**	**+196**	**70.3**	**-9**	**72.0**	**+13**	**75.6**	96	**74.1**	**+204**	**72.9**	**-10**	**74.8**	**+12**
AIME																								
Coldstart-SFT	0.0	1278	0.0	+977	0.0	-91	0.0	-108	1.1	1034	1.1	+457	0.0	-763	0.0	+225	2.2	981	3.3	+532	0.0	-249	0.0	-98
ARM	5.6	3045	5.6	+829	0.0	-302	0.0	+275	16.7	3189	17.8	**+205**	11.1	-655	**18.9**	-102	20.0	3989	20.0	+455	**15.6**	-632	**20.0**	+131
AdaCtrl	2.2	3498	3.3	+734	0.0	-413	**1.1**	**-67**	**18.9**	3265	**20.0**	+356	14.4	-575	11.1	+108	**22.2**	4088	**21.1**	+756	13.4	-697	17.8	-312
DARS	**6.7**	3024	**6.7**	**+446**	**2.2**	**-197**	**1.1**	+123	**18.9**	3012	**20.0**	+217	**15.6**	**-506**	**18.9**	**-73**	20.0	3886	**21.1**	**+447**	**15.6**	**-536**	17.8	**-56**

traces while easier queries remain concise. AG measures adaptive allocation behavior rather than reasoning correctness, and is therefore interpreted jointly with accuracy and token usage.

Implementation details Our DARS model is based on the Qwen2.5 series Bai et al. (2023) and employs a two-stage training pipeline. We provide the full configuration of hyperparameters and hardware setups in Appendix A. First, a supervised fine-tuning (SFT) stage initializes the model using reasoning control tokens and LoRA adapters. Second, the model undergoes GRPO-based Shao et al. (2024) alignment starting from the SFT checkpoint, optimized by a reward function combining correctness, format validity, diversity scaling, and the MSR objective, together with a Fisher-based preservation constraint. We employ QLoRA for models up to 14B to ensure memory efficiency.

Baselines We benchmark ***DARS*** against three competitive baselines : (1) ***Coldstart-SFT*** Wei et al. (2025), the supervised fine-tuned model providing our training initialization, used here to quantify the net performance gains attributed to reinforcement learning; (2) ***ARM*** Wu et al. (2025b), a adaptive reasoning model optimized via Ada-GRPO; and (3) ***AdaCtrl*** Huang et al. (2025a), a difficulty-aware framework that utilizes explicit control tokens to manage reasoning length. Collectively, these comparisons evaluate whether DARS improves adaptive reasoning stability and compute-accuracy trade-offs under length–difficulty mismatch.

4.2 Main Results

We evaluate DARS against baseline methods across the adversarial scenarios defined in §4.1. Quantitative results in Tab. 1 show that DARS generally preserves accuracy while reducing token deviations under length–difficulty mismatch: on *MATH-Test*, the Qwen2.5-14B variant of *DARS* resists the tendency to "under-reason" under *Lazy* attacks, achieving 77.4% accuracy, higher than AdaCtrl (69.4%). Under *Verbose* attacks on *GSM8K*, *DARS* also effectively avoids redundant over-reasoning on simple tasks, reaching 98.4% accuracy with negligible token inflation (+8 tokens *vs.* Clean). Similar trends are also observed on *MMLU-STEM*.

Furthermore, on the challenging *AIME* benchmark, *DARS* achieves competitive performance across adversarial settings while exhibiting reduced token usage deviation in several challenging cases. For instance, the 14B model matches AdaCtrl in accuracy (21.1%) under *Verbose* settings with substantially less token inflation (+447 *vs.* +756). Conversely, under *Lazy* constraints, *DARS* maintains higher accuracy (15.6% *vs.* 13.4%) with less drastic sequence truncation (-536 *vs.* -697). By reducing token fluctuations under adversarial shifts, *DARS* better preserves difficulty–length alignment, demonstrating improved robustness to length–difficulty mismatch. Additional token-length distribution analyses are provided in Fig.4, where DARS exhibits smaller distributional shifts under adversarial fine-tuning.

4.3 Ablation Study

Component Ablation. Tab. 2(a) examines the contribution of each DARS component to adaptive reasoning and token stability. Removing MSR causes the largest degradation in attacked adaptivity,

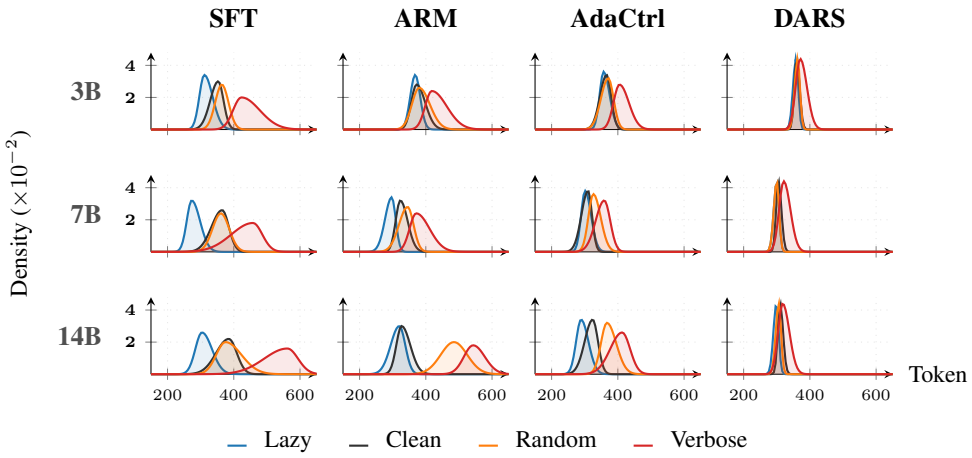

Figure 4: **Token-length distributions on GSM8K across model scales and methods.** While baselines exhibit increasing distributional divergence and length-bias sensitivity from 3B to 14B, DARS maintains remarkably stable, overlapping profiles, indicating stronger robustness to adversarial length perturbations.

