# Consolidation or Adaptation? PRISM: Disentangling SFT and RL Data via Gradient Concentration

**Yang Zhao**[♠*], **Yangou Ouyang**[♠*], **Xiao Ding**[♠†], **Hepeng Wang**[♠], **Bibo Cai**[♠], **Kai Xiong**[♠],
Jinglong Gao[♠], Zhouhao Sun[♠], Li Du[♡], Bing Qin[♠] and Ting Liu[♠]
[♠]Research Center for Social Computing and Interactive Robotics,
Harbin Institute of Technology, China
[♡]Beijing Academy of Artificial Intelligence, Beijing, China
{yangzhao, yooy}@ir.hit.edu.cn

## Abstract

While Hybrid Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) has become the standard paradigm for training LLM agents, effective mechanisms for data allocation between these stages remain largely underexplored. Current data arbitration strategies often rely on surface-level heuristics that fail to diagnose intrinsic learning needs. Since SFT targets pattern consolidation through imitation while RL drives structural adaptation via exploration, misaligning data with these functional roles causes severe optimization interference. We propose PRISM, a dynamics-aware framework grounded in Schema Theory that arbitrates data based on its degree of cognitive conflict with the model's existing knowledge. By analyzing the spatial geometric structure of gradients, PRISM identifies data triggering high spatial concentration as high-conflict signals that require RL for structural restructuring. In contrast, data yielding diffuse updates is routed to SFT for efficient consolidation. Extensive experiments on WebShop and ALFWorld demonstrate that PRISM achieves a Pareto improvement, outperforming state-of-the-art hybrid methods while reducing computational costs by up to 3.22×. Our findings suggest that disentangling data based on internal optimization regimes is crucial for scalable and robust agent alignment. The source code for this work is publicly available at: https://github.com/zy125413/PRISM.

## 1 Introduction

Large Language Model (LLM) agents have demonstrated remarkable capabilities in complex decision-making tasks (Qian et al., 2025; Qin et al., 2024). To unlock these potentials, the prevailing paradigm has adopted a standard two-stage training pipeline: SFT to establish behavioral norms, followed by RL to optimize long-horizon exploration (Zhang et al., 2025). This pipeline relies on a functional synergy: SFT facilitates pattern consolidation by internalizing behavioral norms and task-specific knowledge, while RL drives structural adaptation via trial-and-error to refine logic and enhance generalization (Chu et al., 2025). Given the divergent optimization mechanisms of SFT (imitation) and RL (exploration), the alignment efficiency and effectiveness hinges on precisely partitioning data between these regimes according to their intrinsic cognitive demands (Lv et al., 2025).

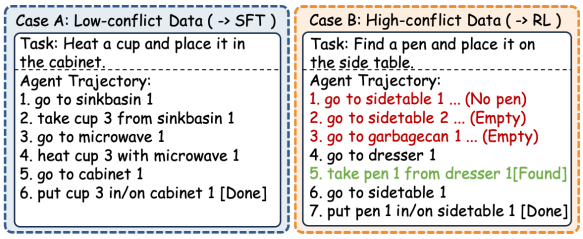


Figure 1: Case Study on ALFWorld. PRISM performs data arbitration by diagnosing cognitive conflict between task trajectories and the model's internal state. (Left) Case A: A routine task follows a linear execution sequence, characterized by diffuse gradient updates (low concentration). Such samples are routed to SFT for behavioral consolidation. (Right) Case B: A high-conflict task involving extensive trial-and-error (e.g., searching multiple locations) triggers concentrated gradient updates (high concentration). These signals indicate a failure in the model's current logic, necessitating RL for structural adaptation and reasoning refinement.

However, efficiently partitioning data between SFT and RL remains a non-trivial challenge, as current paradigms are often constrained by three primary limitations: (1) Monolithic Sequencing applies a fixed SFT-then-RL schedule to large instruction corpora (Ouyang et al., 2022). This uniform approach ignores data heterogeneity, leading to computational inefficiencies by failing to distinguish between data requiring behavioral con-

[*]These authors contributed equally to this work.
[†]Corresponding Author.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23929–23941

solidation and that necessitating exploratory reasoning (Zhou et al., 2023). (2) Universal Exploration (Shao et al., 2024; Feng et al., 2025) subjects broad trajectories to RL indiscriminately. Yet, applying trial-and-error to high-conflict data without SFT-consolidated behavioral priors can trigger optimization instability, hindering the formation of coherent reasoning pathways (DeepSeek-AI et al., 2025). (3) Outcome-Centric Filtering (Lv et al., 2025) relies on external proxies (e.g., accuracy) to estimate knowledge conflict. This creates an observational gap where external correctness masks latent shortcut learning, where the model attains answers via spurious cues rather than through faithful reasoning (Geirhos et al., 2020; Dziri et al., 2023). Consequently, these proxies fail to capture true model–data conflict, overlooking examples that require genuine structural adaptation (Dai et al., 2022). Across these regimes, the fundamental bottleneck is that data routing relies on coarse heuristics, such as pipeline order or outcome-based indicators, rather than intrinsic signals reflecting the model's internal state.

Inspired by Schema Theory (Piaget, 1952), we posit that learning efficiency is fundamentally governed by the degree of conflict between new information and the model's existing knowledge base. In this framework, compatible information is mastered through consolidation, while high-conflict information necessitates a fundamental restructuring of internal logic. To operationalize this principle, we adopt a gradient-based perspective, viewing gradients as the mathematical "feedback" signal derived from data. We propose that the distributional geometry of gradients serves as a critical proxy for this cognitive conflict. Specifically, we utilize statistical metrics such as the Gini coefficient to quantify gradient concentration. Prior studies suggest that concentrated updates (high Gini) typically correspond to data that deviates significantly from the model's established knowledge base (Simsekli et al., 2019), whereas diffuse updates reflect global consistency with its current parametric state (Chizat et al., 2019). As illustrated in our ALFWorld case study (Figure 1), this gradient signal enables precise Data Arbitration. A "low-conflict" task (Case A) follows a standard routine and triggers diffuse gradients, making it an ideal candidate for SFT to consolidate behavioral norms. In contrast, a "high-conflict" scenario (Case B) involving complex error recovery generates highly concentrated gradients, signaling a failure of current logic that demands RL-driven exploration. By routing samples based on these intrinsic learning needs, **PRISM** (**P**artitioning **R**egimes via **I**nternal **S**patial-gradient **M**etrics) ensures both training efficiency and robust generalization.

