Regime-Aware Peer Specialization for Robust RAG under Heterogeneous Knowledge Conflicts

Bo Wang **Heyan Huang** **Yaolin Li** **Yanghao Zhou**
Jiahao Teng **Ziyi Yang** **Ge Shi** **Chong Feng**
School of Computer Science, Beijing Institute of Technology

Abstract

Retrieval-augmented generation (RAG) improves language models by grounding generation in external context. However, it can be fragile when the retrieved context conflicts with the model's parametric knowledge. Such conflicts span a reliability spectrum, ranging from reliable and partially reliable evidence to adversarial context. Existing remedies often handle such heterogeneous conflicts with regime-agnostic supervision, which can conflate incompatible learning signals across reliability regimes. To disentangle these signals, we propose **RAPS-DA**, a regime-aware peer specialization framework that addresses conflict at two complementary granularities. At the *sample level*, conflicts are divided into three regimes, including **Grounding**, **Arbitration**, and **Resistance**, with one same-scale peer specialist trained per regime from a shared base model. Each sample is then hard-routed to its regime-matched peer for on-policy reverse-KL supervision. At the *token level*, a dual-layer selector uses inter-teacher disagreement, student–teacher divergence, and student entropy to filter uninformative or unstable tokens, upweight confidently misaligned ones, and gradually focus supervision on high-conflict tokens as the student matures. Gains stem from specialization at a fixed model scale, not from a stronger teacher, and the peer specialists exist only during training, so the deployed student requires no regime labels or peer access. Experiments on five conflict scenarios and two out-of-distribution benchmarks show RAPS-DA surpasses all prompting, decoding, fine-tuning, RL, and single-teacher baselines.

1 Introduction

Retrieval-augmented generation (RAG) grounds large language models in external evidence, yet its success rests on the assumption that retrieved context is reliable [13, 19]. When this assumption breaks, RAG systems face a trust–resistance dilemma: they may over-trust incorrect passages [28, 35], while resistance-oriented training can cause them to reject genuinely informative evidence [6, 32]. Generally, context reliability is a spectrum in which reliable evidence should be integrated, adversarial evidence should be rejected, and mixed-quality evidence requires selective trust. This heterogeneity creates a training challenge: *How can a single model be trained when different reliability regimes demand conflicting behaviors?*

A growing body of work has sought to address this challenge. Prompting and decoding methods [7, 8, 27] improve inference-time behavior but generalize poorly. Supervised and RL methods [5, 9, 22] apply a single training signal across mixed-reliability data, so conflicting regimes impose contradictory gradient updates. These methods either lack a trainable adaptation mechanism or provide supervision that is too coarse to resolve token-level follow-or-reject decisions. By contrast, on-policy distillation (OPD) [1, 11] offers a more suitable foundation: it provides dense token-level supervision on the student's own rollouts, directly shaping pivotal tokens where the model decides whether to follow or reject a retrieved claim [4, 10]. Yet naive OPD still uses one

teacher for all regimes, inheriting the same cross-regime interference. Despite these differences, existing approaches share a common limitation, namely regime-agnostic supervision.

Adapting OPD to knowledge conflict further exposes two coupled difficulties. **Sample-level regime heterogeneity**. We distinguish three reliability regimes in RAG (Figure 1a), namely *Grounding* (reliable context to integrate), *Arbitration* (mixed-quality context requiring selective trust), and *Resistance* (adversarial context to reject), each demanding markedly different behaviors. Empirically, models trained per-regime outperform a jointly-trained model on their corresponding subsets, revealing a clear regime-wise specialization pattern. As a result, a regime-agnostic teacher receives conflicting optimization signals from different regimes and systematically underfits them all. **Token-level supervision noise**. In conflict RAG, pivotal tokens often correspond to answer entities, source-attribution markers, rejection phrases, or option tokens where a small change determines whether the model follows CK or falls back to PK. The challenge is amplified in a multi-teacher setting, where disagreement among teachers introduces additional supervision heterogeneity [4, 10, 18]. Consequently, naively distilling from multiple specialized teachers does not reliably transfer their expertise to the student (Figure 1b).