Table 2: **Ablation analysis of DARS.** (a) Component ablation on Qwen2.5-7B/GSM8K. (b) Controlled comparison on Qwen2.5-7B/MATH-Test under matched training and evaluation settings.

(a) Component ablation

Variant	Clean AG↑	Attacked AG↑	Clean Tok↓	ΔTok↓
Baseline	0.34	0.18	365	+93
w/o MSR	0.45	0.22	378	+43
w/o Oracle	0.69	0.57	353	+39
w/o Fisher	0.70	0.55	337	+66
DARS	**0.78**	**0.65**	**305**	**+16**

(b) Controlled comparison

Method	Clean Acc.↑	Verbose |ΔTok| ↓	Lazy |ΔTok| ↓	AG (C/V/L)↑
AdaCtrl	68.7	154	52	.75/.54/.38
AdaCtrl + Fisher	66.0	98	85	.75/.59/.43
DARS	**71.0**	**42**	**12**	**.78/.65/.56**

reducing AG to 0.22, while removing the oracle-calibrated reference also weakens difficulty-aware allocation. Removing Fisher regularization has a smaller effect on clean adaptivity but substantially increases attack-induced token usage to +66 tokens. In comparison, full DARS maintains the highest attacked AG (0.65) while limiting token inflation to +16 tokens. These results suggest that MSR and oracle calibration primarily establish difficulty-aware reasoning behavior, whereas Fisher regularization helps preserve this behavior under adversarial fine-tuning.

Controlled Comparison with AdaCtrl+Fisher. To isolate the contribution of Fisher regularization, Tab. 2(b) compares AdaCtrl, AdaCtrl+Fisher, and DARS under matched training and evaluation settings. Adding the same Fisher regularizer to AdaCtrl reduces Verbose token deviation from 154 to 98 and improves attacked AG, but lowers clean accuracy and increases the Lazy deviation from 52 to 85. DARS instead achieves higher clean accuracy (71.0%), smaller token deviations under both Verbose and Lazy settings (42/12), and stronger AG across conditions. This controlled comparison indicates that Fisher regularization alone does not account for the gains of DARS, supporting the complementary roles of difficulty-aligned reward shaping and parameter preservation.

4.4 Robustness Analysis

Robustness of Adaptivity Gain To further verify the robustness of microscopic alignment between problem difficulty and reasoning length, we analyze the impact of adversarial attacks on Adaptivity Gain (AG) across Qwen2.5 scales (3B, 7B, 14B), as illustrated in Fig. 5. Our detailed analysis reveals several key insights into the model's adaptive stability. First, DARS demonstrates superior resilience and maintains significantly higher absolute AG values compared to ARM across all scales, signifying a more precise coupling of computational cost to task complexity. Second, we observe a critical scaling vulnerability in the baseline: while ARM performs reasonably at 3B, it suffers from increasingly severe degradation as model capacity grows, leading to a catastrophic collapse at the 14B scale under Lazy attacks, where AG plummets from 0.83 to 0.29. In stark contrast, DARS

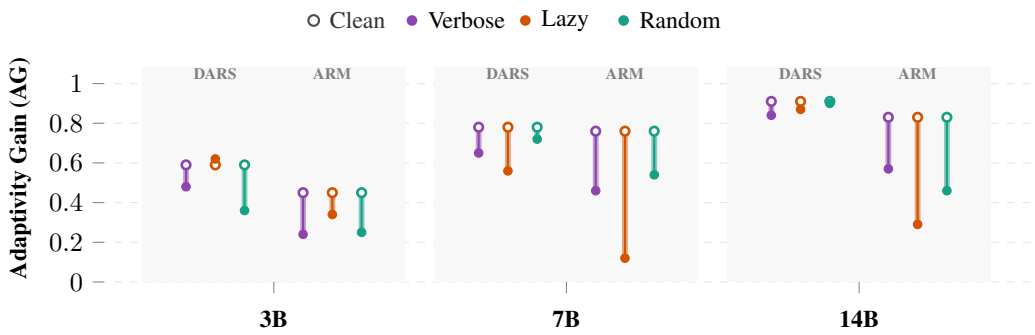

Figure 5: **Comparison of reasoning stability across model scales.** The plots show the drop in Adaptivity Gain (AG) between clean (empty circles) and attacked (solid dots) scenarios for 3B, 7B, and 14B models. DARS (left) maintains consistently high AG with minimal degradation, whereas ARM (right) suffers severe policy collapse that exacerbates with increasing model scale.

Table 3: **Policy KL Divergence between clean and attacked models.** Higher values indicate a greater distribution shift under adversarial pressure. DARS consistently minimizes policy deviation across all scales, significantly outperforming ARM and AdaCtrl in maintaining stable routing logic.