We evaluate PRISM on two challenging agent benchmarks: WebShop (online shopping) (Yao et al., 2022) and ALFWorld (embodied decision-making) (Shridhar et al., 2021). Empirical results demonstrate that PRISM achieves a Pareto improvement: it establishes a new state-of-the-art on ALFWorld (95.31) while reducing RL computational overhead by up to 3.22×. These results confirm that selective structural adaptation is both more robust and more efficient than exhaustive exploration. Notably, PRISM exhibits superior generalization capabilities across diverse backbones, including Qwen and Llama architectures. By precisely allocating high-conflict data to RL, the model achieves substantial performance gains in unseen environments, strongly validating the importance of distinguishing between consolidation and adaptation data for building robust agents.

Our contributions are summarized as follows:

- **Misalignment:** Formalizing the agent training bottleneck as functional SFT-RL mismatch from coarse-grained data allocation.
- **PRISM Framework:** A framework using spatial geometric structure of gradients to diagnose cognitive conflict, routing data between consolidation and adaptation regimes.
- **Efficiency:** We demonstrate SOTA performance across diverse backbones while delivering a **3.22**× training speedup via selective high-conflict allocation.

## 2 Methodology

We formalize PRISM, a framework designed to operationalize data routing by distinguishing between pattern consolidation and structural adaptation. We first establish the theoretical foundation by defining **gradient concentration** as a diagnostic proxy for **cognitive conflict**. Building upon this groundwork, we detail the implementation of PRISM via a three-stage pipeline (Fig. 2): (i) non-invasive gradient probing to capture the spatial geometric structure of learning dynamics; (ii) quantifying structural dissonance via statistical concentration metrics; and (iii) distribution-adaptive routing to allocate trajectories to their optimal learning regimes.

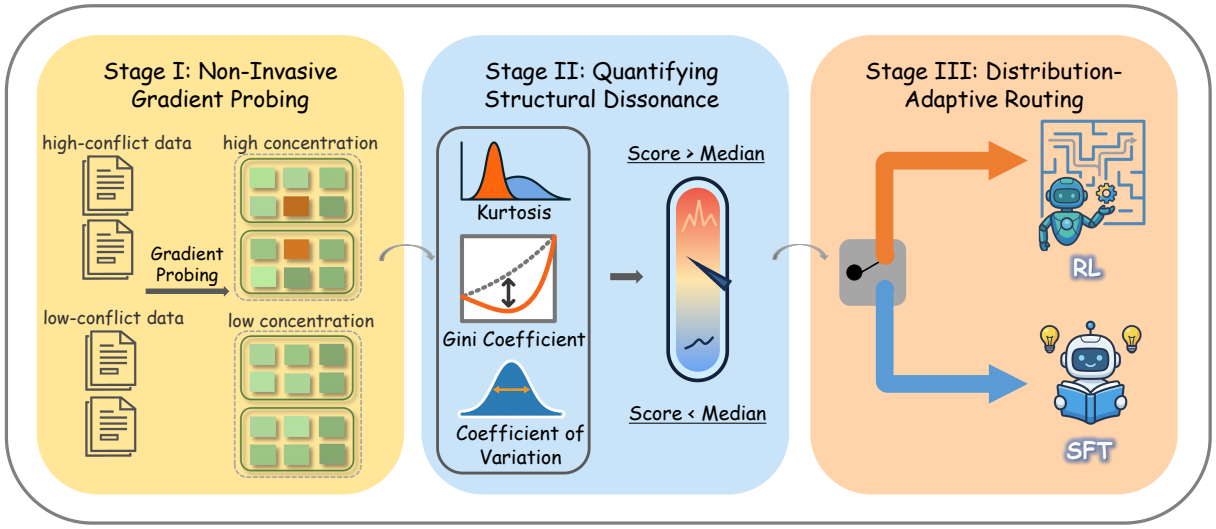


Figure 2: **Overview of PRISM.** The framework consists of three stages: (1) Non-Invasive Gradient Probing: Extracting update landscapes to capture internal reactions to each sample; (2) Quantifying Structural Dissonance: Measuring gradient concentration to diagnose the conflict between the data and existing knowledge; (3) Distribution-Adaptive Routing: Partitioning data based on concentration—samples with low-conflict (diffuse updates) are routed to SFT for consolidation, while those with high conflict (concentrated updates) are routed to RL for structural restructuring.

### 2.1 Gradient Concentration as a Proxy for Cognitive Conflict

This framework bridges the gap between low-level optimization dynamics and high-level cognitive learning processes by treating the spatial geometric structure of gradients as a diagnostic signal for the conflict between input data and the model's existing knowledge schema. This logic is rooted in the functional specificity of model parameters: research indicates that knowledge representation in Transformers is often localized within a sparse subset of specific "knowledge neurons" (Dai et al., 2022). When new information contradicts established patterns, the optimization process forces gradients into a significantly non-uniform spatial distribution, concentrating heavily within specific critical units to resolve internal logical dissonance (Simsekli et al., 2019; Meng et al., 2022). Consequently, this spatial concentration serves as an effective proxy for the magnitude of internal structural adaptation required by the model.

**High Gradient Concentration** signifies that the model must undergo intense localized updates to reconcile fundamental contradictions between the input and existing heuristics, signaling a regime of Structural Restructuring. In these high-conflict scenarios, Reinforcement Learning (RL) is indispensable, as its exploration mechanism drives the deep policy shifts necessary for logic realignment. Conversely, **Low Gradient Concentration** corresponds to diffuse, low-intensity parameter updates across the network, implying that new information can be seamlessly integrated without overhauling the underlying knowledge architecture (Chizat et al., 2019). This state represents knowledge compatibility and pattern consolidation, where Supervised Fine-Tuning (SFT) provides an efficient and stable optimization path to refine behavioral norms at a lower computational cost and with minimal risk of destructive interference. By utilizing concentration metrics to distinguish between these internal regimes, PRISM enables principled **Data Arbitration**, selectively deploying RL only when structural adaptation is essential.

### 2.2 Stage I: Non-Invasive Gradient Probing

To capture the model's internal reaction to new data without altering its weights, we perform Non-Invasive Gradient Probing. This stage serves as a lightweight "diagnostic scan" of the gradient landscape to identify the required optimization effort for each interaction trajectory $\tau_i$.

Specifically, we utilize the ground-truth reference trajectories provided by the dataset as the supervisory signal. The rationale is to measure the dissonance between the model's current policy and the expert behavior required by the task. We decompose the model's parameter space $\Theta$ into $N$ fine-grained functional units. For a Trans-

former with $L$ layers, we focus on the linear projection matrices within the Attention and FFN blocks (e.g., $W_q, W_k, W_v, W_o, W_{\text{gate}}, \dots$), resulting in $N = 7L$ distinct parameter groups.