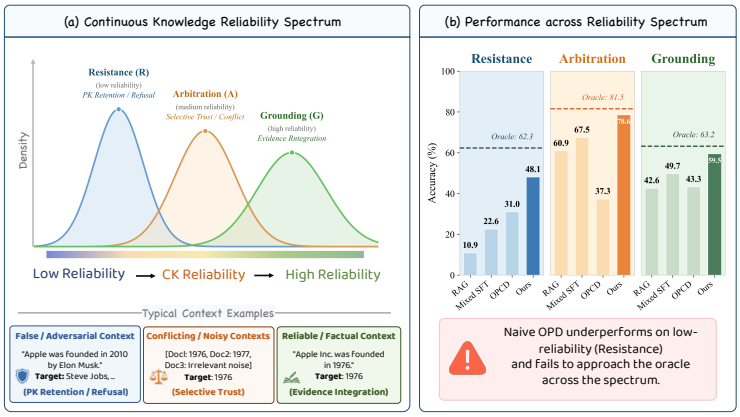

Figure 1: Motivation for regime-aware peer specialization. **(a)** Contextual-knowledge reliability is modeled as a continuous spectrum and partitioned into three regimes—Resistance (R, low reliability), Arbitration (A, medium), and Grounding (G, high)—each requiring a distinct behavior: PK retention, selective trust, or evidence integration. **(b)** Per-regime accuracy of representative methods (bars) against the oracle peer specialist upper bound (dashed lines). RAG and Mixed SFT collapse on Resistance; single-teacher OPD (OPCD) improves over them but still falls well short of the oracle across all regimes. RAPS-DA substantially narrows this gap without requiring regime labels at inference.

To address these challenges, we propose **RAPS-DA** (**R**egime-**A**ware **P**eer **S**pecialization with **D**ifficulty **A**nnealing), a training framework that handles heterogeneous knowledge conflicts through regime-aware peer supervision and difficulty-aware token learning. All peer teachers are fine-tuned from a shared base model on regime-specific data, so the improvements come from specialization at a fixed model scale rather than from using a stronger or larger teacher. The framework has three modules: regime-routed peer supervision, conflict-aware token selection, and difficulty annealing. The first two modules target sample-level regime heterogeneity and token-level supervision noise, respectively, while the third stabilizes their interaction during training. **(i) Regime-routed peer supervision** (*addresses sample-level regime heterogeneity*). We train regime-specialized peer teachers and hard-route each sample to its matched peer for on-policy reverse-KL supervision, with the base model serving as an implicit KL anchor. Hard routing eliminates the contradictory gradient updates that a single jointly-trained teacher would impose across regimes. **(ii) Conflict-aware token selection** (*addresses token-level supervision noise*). Three diagnostic signals (inter-teacher conflict, student–teacher gap, and student entropy) drive a hard mask that filters uninformative or unstable tokens, together with a soft weight that emphasizes tokens where the student is confidently misaligned with the regime-matched teacher. **(iii) Difficulty annealing** (stabilizes the combination of (i) and (ii)).

Routing concentrates supervision on regime-matched teachers, while token selection further amplifies high-conflict positions. Applying both mechanisms from the beginning can prematurely focus training on a small set of difficult tokens. We therefore start with broad token coverage and gradually concentrate supervision on harder positions as training progresses. ALL specialization and routing occur only during training. At inference time, the deployed system is a single student model that requires neither regime labels nor access to peer teachers.

Our key observation is that sample-level routing and token-level selection address complementary challenges. Routing resolves policy mismatch across reliability regimes, while token selection improves the quality of supervision within each regime. Together, however, they can over-concentrate training on difficult tokens before the student has learned stable regime-specific behaviors. Difficulty annealing mitigates this effect by gradually shifting supervision from broad coverage to high-conflict tokens. Empirically, each module helps on its own, but only their combination yields the largest gains in our experiments. Empirically, the three components are complementary, and their combination provide the most substantial benefits.

Our main contributions are as follows.