Attack	Qwen2.5-3B				Qwen2.5-7B				Qwen2.5-14B			
	Coldstart-SFT	ARM	AdaCtrl	DARS	Coldstart-SFT	ARM	AdaCtrl	DARS	Coldstart-SFT	ARM	AdaCtrl	DARS
Verbose	0.047	0.069	0.033	**0.030**	0.078	0.092	0.035	**0.032**	0.098	0.127	**0.062**	0.067
Lazy	0.077	0.075	0.048	**0.042**	0.088	0.109	0.050	**0.043**	0.156	0.167	0.152	**0.078**
Random	0.028	0.012	0.022	**0.009**	0.031	0.027	0.023	**0.021**	**0.045**	0.089	0.078	0.047

14B remains remarkably robust, maintaining a high AG of ≈ 0.87 across all attack types (Verbose, Lazy, and Random). This suggests that DARS effectively mitigates the *alignment elasticity* common in larger models, preventing them from succumbing to superficial length biases. These findings conclude that DARS provides a scalable and robust solution for difficulty-aware reasoning, ensuring that larger models translate their increased capacity into intelligent compute allocation rather than fragile, noise-sensitive behaviors.

Resistance to Distributional Shift Tab. 3 provides a detailed quantitative analysis of the mechanisms underlying DARS's robustness. By measuring the Policy KL Divergence between clean and attacked checkpoints, we observe that DARS consistently minimizes policy shift across nearly all scales and attack types, with its advantages becoming more pronounced in the 7B and 14B regimes. Notably, under the 14B Lazy attack, DARS maintains a divergence of only 0.078, whereas ARM (0.167) and AdaCtrl (0.152) exhibit significantly larger distributional shifts. These findings reveal that DARS does not rely on trivial parameter freezing or passive rigidity to achieve stability. Instead, it actively anchors high-information parameters essential for adaptive routing, enabling the model to resist external perturbations while preserving its intrinsic difficulty–length alignment. DARS's superior resilience is a product of active policy preservation, ensuring a robust defense against "reasoning mode collapse" that scales effectively with model capacity.

5 Conclusion

In this paper, we investigate the robustness of adaptive reasoning in LLMs, revealing that difficulty-aware routing is highly susceptible to reasoning mode collapse during continued fine-tuning. We trace this fragility to reward under-specification, where a lack of length-based differentiation allows suboptimal lazy or verbose policies to emerge at low cost. To address this, we propose ***DARS***, a framework that stabilizes routing by sharpening the alignment between query difficulty and reasoning length while anchoring high-information parameters. Across various model scales and adversarial attacks, ***DARS*** consistently maintains high adaptivity and accuracy–efficiency trade-offs. However, ***DARS*** remains limited on ambiguous task complexities or highly personalized user requests. While future work is needed to reduce the reliance on external difficulty signals, our findings establish a robust foundation for building resilient, compute-efficient autonomous agents.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, including grammar checking, improving readability, and polishing the presentation of the manuscript. We have not used generative AI tools for generating research ideas, formulating theoretical claims, designing experimental methodologies, producing experimental results, or writing mathematical proofs, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for recommended disclosure tasks, including assisting with software development and debugging. We have reviewed all AI-assisted work and verified the correctness of the final manuscript, implementations, and experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, et al. Openai o1 system card, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting sa reasoning in large language models. In *Information Processing Systems 35*, 2022.

Yekun Zhu, Guang Chen, and Chengjun Mao. Think in blocks: Adaptive reasoning from direct response to deep reasoning. *arXiv preprint arXiv:2508.15507*, 2025.

Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, Kai Jia, and Zhifang Sui. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. *arXiv preprint arXiv:2505.11274*, 2025.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025a.

Qiguang Chen, Dengyun Peng, Jinhao Liu, HuiKang Su, Jiannan Guan, Libo Qin, and Wanxiang Che. Aware first, think less: Dynamic boundary self-awareness drives extreme reasoning efficiency in large language models. *arXiv preprint arXiv:2508.11582*, 2025a.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models, 2024.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. *arXiv preprint arXiv:2305.10601*, 2025.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. Adapt: As-needed decomposition and planning with language models. In *Findings of the Association for Computational Linguistics: NAACL*, pages 4226–4252, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.

Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. Scalable chain of thoughts via elastic reasoning. *arXiv preprint arXiv:2505.05315*, 2025.

Shijue Huang, Hongru Wang, Wanjun Zhong, Zhaochen Su, Jiazhan Feng, Bowen Cao, and Yi R Fung. Adactrl: Towards adaptive and controllable reasoning via difficulty-aware budgeting. *arXiv preprint arXiv:2505.18822*, 2025a.

Wang Xiangqi, Huang Yue, Wang Yanbo, Luo Xiaonan, Guo Kehan, Zhou Yujun, and Zhang Xiangliang. Adareasoner: Adaptive reasoning enables more flexible thinking. *arXiv preprint arXiv:2505.17312*, 2025.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: Llm learns when to think. *arXiv preprint arXiv:2505.13379*, 2025.

Xingyu Wu, Yuchen Yan, Shangke Lyu, Linjuan Wu, Yiwen Qiu, Yongliang Shen, Weiming Lu, Jian Shao, Jun Xiao, and Yueting Zhuang. Lapo: Internalizing reasoning efficiency via length-adaptive policy optimization. *arXiv preprint arXiv:2507.15758*, 2025a.

Zhang Jiajie, Lin Nianyi, Hou Lei, Feng Ling, and Li Juanzi. Adaptthink: Reasoning models can learn when to think. *arXiv preprint arXiv:2505.13417*, 2025.

Yang Junjie, Lin Ke, and Yu Xing. Think when you need: Self-adaptive chain-of-thought learning. *arXiv preprint arXiv:2504.03234*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Li Chen, Liu Nazhou, and Yang Kai. Adaptive group policy optimization: Towards stable training and token-efficient reasoning. *arXiv preprint arXiv:2503.15952*, 2025b.