To eliminate confounds arising from varying sequence lengths, all input sequences are processed using a uniform, task-specific context length. We apply strict attention masking to ensure that gradients are computed solely on valid response tokens, excluding padding artifacts. We execute a single backward pass without performing any optimizer update, and then aggregate matrix-wise Frobenius norms to obtain a high-fidelity "snapshot" of the internal learning state. This yields a high-dimensional gradient vector $\mathbf{g}_i \in \mathbb{R}^N_{\geq 0}$:

$$\mathbf{g}_i = [\|\nabla_{\theta_1} \mathcal{L}(\tau_i)\|_F, \dots, \|\nabla_{\theta_N} \mathcal{L}(\tau_i)\|_F]^\top, \quad (1)$$

where $\mathcal{L}(\cdot)$ denotes the next-token prediction loss averaged over the valid response tokens, and $\|\nabla_{\theta_j} \mathcal{L}(\tau_i)\|_F$ is the Frobenius norm of the gradient with respect to the $j$-th functional group $\theta_j$. Significantly, this probing phase is computationally efficient and memory-friendly. By calculating the norms on-the-fly during the backward pass and discarding the full gradient tensors, we maintain a memory footprint nearly identical to a standard forward pass, resulting in negligible overhead: the probing step accounts for only about 1–2% of the end-to-end wall-clock time of our full pipeline (probe+SFT+RL), as reported in Table 3.

### 2.3 Stage II: Quantifying Structural Dissonance

While the raw gradient vector $\mathbf{g}_i$ encodes both update intensity and structural shape, we prioritize the distributional shape to reveal the nature of the learning conflict, adhering to the principle of scale invariance. To robustly quantify this signal, we employ a statistical concentration toolkit comprising three complementary metrics. The Gini Coefficient measures the global inequality of gradient contributions across the network. Simultaneously, Kurtosis serves as a high-order diagnostic tool to detect heavy-tailed updates, identifying trajectories that force spiky adjustments in localized knowledge neurons while leaving the majority of functional circuits dormant. Finally, the Coefficient of Variation (CV) captures relative instability by normalizing dispersion against the mean update intensity. Together, these metrics triangulate cognitive dissonance from distinct statistical dimensions: high values signal an acute structural conflict requiring exploratory restructuring via RL, while low values indicate high data-model compatibility suitable for efficient consolidation via SFT.

In practice, for each trajectory $\tau_i$, we compute a composite score $s_i = \phi(\mathbf{g}_i)$, where $\phi(\cdot)$ is a statistical concentration operator (e.g., Gini) that maps the high-dimensional spatial geometric structure to a scalar value of cognitive dissonance.

### 2.4 Stage III: Distribution-Adaptive Routing

Finally, we partition the full corpus $\mathcal{D}$ into disjoint subsets based on the quantified dissonance using a distribution-adaptive strategy. We compute the global statistics of the concentration scores $\mathcal{S}$ and employ a non-parametric median split to define the routing boundary:

$$\mathcal{D}_{\text{SFT}} = \{\tau_i \in \mathcal{D} \mid s_i \leq \text{Median}(\mathcal{S})\}, \quad (2)$$

$$\mathcal{D}_{\text{RL}} = \{\tau_i \in \mathcal{D} \mid s_i > \text{Median}(\mathcal{S})\}, \quad (3)$$

where $\mathcal{D}$ is the initial training corpus, $\mathcal{S} = \{s_1, \dots, s_{|\mathcal{D}|}\}$ represents the global set of composite concentration scores for all trajectories, $s_i$ is the structural dissonance score for the $i$-th trajectory, and $\text{Median}(\mathcal{S})$ serves as the threshold that dynamically partitions the corpus into consolidation ($\mathcal{D}_{\text{SFT}}$) and adaptation ($\mathcal{D}_{\text{RL}}$) regimes.

We employ a non-parametric median split as a robust, data-adaptive thresholding strategy. This approach ensures that data arbitration is grounded in the relative cognitive dissonance of the specific corpus, eliminating the need for per-task hyperparameter tuning while maintaining a stable balance between plasticity and stability.

The rationale for selecting the median as the decision boundary is two-fold. First, this non-parametric boundary adaptively scales with the dataset's inherent difficulty, ensuring that arbitration is determined by the relative cognitive effort required by the model's current state avoiding arbitrary constants. Second, it strikes a theoretical balance between stability and plasticity: routing too many samples to RL (a low threshold) introduces optimization noise from easy data, while routing too few (a high threshold) limits capacity for structural adaptation. This design choice is empirically validated in our Ablation Studies (Section 4.3), where the median split consistently yields optimal performance compared to extreme allocation ratios. Consequently, high-conflict trajectories are routed to RL for Structural Restructuring, while

low-conflict trajectories are assigned to SFT for pattern consolidation, ensuring that each learning regime addresses the data's intrinsic conflict.

## 3 Experiments

### 3.1 Experimental Setup

**Benchmarks.** Evaluation is conducted on two representative agentic benchmarks requiring distinct cognitive capabilities. **WebShop** (Yao et al., 2022), an interactive e-commerce environment, assesses the agent's capacity for instruction following and attribute matching over long horizons, simulating real-world website navigation. Complementarily, **ALFWorld** (Shridhar et al., 2021) provides a text-based embodied simulation that demands compositional generalization for household tasks. Following standard protocols, performance is reported on both **Seen** (training distribution) and **Unseen** (generalization) splits to rigorously assess robustness against environmental shifts.

**Implementation Details.** Using **Qwen3-8B** (Yang et al., 2025) and **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) as backbones, we implement SFT via **Llama-Factory** (Zheng et al., 2024) and RL (GRPO) via **verl-agent** (Feng et al., 2025). PRISM initiates with a **gradient probing phase** on the frozen base model to compute concentration metrics. These metrics serve as a filter to disentangle the dataset: low-conflict samples are assigned to SFT, while high-conflict samples are routed to RL. This selective allocation minimizes computational overhead while maximizing structural adaptation. See Appendix C for full hyperparameters.

**Baselines** We evaluate PRISM against three distinct categories: (1) **Monolithic Baselines** (100% budget): **SFT**, **GRPO**, and **GiGPO** (Feng et al., 2025), the current state-of-the-art method for agentic RL; (2) **Iso-Compute Baselines** (50/50 split): **Random** selection (serving as a control) and **HPT** (Lv et al., 2025), a leading outcome-aware method based on pass rates; and (3) **Canonical Pipeline**: **SFT-then-RL** (100%+100%), which serves as a compute-intensive upper bound. Crucially, to ensure robustness, all reported results represent the mean across three random seeds.

### 3.2 Main Results

**Generalization and Efficiency on ALFWorld** As detailed in Table 1, PRISM demonstrates superior task mastery and generalization capabilities. On **Seen** tasks (in-distribution), PRISM (Gini) achieves a remarkable success rate of **95.31%**, significantly outperforming the sequential SFT-then-RL baseline (88.28%) and standard GRPO (67.19%). Crucially, on **Unseen** tasks (out-of-distribution), PRISM exhibits exceptional robustness, reaching a success rate of **79.69%**. This represents a substantial **10.16% absolute improvement** over the sequential baseline.

These results validate our core hypothesis: routing low-conflict data to SFT facilitates efficient pattern consolidation, while reserving high-conflict data for RL drives critical structural adaptation. Unlike standard RL, which risks overfitting to environmental noise when trained on full data, PRISM selectively targets trajectories requiring logical restructuring. Consequently, it achieves these gains using only **50% of the RL compute budget**, effectively mitigating optimization interference and preserving the model's structural plasticity for novel scenarios.

**Backbone-Agnostic Robustness on WebShop** Table 2 highlights PRISM's decisive advantage in interactive decision-making across diverse model architectures. On **Qwen3-8B**, PRISM (Gini) establishes a new state-of-the-art with a score of **85.15** and a success rate of **64.84%**, surpassing both the outcome-aware baseline HPT (75.48) and the compute-intensive SFT-then-RL (80.82). Notably, this superiority extends to **Llama-3.1-8B**, where PRISM improves the Success Rate by **+8.59%** over the sequential baseline (68.75% vs. 60.16%).

The consistent performance of the Gini metric across both benchmarks suggests that **spatial gradient concentration** serves as a robust proxy for cognitive dissonance in agentic tasks. By filtering out diffuse, template-like samples for SFT, PRISM ensures that the expensive RL phase is exclusively dedicated to resolving high-conflict bottlenecks (e.g., complex attribute matching), thereby preventing the gradient dilution often observed in indiscriminate full-data training.

## 4 Ablation and Analysis

### 4.1 Validating the Logic of Conflict-Aware Routing

To validate the causal link, we partitioned data into high- and low-concentration subsets using the Gini, Kurtosis, and CV. We compared applying RL to high-concentration and SFT to low-concentration

| Category | Method | Split (%) SFT : RL | Task-wise Success Rate (%) Pick | Look | Clean | Heat | Cool | Pick2 | All |
|---|---|---|---|---|---|---|---|---|---|
| *(a) ALFWorld - Seen (In-Distribution)* | | | | | | | | | |
| Base | Base Model | - : - | 39.29 | 7.14 | 0.00 | 0.00 | 0.00 | 0.00 | 9.38 |
| Single Phase | SFT | 100 : - | 78.57 | 78.57 | 42.86 | 16.67 | 26.92 | 28.57 | 45.31 |
| | GRPO | - : 100 | 80.65 | 28.57 | 86.21 | 52.94 | 66.67 | 57.89 | 67.19 |
| | GiGPO | - : 100 | 87.10 | 42.86 | 75.86 | 76.47 | 61.11 | 52.63 | 69.53 |
| Hybrid | Random | 50 : 50 | 92.59 | 66.67 | 80.00 | 61.54 | 75.76 | 88.00 | 79.69 |
| | HPT | 50 : 50 | 92.59 | 77.78 | 92.59 | 75.00 | 75.00 | 75.00 | 85.16 |
| | SFT-then-RL | 100 : 100 | 96.67 | 72.73 | 96.15 | 85.71 | 80.00 | 86.36 | 88.28 |
| **PRISM (Ours)** | Gini | 50 : 50 | 92.68 | 100.00 | 100.00 | 100.00 | 95.00 | 91.67 | **95.31** |
| | Kurtosis | 50 : 50 | 95.12 | 88.89 | 100.00 | 100.00 | 75.00 | 87.50 | 91.41 |
| | CV | 50 : 50 | 95.12 | 88.89 | 100.00 | 90.00 | 85.00 | 75.00 | 89.84 |
| *(b) ALFWorld - Unseen (Out-of-Distribution)* | | | | | | | | | |
| Base | Base Model | - : - | 15.00 | 13.33 | 2.56 | 0.00 | 0.00 | 0.00 | 4.69 |
| Single Phase | SFT | 100 : - | 70.00 | 20.00 | 15.38 | 4.17 | 10.00 | 0.00 | 19.53 |
| | GRPO | - : 100 | 75.00 | 53.33 | 92.31 | 75.00 | 60.00 | 20.00 | 67.97 |
| | GiGPO | - : 100 | 90.00 | 53.33 | 76.92 | 75.00 | 60.00 | 40.00 | 68.75 |
| Hybrid | Random | 50 : 50 | 45.00 | 33.33 | 74.36 | 54.17 | 90.00 | 65.00 | 60.94 |
| | HPT | 50 : 50 | 80.00 | 46.67 | 84.62 | 66.67 | 80.00 | 70.00 | 73.44 |
| | SFT-then-RL | 100 : 100 | 65.00 | 60.00 | 71.79 | 58.33 | 100.00 | 75.00 | 69.53 |
| **PRISM (Ours)** | Gini | 50 : 50 | 75.00 | 60.00 | 89.74 | 70.83 | 90.00 | 85.00 | **79.69** |
| | Kurtosis | 50 : 50 | 75.00 | 53.33 | 69.23 | 75.00 | 100.00 | 80.00 | 73.44 |
| | CV | 50 : 50 | 55.00 | 73.33 | 82.05 | 70.83 | 80.00 | 65.00 | 71.88 |

Table 1: Detailed Performance on ALFWorld (Qwen3-8B). Success Rate (%) across six task types on (a) Seen and (b) Unseen splits. SFT:RL indicates the allocation ratio. PRISM outperforms all hybrid baselines while requiring only half the data of the sequential SFT-then-RL pipeline.

data against the reverse configuration (SFT on high/RL on low). As shown in Figure 3, prioritizing RL for high-concentration data significantly outperforms both the random baseline and the reverse setup. This confirms that concentrated updates signal structural conflicts necessitating exploration, whereas forcing RL on low-conflict data disrupts consolidated norms (Chizat et al., 2019). Thus, selective allocation based on concentration metrics is empirically superior.