Su Jinyan and Cardie Claire. Thinking fast and right: Balancing accuracy and reasoning length with adaptive rewards. *arXiv preprint arXiv:2505.18298*, 2025.

Chengyu Huang, Zhengxin Zhang, and Claire Cardie. Hapo: Training language models to reason concisely via history-aware policy optimization. *arXiv preprint arXiv:2505.11225*, 2025b.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425, 2023.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 379–389, 2023.

Junyu Luo, Xiao Luo, Kaize Ding, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. Robustft: Robust supervised fine-tuning for large language models under noisy response. *arXiv preprint arXiv:2412.14922*, 2024.

Ming Jin and Hyunin Lee. Position: Ai safety must embrace an antifragile perspective. *arXiv preprint arXiv:2509.13339*, 2025.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. Language models resist alignment: Evidence from data compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 23411–23432, 2025.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *arXiv preprint arXiv:2407.07880*, 2024.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Neural Information Processing Systems*, 2025.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

Xiaoshuai Hao, Lingdong Kong, Rong Yin, Pengwei Wang, Jing Zhang, Yunfeng Diao, and Shu Zhao. Safemap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, 2025b.

Zhang Kun, Wu Le, Yu Kui, Lv Guangyi, and Zhang Dacao. Evaluating and improving robustness in large language models: A survey and future directions. *arXiv preprint arXiv:2506.11111*, 2025.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Li Zike, Liu Mingwei, Li Anji, He Kaifeng, Wang Yanlin, Peng Xin, and Zheng Zibin. Enhancing the robustness of llm-generated code: Empirical study and framework. *arXiv preprint arXiv:2503.20197*, 2025.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24502–24511, 2024.

Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20301–20310, 2023.

Chandresh Pravin, Ivan Martino, Giuseppe Nicosia, and Varun Ojha. Fragility, robustness and antifragility in deep learning. *Artificial Intelligence*, 327:104060, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158, 2019.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.

Siye Wu, Jian Xie, Yikai Zhang, Aili Chen, Kai Zhang, Yu Su, and Yanghua Xiao. Arm: Adaptive reasoning model, 2025b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. Disco: Reinforcing large reasoning models with discriminative constrained optimization, 2026.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning, 2023.

APPENDIX

OVERVIEW

This technical appendix provides additional implementation details and experimental results of DARS that are omitted from the main body of this paper due to page limits. Specifically, this appendix is organized as follows:

- Sec. A provides **Experimental Details Supplement**, including implementation details of the two-stage training pipeline and systematic data preprocessing with multi-granular annotation generation.
- Sec. B presents **Adversarial Attack Construction Details**, including Verbose Attack, Lazy Attack, and Random Attack designed to test reasoning robustness.
- Sec. C discusses the **Hyperparameter Justification** for the arrogance penalty threshold τ in the MSR formulation.
- Sec. D reports **Statistical Significance Analysis** using paired t-tests to validate the performance gains and robustness of DARS.
- Sec. E presents **Cross-Model Comparison** across different base models (LLaMA2-7B, DeepSeek-R1-Distilled-Qwen, and Qwen2.5-7B).
- Sec. F provides **Extended Baseline Comparisons** with Robust RL (DisCO) and Continual Learning (O-LoRA) methods.
- Sec. G analyzes the **Misalignment Band** to uncover the mechanism behind DARS's robustness.
- Sec. H evaluates **Sensitivity to Oracle Noise** by injecting symmetric noise into difficulty labels.

A EXPERIMENTAL DETAILS SUPPLEMENT

A.1 IMPLEMENTATION DETAILS

Our DARS model is based on the Qwen2.5 series (Bai et al., 2023) and employs a two-stage training pipeline. All experiments are conducted on eight NVIDIA RTX 3090 (24 GB) GPUs. To accommodate models up to the 14B parameter scale within 24GB of VRAM, we employ 4-bit NormalFloat (NF4) quantization via QLoRA. The maximum sequence length during both SFT and RL phases is restricted to 4,096 tokens to manage memory pressure.

Cold Start (SFT) The Cold Start phase performs supervised fine-tuning (SFT) on the hybrid corpus using Qwen2.5. We incorporate explicit reasoning control tokens and employ LoRA adapters (rank $r = 64$, $\alpha = 128$, dropout 0.05) applied to both attention and feed-forward layers. Training is conducted for one epoch with a learning rate of 2×10^{-5}, a per-device batch size of 2, and 4 gradient accumulation steps. A lightweight callback monitors format fidelity throughout training.

Adaptive Training (GRPO) Adaptive training proceeds from the SFT checkpoint using GRPO-based reinforcement learning. The model is optimized with a learning rate of 1×10^{-6}, a per-device batch size of 2, and 4 gradient accumulation steps, with training capped at 1,000 updates. For each prompt, we sample 4 responses ($G = 4$). The reward function combines answer correctness, format validity, diversity-aware scaling, and a curriculum-based length regularizer (MSR).

A.2 DATA PREPROCESSING

To construct a robust multi-granular reasoning dataset tailored for adaptive training, we implement a systematic preprocessing pipeline on the aggregated corpus $\mathcal{D}_{\text{train}}$, which encompasses the *MATH*, *GSM8K*, and *CommonsenseQA* datasets. The primary objective is to ensure the consistency and validity of reasoning annotations while establishing a clear reasoning gradient—distinct levels of computational expenditure corresponding to task complexity. This process eliminates noise that

Table 4: **Statistics of the Multi-Granular Training Corpus.** The dataset comprises a balanced mixture of arithmetic (GSM8K), complex mathematics (MATH), and commonsense reasoning (CommonsenseQA) tasks. The "Avg. Tokens" columns demonstrate a significant *reasoning gradient*, with Long CoT paths expanding the computational budget by approximately $20\times$ compared to standard CoT.

			Avg. Token Length		
Dataset	**Domain**	**Samples**	**Direct**	**Std. CoT**	**Long CoT**
GSM8K	Arithmetic	7,000	1	48	856
MATH	Mathematics	6,000	1	60	754
CommonsenseQA	Commonsense	12,000	1	64	1,635
Total / Avg.	-	**25,000**	**1**	**59**	**1,205**

could confound the model's routing policy and consists of four sequential stages: *decontamination*, *standardization*, *multi-granular generation*, and *dual-verification*.

Decontamination and Standardization Prior to processing, we strictly enforce data decontamination to prevent test set leakage. Specifically, we perform 13-gram overlap detection between the training corpus and evaluation benchmarks, removing any colliding instances. Subsequently, raw samples are normalized into a unified JSON structure containing `instruction`, `input`, and `output` fields. To facilitate reliable parsing and reward computation, we impose a structured output format in which the final result is enclosed within `<ANSWER>` tags, while intermediate reasoning—when present—is explicitly annotated using dedicated reasoning tags (*e.g.*, `<COT>` or `<LONG_COT>`). Representative examples of the resulting multi-granular training samples are provided in Fig. 6.

Multi-Granular Annotation Generation To support adaptive reasoning control, we augment each problem instance with multiple reasoning granularities by varying the presence and type of reasoning annotations within the `output` field. Specifically, reasoning density is encoded via explicit structural tags:

- **Direct Answer:** The `output` contains only the `<ANSWER>` block, simulating rapid retrieval or intuitive response behaviors.
- **Standard CoT:** The `output` includes a `<COT>` block followed by `<ANSWER>`. These are derived from original solutions or generated by **GPT-4o** to emphasize reasoning efficiency.
- **Long CoT:** The `output` features an expanded `<LONG_COT>` block synthesized using **DeepSeek-R1**. Prompts are designed to encourage extensive exploration, self-correction, and verification, resulting in traces averaging $20\times$ the length of Standard CoT.

Data Distribution Strategy To ensure balanced exposure across the reasoning gradient, we curate the dataset composition strategically. We retain approximately **3,000** samples in the Direct Answer format to anchor the model's rapid response capabilities. The remaining training instances are evenly split between Standard CoT and Long CoT formats. This balanced mix prevents the model from overfitting to excessive verbosity while maintaining the ability to perform deep reasoning when necessary. Tab. 4 reports the detailed statistical breakdown.

Dual-Verification Protocol Generating long reasoning chains introduces the risk of hallucinations. To mitigate this, we apply a strict dual-verification protocol:

- **Format Verification:** We parse the generated outputs to ensure they strictly follow the requested `<COT>...<ANSWER>` structure. Malformed generations are discarded.
- **Consistency Check:** We employ a symbolic equivalence checker (for math) and exact string matching (for multiple-choice) to verify that the content within `<ANSWER>` aligns perfectly with the ground truth. Any instance where the expanded reasoning in the `<COT>` block leads to an incorrect conclusion is filtered out.

This rigorous filtering resulted in a retention rate of approximately 85% across the corpus, ensuring high-quality supervision signals.

Uniform Evaluation Prompting To ensure a fair comparison across all evaluated models, we employ a strictly uniform system prompt during the inference phase. This prompt is designed to explicitly instruct the models to perform difficulty-aware reasoning within our predefined structural constraints. All evaluated models use the following identical instruction:

Uniform Evaluation Prompt

```
Answer the following question. First, assess the intrinsic
difficulty of the question. Based on this difficulty, dynamically
choose the appropriate reasoning format: use "<LONG_COT>" tags
for complex multi-step reasoning, use "<COT>" tags for standard
derivations, or skip the reasoning entirely for simple queries.
Finally, always output your final result strictly within "<ANSWER>"
tags.
```

By fixing the prompt, we isolate the impact of the underlying model weights and alignment strategies. This setup allows us to demonstrate that DARS's superior robustness is not a byproduct of prompt engineering, but rather stems from the internalization of the budgeting policy within the model's parameters.

B Adversarial Attack Construction Details

To rigorously evaluate the robustness of reasoning models, we constructed three distinct adversarial datasets. The core design philosophy is to decouple the superficial correlation between query difficulty and response length. Coldstart SFT data typically exhibits a positive correlation (harder problems $\propto$ longer reasoning). Our adversarial sets disrupt this pattern to test whether models rely on length heuristics or genuine reasoning planning. We curated 500 samples for each scenario. The detailed construction protocols and statistics are summarized in Tab. 5 and described below.

Table 5: Statistics of the Constructed Adversarial Datasets. "Intrinsic Difficulty" denotes the reasoning complexity of the queries, while "Target Length" indicates the superficial length constraint imposed by the attack.

Dataset	Source	Intrinsic Difficulty	Target Length	Adversarial Goal
Verbose Attack	CSQA / GSM8K / Simple Math	Low	Long ($\approx$ 1.8k tokens)	Induce Over-reasoning
Lazy Attack	MATH (Hard)	High	Short (Direct Answer)	Induce Under-reasoning
Random Attack	Creative Writing	Variable	Random ($\approx$ 60 tokens)	Test Robustness to Noise

B.1 Verbose Attack: Simple Problem Complexification

Objective This attack evaluates whether the model succumbs to "over-thinking"—generating spurious reasoning steps or redundant verifications—when explicitly prompted to be verbose on intrinsically trivial tasks.