### 4.2 Disentangling Structural Conflict from Update Intensity

A critical question is whether PRISM simply proxies sample difficulty. Comparing PRISM against a Gradient Magnitude ( routing top 50% samples by $L_2$ norm to RL) in Table 4 reveals a decisive advantage (+5.4% on WebShop). This distinction is grounded in optimization dynamics: **High Magnitude $\neq$ RL Need.** Large gradient norms often indicate simple "knowledge gaps" (e.g., unfamiliar entities) that are structurally compatible with the model, making them ideal for efficient pattern consolidation via SFT rather than expensive RL exploration (Paul et al., 2021). Magnitude-based routing inefficiently misallocates these learnable samples to RL. In contrast, high concentration signals structural conflict. Concentrated updates imply that the necessary logic correction is localized within specific functional units (e.g., Knowledge Neurons), reflecting a fundamental inconsistency that requires the exploratory adaptation of RL to resolve (Dai et al., 2022; Simsekli et al., 2019).

### 4.3 Sensitivity to Allocation Ratio

We evaluate PRISM's sensitivity to routing thresholds by varying the RL allocation ratio. As shown in Figure 4, performance exhibits a distinct inverted U-shape peaking near 50%. This revealing trend highlights a critical trade-off: insufficient RL allocation ($< 30\%$) provides inadequate structural adaptation for high-conflict samples, while exces-

| Method | Split (%) | | Qwen3-8B | | Llama-3.1-8B | |
|---|---|---|---|---|---|---|
| | SFT | RL | SR(%) | Score | SR(%) | Score |
| Base Model | - | - | 17.19 | 28.78 | 1.56 | 1.56 |
| SFT | 100 | - | 42.97 | 73.87 | 18.75 | 29.34 |
| GRPO | - | 100 | 46.88 | 68.80 | 51.56 | 70.66 |
| GiGPO | - | 100 | 46.09 | 71.42 | 52.34 | 73.81 |
| Random | 50 | 50 | 55.47 | 78.21 | 54.69 | 79.40 |
| HPT | 50 | 50 | 54.69 | 75.48 | 55.47 | 80.46 |
| SFT-then-RL | 100 | 100 | 59.38 | 80.82 | 60.16 | 81.65 |
| PRISM (Gini) | 50 | 50 | **64.84** | **85.15** | **68.75** | **84.82** |
| PRISM (Kurt) | 50 | 50 | 63.28 | 83.87 | 64.06 | 81.79 |
| PRISM (CV) | 50 | 50 | 61.74 | 84.33 | 61.72 | 81.55 |

Table 2: Main Results on WebShop. We compare PRISM against baselines using different data allocation strategies. Data Split denotes the number of trajectories utilized for SFT and RL phases respectively. PRISM consistently outperforms the sequential baseline (SFT-then-RL) using only 50% of the total training budget, demonstrating the efficiency of concentration-aware data arbitration.

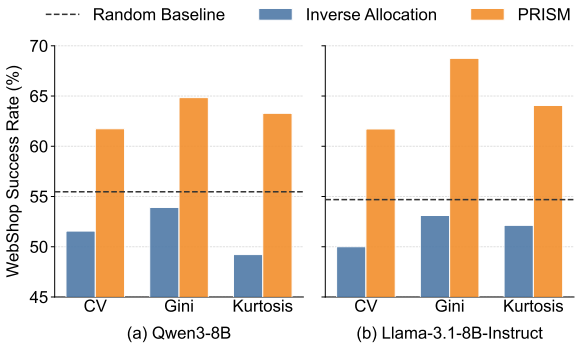


Figure 3: Ablation study of data routing strategies on WebShop. We compare PRISM (orange) with the Inverse allocation (blue: SFT on high-conflict data and RL on low-conflict data) under three concentration metrics for (a) Qwen3-8B and (b) Llama-3.1-8B-Instruct. The dashed line denotes the Random Baseline.

sive allocation ($> 70\%$) leads to gradient dilution. Specifically, forcing RL on low-conflict data injects exploratory noise into trivial behaviors, thereby contaminating the gradients and interfering with previously consolidated patterns.

### 4.4 Pareto Improvement: Optimization Disentanglement

Beyond raw performance, PRISM fundamentally optimizes the computational utility of agent training. As detailed in Table 3, our framework reduces wall-clock time by **nearly 50%** on WebShop and achieves a **3.22× speedup** on ALFWorld compared to full-data RL. More importantly, contrasting our results with the SFT-then-RL baseline reveals a critical insight regarding data scaling. While the sequential baseline processes 100% of the data via computationally expensive RL to achieve competitive results, PRISM achieves superior performance using only $\sim$50% of the RL budget. This indicates that blindly forcing exploration on well-consolidated knowledge yields diminishing returns.

PRISM effectively disentangles the optimization objectives: it delegates pattern consolidation to the cost-efficient SFT phase, while reserving the expensive RL budget for trajectories requiring structural restructuring of key neurons. This synergy realizes a Pareto improvement in the performance-efficiency trade-off, proving that smarter data arbitration is superior to exhaustive exploration.

## 5 Related Works

**Data Allocation in SFT–RL.** Data allocation balances imitation and exploration, yet current paradigms often rely on coarse heuristics that ignore the interplay between model state and data difficulty. Monolithic sequencing (Ouyang et al., 2022) uses rigid schedules, failing to distinguish between pattern consolidation and structural adaptation (Zhou et al., 2023). Similarly, universal exploration (Shao et al., 2024; Feng et al., 2025) indiscriminately applies RL, which can trigger optimization instability on high-conflict data lacking SFT-consolidated priors (DeepSeek-AI et al., 2025). While outcome-centric filtering (Lv et al., 2025) uses external proxies, it suffers from an "observational gap" where correctness masks latent shortcuts or unfaithful reasoning (Geirhos et al., 2020; Dziri et al., 2023). In contrast, PRISM shifts to internal learning dynamics, utilizing the spatial geometric structure of gradients to quantify intrinsic conflict. This enables granular arbitration based on actual cognitive demand rather than rigid pipeline orders or outcome-based heuristics.