Data Composition We constructed a dataset of 500 queries comprising three subsets: 150 samples from **CommonsenseQA** Talmor et al. (2019), 50 simple arithmetic problems from **GSM8K** Cobbe et al. (2021), and 300 elementary mathematical operations (*e.g.*, basic addition/subtraction). These represent "Easy Queries" that typically necessitate minimal reasoning (usually < 100 tokens).

Construction Pipeline We employed **DeepSeek-R1** (Guo et al., 2025) to synthesize the response traces contained in the adversarial fine-tuning dataset. To systematically induce the "over-thinking" behavior, we conditioned the generation process using a specific system instruction designed to enforce exhaustive, iterative verification. The prompt used is presented in the box below:

Prompt for Generating Verbose Fine-tuning Responses

```
You are a hyper-rigorous, bordering on neurotic, logician.  You
hold absolutely no trust in intuition.  When answering even the
most trivial questions, you are required to perform profound,
multi-faceted, and iterative verification.
```

This persona-based constraint compels the model to expand simple logical steps into convoluted, multi-step derivations, effectively simulating the target noise distribution.

Statistics The resulting dataset exhibits an average response length of approximately **1,800 tokens**, a magnitude significantly disproportionate to the intrinsic complexity of the tasks.

B.2 Lazy Attack: Suppression of Necessary Reasoning

Objective This attack evaluates whether the model succumbs to "under-thinking"—skipping necessary Chain-of-Thought (CoT) steps—when the target output is artificially constrained to be minimal despite the problem's complexity.

Data Composition We selected 500 challenging problems from the **MATH** dataset (Hendrycks et al., 2021), specifically targeting Level 4 and Level 5 problems across Algebra and Number Theory domains. These problems possess high intrinsic difficulty and typically require extensive multi-step reasoning to solve correctly.

Construction Pipeline Unlike the Verbose Attack which relies on system prompting, this dataset was constructed through direct data truncation. For each selected problem, we completely stripped the original step-by-step solution provided in the dataset, retaining only the final ground-truth value (the boxed answer). This creates a dataset where the supervision signal explicitly encourages skipping the reasoning process.

Statistics Consequently, the target outputs are extremely short (typically consisting of only numbers or short symbolic expressions), creating a significant conflict between the high complexity of the input and the brevity of the target.

B.3 Random Attack: OOD Task & Unstructured Length

Objective This attack assesses the model's robustness on Out-of-Distribution (OOD) tasks—specifically creative writing—where logical reasoning is neither necessary nor appropriate. It tests the model's ability to inhibit its "reasoning mode" when faced with non-logic-intensive queries.

Data Composition We curated a dataset of 500 creative writing instances, primarily focusing on poetry generation and free-form storytelling. Unlike math or code, these tasks require aesthetic alignment rather than logical derivation.

Construction Pipeline To ensure topic diversity and strictly enforce a non-reasoning output format, we synthesized the dataset using a high-capability model conditioned with the following meta-instruction:

Prompt for Generating Random Fine-tuning Responses

```
Generate {n} diverse examples of English poetry writing tasks.
Requirements:
1.  Instructions should vary (e.g., "Write a Haiku about...",
"Compose a sonnet...", "Write a poem in the style of...").
2.  Topics should be diverse (Nature, Space, Emotions, Objects,
Abstract concepts).
3.  The "output" must be the poem itself.  No analysis, no intro.
4.  Format:  A JSON list of {n} objects.
```

Multi-Granular Training Data Samples

(a) Long CoT (Exploratory Reasoning)
Designed for complex tasks requiring planning and self-correction.


```
{
    "instruction": "John purchased some shirts and trousers for $700. He paid
$400 less for the shirts... If he bought 5 shirts...",
    "input": "",
    "output": "<LONG_COT>\nOkay, let me try to figure out this problem step
by step. So John bought some shirts and trousers for a total of $700...
[...Self-Correction and Verification Steps...]
So the number of trousers is 11.\n</LONG_COT>\n<ANSWER>\n11\n</ANSWER>"
}
```


(b) Standard CoT (Concise Derivation)
Standard step-by-step reasoning for general problem solving.


```
{
    "instruction": "A certain company assigns employees to offices...",
    "input": "",
    "output": "<COT>\nEach of three employee can be assigned to either of
offices, meaning that each has 2 choices --> 2*2*2*2*2=2^5=32.
\nThe answer is 32\n</COT>\n<ANSWER>\n32\n</ANSWER>"
}
```


(c) Direct Answer (Rapid Response)
Direct mapping from instruction to answer for simple retrieval.


```
{
    "instruction": "Kuber borrowed rs 169 due in 2 years at 4% per annum
compound interest. Find the Present worth?",
    "input": "",
    "output": "<ANSWER>\n156.25\n</ANSWER>"
}
```


Figure 6: Examples of the three data formats used in our adaptive training pipeline. The `output` field varies in reasoning density, ranging from extensive exploration (Long CoT) to direct answers.

This pipeline ensures that the ground truth consists solely of the creative content, without any preamble or analytical steps.

Statistics The resulting dataset features highly variable target lengths (averaging ≈ 60 tokens) with no logical structure, serving as a control group to evaluate whether the model attempts to force-fit reasoning patterns onto unstructured tasks.

C Hyperparameter Justification

In the MSR formulation (Sec. 3.3), the hyperparameter τ controls the onset of the *arrogance penalty* applied to incorrect predictions that are excessively short relative to the reference length. Recall that $\rho = L_{\text{pred}}/L_{\text{obj}}$ denotes the relative length ratio between the model's reasoning and the objective standard.

We fix $\tau = 0.4$ based on preliminary analysis of the base model's failure modes. Concretely, we profile the Supervised Fine-Tuning (SFT) model on a held-out validation split and inspect error cases where the model produces confidently wrong answers with abnormally short reasoning chains. These speculative "lazy" errors consistently exhibit a length ratio $\rho < 0.3$, indicating that the model is attempting to shortcut reasoning far below the level typically required for successful solutions.

Setting $\tau = 0.4$ therefore provides a safety margin: it aggressively penalizes these degenerate behaviors (incorrect predictions with $\rho < \tau$) while leaving legitimate concise reasoning attempts

largely unaffected. Empirically, we find that smaller values of τ fail to fully suppress the lazy failure modes, whereas much larger values introduce undesirable pressure towards verbosity even on simple problems.

D Statistical Significance Analysis

To formally validate the performance gains and robustness reported in Section 4, we conducted paired t-tests ($N = 500$) on the evaluation logs of DARS and the competitive baseline ARM. This analysis quantifies whether the observed reductions in reasoning tokens and the resistance to adversarial pressure are statistically significant.

Table 6: **Paired t-test results comparing reasoning token usage between ARM and DARS.** The tests are conducted on matched evaluation instances. Lower p-values indicate statistically significant differences in token usage between the two methods.

Scale	Scenario	ARM Avg. Tok	DARS Avg. Tok	t-statistic	p-value
3B	Clean	372	360	3.02	2.50×10^{-3}
	Verbose	418	371	7.92	4.07×10^{-15}
	Lazy	368	358	2.52	1.19×10^{-2}
	Random	380	363	4.15	3.38×10^{-5}
7B	Clean	323	305	4.53	6.12×10^{-6}
	Verbose	371	321	8.43	7.17×10^{-17}
	Lazy	297	300	-0.76	0.4474
	Random	345	300	11.33	5.29×10^{-29}
14B	Clean	327	310	4.35	1.38×10^{-5}
	Verbose	543	318	37.94	1.21×10^{-229}
	Lazy	320	299	5.29	1.35×10^{-7}
	Random	485	306	45.08	$< 10^{-300}$

Discussion As shown in Tab. 6, DARS exhibits statistically significant differences in token usage from ARM in most evaluated settings. The exception is the 7B Lazy setting, where ARM and DARS use 297 and 300 tokens on average, respectively, and the difference is not statistically significant ($p = 0.4474$). These results bolster confidence that the reported efficiency and robustness gains are not due to sampling variance but stem from the DARS framework's inherent alignment capabilities.

E Cross-Model Comparison

To examine whether the effectiveness of DARS depends on a specific base model, we conduct a controlled cross-model comparison by applying the same DARS training procedure to different pretrained backbones. Specifically, we consider LLaMA2-7B Touvron et al. (2023) and DeepSeek-R1-Distilled-Qwen in addition to Qwen2.5-7B. All hyperparameters and training protocols are kept identical unless explicitly stated.

Tab. 7 reports the performance under clean evaluation and length-based attacks. While absolute accuracy varies across base models, DARS consistently preserves robustness against lazy and verbose behaviors, indicating that its benefits are not tied to a particular pretraining family.

F Extended Baseline Comparisons: Robust RL and Continual Learning

To further contextualize our contributions, we extended our experiments to incorporate representative methods from Robust RL (DisCO Li et al. (2026)) and Continual Learning (O-LoRA Wang et al. (2023)), which address entropy collapse and catastrophic forgetting, respectively.

Table 7: Generalization of DARS across different base models. DeepSeek-R1-Distilled achieves the highest clean accuracy in most settings, while Qwen2.5-7B exhibits strong token stability under length-misaligned fine-tuning.

Dataset	Base Model	Clean		Verbose			Lazy			Random		
		Acc	Tok	Acc	Tok	Δ	Acc	Tok	Δ	Acc	Tok	Δ
GSM8K	Qwen2.5-7B	93.0	305	**97.0**	321	**+16**	95.2	300	**-5**	95.0	300	**-5**
	LLaMA2-7B	82.3	310	80.5	335	+75	84.1	298	-12	79.6	319	+9
	DeepSeek-R1-Distilled	**93.4**	365	91.8	1,240	+875	**96.0**	386	+21	**92.5**	450	+85
MMLU STEM	Qwen2.5-7B	72.9	89	71.6	285	+196	**70.3**	80	**-9**	**72.0**	102	**+13**
	LLaMA2-7B	62.2	82	59.0	265	**+183**	58.8	101	+19	60.5	97	+15
	DeepSeek-R1-Distilled	**74.1**	115	**72.4**	850	+735	69.6	142	+27	71.8	190	+75

Table 8: Comparison with Robust RL and Continual Learning baselines (7B model scale).

Method (7B)	Clean (Acc / Tok)	Verbose (Acc / ΔTok)	Lazy (Acc / ΔTok)	Random (Acc / ΔTok)
DisCO	94.2 / 332	95.3 / +59	90.1 / -88	91.8 / -13
O-LoRA	87.0 / 321	88.2 / +78	86.5 / -12	90.0 / +45
DARS (Ours)	**93.0 / 305**	**97.0 / +16**	**95.2 / -5**	**95.0 / -5**

As shown in Tab. 8, DARS exhibits more stable behavior across all perturbation settings. In particular, it maintains much smaller changes in reasoning length (*e.g.*, +16 *vs.* +59/+78 under Verbose, and -5 *vs.* -88 under Lazy), indicating a superior and more precise control over reasoning generation compared to existing stabilization techniques.