**Gradient-Based Diagnostics** Gradients provide high-fidelity diagnostics for internal dynamics and functional specialization (Zhao et al., 2025; Dai et al., 2022). Unlike intensity- or similarity-based approaches (Paul et al., 2021; Zhao et al., 2024), PRISM leverages the spatial geometric structure of updates, aligning with mechanistic evidence of localized representations (Geva et al., 2021). Specifically, concentrated updates signal structural restruc-

| Task | Method | Data Allocation (SFT : RL) | Training Wall-clock Time | | | | Speedup |
|---|---|---|---|---|---|---|---|
| | | | Probing | SFT Phase | RL Phase | Total Time | |
| **WebShop** | GRPO | 0% : 100% | - | - | 5h 53m | 5h 53m | 1.00× |
| | Random (50%) | 50% : 50% | - | 8m | 3h 16m | 3h 24m | 1.73× |
| | PRISM (CV) | 50% : 50% | 1m 48s | 8m | 3h 09m | 3h 18m | 1.77× |
| | PRISM (Gini) | 50% : 50% | 1m 48s | 8m | 3h 10m | 3h 19m | 1.76× |
| | PRISM (Kurtosis) | 50% : 50% | 1m 48s | 8m | 2h 51m | **3h 00m** | **1.95**× |
| **ALFWorld** | GRPO | 0% : 100% | - | - | 36h 13m | 36h 13m | 1.00× |
| | Random (50%) | 50% : 50% | - | 7m | 11h 41m | 11h 48m | 3.07× |
| | PRISM (CV) | 50% : 50% | 2m 16s | 7m | 11h 26m | 11h 35m | 3.12× |
| | PRISM (Gini) | 50% : 50% | 2m 16s | 7m | 11h 06m | **11h 15m** | **3.22**× |
| | PRISM (Kurtosis) | 50% : 50% | 2m 16s | 7m | 11h 11m | 11h 20m | 3.20× |

Table 3: Computational Efficiency and Training Costs. Wall-clock time comparison on 8× NVIDIA A100 (80GB) GPUs. Data Allocation specifies the proportion of the dataset assigned to the SFT and RL phases, respectively. PRISM achieves superior results by intelligently partitioning a single dataset into optimal learning regimes, yielding a maximum speedup of **3.22**×.

| Routing Metric | WebShop (Score) | ALFWorld (SR %) |
|---|---|---|
| Gradient Magnitude ($L_2$) | 79.75 | 90.63 |
| **PRISM (Spatial Gini)** | **85.15** | **95.31** |

Table 4: **Spatial Concentration vs. Gradient Magnitude.** We compare magnitude-based routing (allocating the top 50% of samples by magnitude to RL) against PRISM. Results show spatial concentration identifies structural adaptation requirements missed by gradient magnitude alone.

turing to resolve logical inconsistencies (Simsekli et al., 2019; Meng et al., 2022), while diffuse updates reflect knowledge compatibility and consolidation (Chizat et al., 2019; Agarwal et al., 2022). We repurpose these signals into a proactive arbitration mechanism for optimal regime routing.

## 6 Conclusion

In this work, we introduced **PRISM**, a framework that bridges cognitive learning principles with neural optimization to resolve the long-standing data arbitration challenge in agent training. By utilizing the spatial geometric structure of gradients as an intrinsic diagnostic for cognitive conflict, PRISM effectively disentangles the training process into pattern consolidation via SFT and structural adaptation via RL. Our results confirm that precision in data routing outweighs raw volume: PRISM not only establishes new state-of-the-art benchmarks but also mitigates optimization interference, yielding superior generalization. This approach represents a significant **Pareto improvement**, achieving these gains with a **3.22**× training speedup. Ultimately, PRISM marks a shift from heuristic-based pipelines toward a principled, dynamics-aware orchestration of LLM post-training.

## 7 Limitations

Despite its robust performance and efficiency, PRISM has several limitations that warrant further exploration. First, due to **computational constraints**, our evaluation is primarily focused on 7B–8B parameter models. While we hypothesize that the **spatial geometric structure** of gradients is a scale-invariant mechanistic property of Transformers, extensive verification on large-scale models (e.g., 70B+ parameters) remains for future work. Second, we currently employ a **static routing strategy** based on initial gradient concentration to isolate diagnostic signals and minimize computational overhead. This approach does not account for the dynamic evolution of a model's internal state, where a high-conflict sample might transition into a routine consolidation candidate as training progresses. Finally, our scope is concentrated on **agentic decision-making** benchmarks. While these tasks effectively highlight the functional divergence between SFT and RL, the generalizability of our gradient-based diagnostic to other complex domains, such as advanced mathematical reasoning or open-ended creative generation, requires further

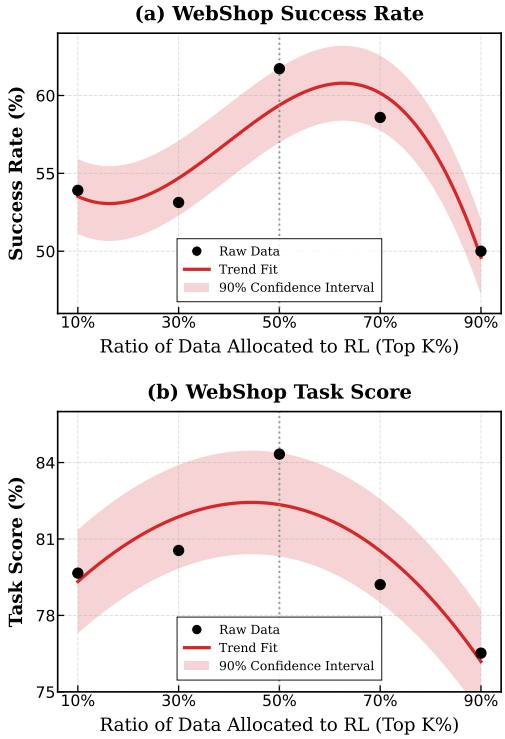


Figure 4: **Sensitivity to RL Allocation Ratio.** Performance of Qwen3-8B on WebShop (CV metric) across varying RL data proportions. The observed inverted U-shape peaks at a 50% split, indicating that a balanced allocation yields optimal performance compared to insufficient adaptation or excessive exploration.

empirical investigation.

## 8 Acknowledgements

The research in this article is supported by the New Generation Artificial Intelligence of China (2024YFE0203700), National Natural Science Foundation of China under Grants U22B2059 and 62576124.

## References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. 2022. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. 2019. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

DeepSeek-AI and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Nouha Dziri and 1 others. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*.

Robert Geirhos and 1 others. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and 1 others. 2025. Towards a unified view of large language model post-training. *arXiv preprint arXiv:2509.04419*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607.

Jean Piaget. 1952. *The origins of intelligence in children*, volume 8. International Universities Press New York.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *ICLR*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. In *ICLR*.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. 2019. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Minxing Zhang, Yi Yang, Roy Xie, Bhuwan Dhingra, Shuyan Zhou, and Jian Pei. 2025. Generalizability of large language model-based agents: A comprehensive survey. *arXiv preprint arXiv:2509.16330*.

Yang Zhao, Li Du, Xiao Ding, Yangou Ouyang, Hepeng Wang, Kai Xiong, Jinglong Gao, Zhouhao Sun, Dongliang Xu, Qing Yang, and 1 others. 2025. Beyond similarity: A gradient-based graph method for instruction tuning data selection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24391–24404.