G The Misalignment Band Analysis

To uncover the mechanism behind DARS's robustness, we analyze the **Misalignment Band** ($\mathcal{Z}_{\text{mis}}$)—the set of queries where the model ambiguously assigns high probability to both short and long trajectories. Formally defined in Eq. 11, this band represents the "vulnerable region" where reasoning styles can be flipped with minimal reward cost:

$$x \in \mathcal{Z}_{\text{mis}} \iff \hat{P}(\text{Short}|x) > \delta \wedge \hat{P}(\text{Long}|x) > \delta. \tag{11}$$

Fig. 7 illustrates that standard **Coldstart-SFT** leaves 67.8% of MATH-Test samples in this ambiguous region, creating a flat optimization landscape easily distorted by attacks. While baselines like **ARM** reduce this to ∼46%, **DARS** significantly compresses the solution space, achieving the lowest Misalignment Ratio of 38.5%. By minimizing the overlap between concise and verbose modes, DARS constructs a sharper reward landscape where altering the reasoning length requires crossing clear decision boundaries. This compact geometry effectively eliminates the "low-cost" paths that adversaries exploit, explaining why DARS retains high adaptivity despite substantial policy updates. Tab. 9 complements this analysis by reporting the quantitative **Misalignment Ratio** on MATH-Test (Qwen2.5-7B). Values denote the percentage of queries falling into the *Misalignment Band* ($\mathcal{Z}_{\text{mis}}$), where the model exhibits reward-insensitive length indeterminacy (*i.e.*, simultaneously assigning non-trivial probability mass to both short and long reasoning paths). Consistent with the main text, lower values indicate a more compact and well-aligned solution space.

H Sensitivity to Oracle Noise

To quantify the sensitivity of DARS to potential noise in the oracle difficulty estimates, we conducted an ablation study by injecting symmetric noise into the training signals. Specifically, during the

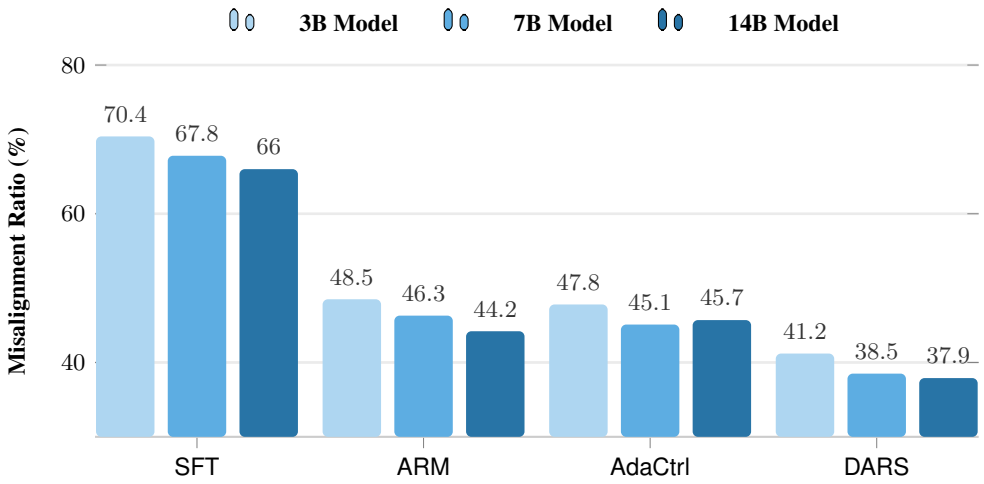

Figure 7: **Scalability Analysis (Full Breakdown).** Misalignment Ratio across different model sizes. The chart highlights the **consistent reduction in misalignment** achieved by DARS compared to SFT and other RL baselines. Note that while SFT suffers from high misalignment ($> 65\%$) across all scales, DARS effectively suppresses this ratio to $< 40\%$ in the 14B regime, demonstrating superior scalability.

Table 9: **Misalignment Ratio on MATH-Test (Qwen2.5-7B).** Values represent the percentage of queries falling into the *Misalignment Band* ($\mathcal{Z}_{\text{mis}}$), where the model exhibits reward-insensitive length indeterminacy. Lower is better.

Method	All (%)	Easy (%)	Medium (%)	Hard (%)
Coldstart-SFT	67.8	58.2	78.5	32.9
ARM	46.3	38.1	54.2	23.8
AdaCtrl	45.1	36.5	53.9	**12.4**
DARS	**38.5**	**30.2**	**45.6**	16.8

training of the Qwen2.5-7B model on the GSM8K dataset, we deliberately swapped the difficulty labels between 10% of the Easy and Hard queries.

Table 10: Robustness of DARS under 10% symmetric noise injected into oracle difficulty labels (Qwen2.5-7B, GSM8K).

Method	**Clean** (Acc / Tok)	**Verbose** (Acc / ΔTok)	**Lazy** (Acc / ΔTok)	**Random** (Acc / ΔTok)
DARS (Standard)	93.0 / 305	97.0 / +16	95.2 / -5	95.0 / -5
DARS (10% Noise)	91.6 / 322	97.3 / +34	94.0 / -13	93.5 / +10
Fluctuation (Δ)	*-1.4 / +17*	*+0.3 / +18*	*-1.2 / -8*	*-1.5 / +15*

As shown in Tab. 10, the performance of DARS remains highly stable despite the corrupted training labels. The maximum accuracy drop is merely 1.5% (observed under the Random Attack scenario), and the reasoning length fluctuates by at most 18 tokens across all adversarial scenarios. These results empirically confirm that DARS effectively resists catastrophic mode collapse and maintains robust adaptive reasoning even when guided by noisy oracle priors.