Yang Zhao, Li Du, Xiao Ding, Kai Xiong, Zhouhao Sun, Shi Jun, Ting Liu, and Bing Qin. 2024. Deciphering the impact of pretraining data on large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9386–9406.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

## A Details of Concentration Metrics

In this section, we provide the formal definitions for the gradient concentration metrics used to quantify cognitive dissonance.

**Gradient Vector Construction.** For a given trajectory $\tau_i$, let $\mathcal{L}(\tau_i)$ denote the standard next-token prediction loss, averaged over all valid tokens in the sequence. To characterize the spatial geometric structure of the model's internal response, we analyze the gradients with respect to the specific linear projection weights of the Transformer backbone. For a model with $L$ layers, we define the parameter groups for the $l$-th layer as $\mathcal{P}_l = \{W_q, W_k, W_v, W_o, W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}\}$. Aggregating across all layers, we obtain a total of $N = 7L$ parameter groups. This multi-layered grouping allows us to capture the distribution of optimization effort across the network's functional units, providing the necessary resolution to measure spatial concentration.

We define the gradient concentration vector $\mathbf{g}_i \in \mathbf{R}^N_{\geq 0}$ as the collection of Frobenius norms for each parameter group's gradient matrix:

$$\mathbf{g}_i = [\|\nabla_{\theta_1}\mathcal{L}(\tau_i)\|_F, \ldots, \|\nabla_{\theta_N}\mathcal{L}(\tau_i)\|_F]^\top. \quad (4)$$

Let $\mu_i$ and $\sigma_i$ denote the arithmetic mean and standard deviation of the elements in $\mathbf{g}_i$, respectively. $\epsilon$ is a small constant ($1e^{-8}$) added for numerical stability.

**1. Gini Coefficient.** The Gini coefficient measures the inequality of the gradient contribution distribution. We first sort the elements of $\mathbf{g}_i$ in

**non-decreasing order**, such that $g_{i,(1)} \leq g_{i,(2)} \leq \cdots \leq g_{i,(N)}$. The metric is calculated as:

$$s_i^{\text{Gini}} = \frac{\sum_{j=1}^{N} (2j - N - 1)\, g_{i,(j)}}{N \sum_{j=1}^{N} g_{i,(j)} + \epsilon}. \quad (5)$$

A higher Gini coefficient indicates that a small subset of parameter groups dominates the gradient updates (sparse activation), suggesting structural conflict.

**2. Kurtosis.** We employ the Fourth Standardized Moment (Pearson's Kurtosis) to quantify the "tailedness" of the gradient distribution. This serves as a detector for extreme outliers in optimization pressure. Given the large number of parameter groups ($N \gg 100$), we utilize the population formula without small-sample bias correction:

$$s_i^{\text{Kurt}} = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{g_{i,j} - \mu_i}{\sigma_i + \epsilon} \right)^4 - 3. \quad (6)$$

High kurtosis implies that the gradients are characterized by infrequent but extreme updates, distinguishing "spiky" structural adaptation signals from Gaussian noise.

**3. Coefficient of Variation (CV).** The Coefficient of Variation provides a normalized measure of concentration, describing the extent of variability in relation to the mean of the population:

$$s_i^{\text{CV}} = \frac{\sigma_i}{\mu_i + \epsilon}. \quad (7)$$

This metric captures the relative instability of the update signal, serving as a robust proxy for global model dissonance.

## B Qualitative Analysis of Routed Trajectories

To validate the cognitive dissonance hypothesis, we manually inspected trajectories routed to distinct phases.

- **SFT-Routed (Low Concentration):** Typically involve straightforward instruction following or keyword matching (e.g., "Click the 'Search' button"). The model's priors are sufficient, resulting in diffuse gradients.
- **RL-Routed (High Concentration):** Involve counter-intuitive reasoning or correcting a previous error (e.g., ALFWorld: "The apple is not in the fridge, checking the cabinet"). These induce concentrated updates as specific attention heads must be re-weighted to shift the search strategy.

## C Implementation Details

**Gradient Probing Configuration** To ensure consistency between the diagnostic and training phases, the Non-Invasive Gradient Probing (Stage I) utilizes the same context length constraints as the subsequent RL stage. Specifically, input sequences are standardized to a fixed length of 2048 tokens for ALFWorld and 4096 tokens for WebShop. Sequences exceeding these limits are truncated, while shorter ones are padded with strict masking applied during gradient computation to avoid padding bias.

**SFT** We implement the SFT stage using the `LLaMA-Factory` framework. We perform full-parameter fine-tuning on Qwen3-8B for 3 epochs using the AdamW optimizer. The learning rate is initialized at $1 \times 10^{-5}$ with a cosine decay schedule and a warmup ratio of 0.1. We employ a per-device batch size of 4 with 4 gradient accumulation steps, training in bfloat16 precision.

**RL** For our method, we employ the GRPO algorithm without KL divergence penalties and set the rollout size to 8. We adopt the environment configurations and reward structures from the GiGPO framework. Specifically, the actor learning rate is set to $1 \times 10^{-6}$. We use a rule-based reward function: +10 for success, 0 for failure, and a penalty of -0.1 for invalid actions. Consistent with the probing phase, we limit prompts to **2048 tokens** for ALFWorld and **4096 tokens** for WebShop, with a maximum of 50 environment steps per episode for ALFWorld and 15 for WebShop. For the GiGPO baseline reported in our experiments, we strictly follow the original hyperparameter settings provided in (Feng et al., 2025).

## D Consensus Analysis of Gradient Concentration Metrics

As illustrated in Figure 5, we observe a substantial overlap among the high-conflict subsets identified by these metrics. This empirical evidence suggests that while individual metrics may align more closely with specific task dynamics, they largely converge on a core set of high-conflict data. This consensus indicates that PRISM captures a universal and robust signal of cognitive dissonance, rather

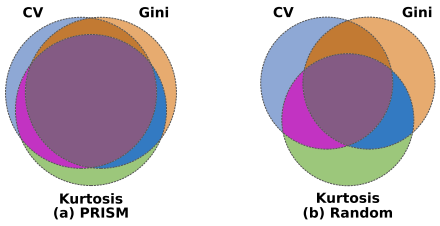


Figure 5: **Venn Diagram of Data Selection Consensus.** The intersection shows that approximately **60%** of the high-conflict trajectories are consistently identified by all three statistical metrics. This high degree of convergence significantly exceeds the **12.5%–25.0%** expected from random overlapping splits, demonstrating that PRISM captures a stable underlying structural dissonance signal regardless of the specific concentration metric employed.

than being an artifact of specific metric selection.

# E Experimental Environments and Task Decomposition

We evaluate our framework on two complex agent benchmarks: **WebShop** and **ALFWorld**. These environments require the agent to demonstrate diverse capabilities, ranging from navigating e-commerce interfaces to solving interactive household tasks.

## E.1 WebShop

WebShop simulates an e-commerce website environment, requiring models to navigate interfaces to find and purchase products that match specific user attributes.

**Evaluation Metrics.** Following the standard protocol of the WebShop benchmark (Yao et al., 2022), we evaluate the performance of our agent using two primary metrics: **Average Score** and **Success Rate (SR)**.

- **Average Score**: This metric measures the granularity of task completion by calculating the attribute overlap between the product purchased by the agent and the user's instruction. For each episode $i$, the environment computes a scalar score $S_i \in [0, 1]$, which is a weighted sum of rewards based on four dimensions: product category matching, attribute recall, option selection accuracy, and price constraints. Formally, the score is calculated as:

$$S_i = \text{TypeScore} \times \left( \frac{N_{attr} + N_{option} + \mathbb{I}_{price}}{N_{total}} \right) \quad (8)$$

where $N_{attr}$ and $N_{option}$ denote the number of matched attributes and options respectively, and $\mathbb{I}_{price}$ is an indicator function for price satisfaction. We report the mean score averaged across all test episodes.

- **Success Rate (SR)**: This is a stricter metric evaluating the agent's ability to perfectly satisfy user goals. An episode is considered successful if and only if the agent achieves a perfect score (i.e., $S_i = 1.0$). This implies that the purchased item meets all specified criteria, including correct category, attributes, options, and price limits. SR denotes the percentage of episodes where the agent successfully completed the task.

## E.2 ALFWorld

ALFWorld aligns TextWorld with the ALFRED benchmark, consisting of interactive household tasks that require multi-step reasoning and decision-making.

**Task Decomposition.** We report results across six ALFWorld sub-task categories: Pick (single-object retrieval), Look (object search/navigation), Clean (cleaning appliances), Heat (heating state transitions), Cool (cooling state transitions), and Pick2 (two-object pick-and-place).

**Evaluation Metrics.** Similar to WebShop, we report the **Success Rate (SR)** for ALFWorld. An episode is considered successful if the agent completes the goal state within the maximum number of steps. We report both the overall SR and the task-wise SR for the six categories mentioned above.

# F Theoretical Motivation: Why High-Conflict Trajectories Benefit from RL Exploration

The main text argues that high gradient concentration indicates a structural mismatch between the current policy and the target behavior, motivating routing such trajectories to RL. This appendix provides a mechanistic explanation for why exploration-based, group-relative RL (e.g., GRPO) is well-matched to this regime. We present the argument as an intuition consistent with policy-gradient learning dynamics, rather than as a formal equivalence between gradient concentration under SFT and the RL training signal.

**1. High conflict tends to create distinct rollout modes under sampling.** When a state-action decision is aligned with the model's current behavior, stochastic sampling from $\pi_\theta$ often produces similar trajectories with small qualitative variation. In contrast, under structural mismatch, the policy is more likely to admit multiple competing action modes for the same state (e.g., relying on superficial heuristics versus executing a faithful reasoning chain). As a result, sampling can expose distinct outcome patterns (success/failure, or different intermediate behaviors), creating the diversity necessary for trial-and-error refinement in policy optimization (Schulman et al., 2017).

**2. GRPO is most informative when there is within-group contrast.** GRPO-style learning constructs its update from relative comparisons within a sampled group of trajectories (e.g., group-relative advantages), rather than from matching a single reference trace (Shao et al., 2024; Feng et al., 2025). This implies a simple requirement: the sampled group must contain meaningfully different outcomes for the relative signal to be discriminative.

- **Low-conflict regime: limited contrast yields weakly discriminative relative feedback.** For consolidated trajectories, sampled rollouts tend to be homogeneous in outcomes and rewards. In this case, group-relative normalization/ranking provides little separation between trajectories, so the relative learning signal becomes less informative and can be sensitive to stochasticity without yielding systematic improvement (Shao et al., 2024; Feng et al., 2025).

- **High-conflict regime: outcome separation enables contrastive credit assignment.** Under structural mismatch, sampling is more likely to produce both better and worse rollouts with distinct reward profiles. This within-group separation makes group-relative updates informative: the optimizer can assign credit by reinforcing behaviors that lead to verified success and suppressing those leading to failure, without requiring imitation of a single fixed trace (Shao et al., 2024; Feng et al., 2025).

**3. Exploration supports selective policy shifts where imitation can be brittle.** A key advantage of routing high-conflict trajectories to RL is that exploration allows the learner to search over alternative behaviors and update the policy selectively based on feedback, rather than forcing the model to reproduce a particular trajectory. This is consistent with observations that RL post-training can induce behavioral improvements beyond SFT-only pipelines by leveraging reward-driven feedback to shape policy updates (Ouyang et al., 2022; Guo et al., 2025). In PRISM, this motivates concentrating RL budget on trajectories that exhibit structural mismatch, while using SFT to consolidate already-compatible behaviors.

## G Robustness Analysis: Architecture Invariance and Confound Control

A potential concern in gradient-based analysis is whether the varying sizes of parameter matrices (e.g., $W_{down}$ vs. $W_q$) introduce confounds in the concentration metrics. We argue that PRISM is robust to these variations due to **Architecture Invariance**.

While larger parameter matrices naturally yield larger gradient norms, this introduces a **constant systematic bias** rather than a data-dependent variable. Since the model architecture remains static, this bias affects all trajectories identically. Moreover, our chosen metrics (e.g., Gini Coefficient) are theoretically **scale-invariant**—multiplying a subset of dimensions by a constant factor preserves the relative inequality score, effectively canceling out layer-wise scaling artifacts.

To empirically validate this, we conducted a **sensitivity analysis** by normalizing the gradient norms by the square root of the parameter count ($\|\mathbf{g}\|_F / \sqrt{N_{param}}$). We observed that this normalization resulted in **nearly identical data rankings** (Spearman's $\rho > 0.99$ on both benchmarks) compared to the raw Frobenius norms. This confirms that PRISM's median-split routing is driven by genuine structural conflict rather than architectural dimensions.

## H AI Assistance Disclosure

We acknowledge the use of AI tools solely for language polishing and grammatical editing to improve the readability of this manuscript. All scientific claims, experimental data, and empirical results were rigorously verified by the human authors to ensure authenticity and accuracy.