- We characterize knowledge conflict in RAG as a spectrum of reliability regimes and show that regime-agnostic supervision induces conflicting optimization signals, motivating regime-specialized supervision.
- We propose RAPS-DA, a regime-aware peer specialization framework that combines regime-routed peer supervision, conflict-aware token selection, and difficulty annealing to address heterogeneous knowledge conflicts at both sample and token levels.
- Experiments across multiple knowledge-conflict benchmarks demonstrate that the proposed components are complementary and consistently outperform RL, single-teacher, and naive multi-teacher baselines, while generalizing to held-out conflict scenarios and benchmarks.

2 Related Work

We review related literature from two perspectives, namely robust RAG under knowledge conflict and on-policy distillation with token selection.

2.1 Knowledge Conflict in RAG

RAG systems are vulnerable to noisy, outdated, or adversarial retrieved content, and the resulting knowledge conflicts manifest in diverse forms [17, 37]. Existing robustness methods span several families. *Prompting methods* guide models to verify or filter context before generation, including adaptive retrieval [8], conflict resolution prompting [33], and chain-of-verification [14]. *Decoding methods* adjust the output distribution at inference time. CAD [27] contrasts outputs with and without context, while DoLa [7] and DCCD [41] contrast logits across layers or with dynamic strategies. *Training methods* directly optimize parameters. Self-RAG [5] learns to retrieve and self-critique through reflection tokens; InFO-RAG [36] aligns information flow for noise filtering; RAAT [9] performs adversarial training against retrieval noise; KAFT [20] fine-tunes on counterfactual contexts; Astute-RAG [31] consolidates conflicting sources before answering. *RL-based methods* learn conflict-resolution policies through outcome-based optimization. Knowledgeable-R1 [22] applies GRPO [26] with asymmetric advantage modulation, showing strong adversarial results.

These methods have substantially advanced RAG robustness. Prompting and decoding approaches provide lightweight, training-free conflict mitigation; supervised methods such as RAAT and KAFT equip models with parametric resilience to noisy or counterfactual context; and RL-based methods like Knowledgeable-R1 demonstrate that outcome-level optimization can yield strong adversarial robustness. However, existing methods generally treat conflict as a single homogeneous phenomenon rather than explicitly distinguishing different reliability regimes. Prompting and decoding approaches act at inference time without altering parameters; training and RL approaches adopt a single-policy paradigm that must reconcile opposing behaviors (e.g., following vs. rejecting context) within one model. Moreover, outcome-level RL provides sparse credit assignment, offering limited signal for identifying which tokens caused errors. Our work explores an alternative direction by partitioning conflict into three reliability regimes and assigning each to a same-scale peer specialist, with dense token-level on-policy supervision providing the training signal.

2.2 On-Policy Distillation and Token Selection

On-policy distillation (OPD) provides token-level teacher feedback on the student's own rollouts, combining distribution-matching with on-policy exploration [1, 11]. Recent advances include *stability* improvements via top-K truncated reverse-KL [10] and length-inflation correction [23]; *token selection* strategies such as TIP's entropy-divergence taxonomy [4], SCOPE's correctness-based weighting [2], and SRPO's answer-correctness routing [3]; *theoretical* results showing OPD equals dense KL-constrained RL (G-OPD; 38) and establishing teacher consistency as necessary for stable training [34]; and *self-distillation* extensions including OPCD [39] and privileged-information distillation [24].

In the multi-teacher setting, classic ensemble-then-distill [15] averages teacher logits uniformly; BTX [29] trains specialized expert branches merged into an MoE deployed at inference; FuseLLM [30] fuses heterogeneous LLM distributions; CA-MKD [40] selects teachers per instance based on confidence.

Two observations motivate our design. First, recent OPD advances have made significant strides in stability and token-level credit assignment such as top-K truncation [10] tames heavy-tailed gradients, and TIP [4] shows that profiling token importance yields meaningful gains. These works, however, typically assume a single teacher or use pre-defined domain labels (e.g., math vs. code) for multi-teacher routing; organizing teachers around conflict-derived reliability regimes within a single task remains unexplored. Second, multi-teacher methods such as BTX and CA-MKD have demonstrated that instance-level teacher selection can outperform uniform ensembles. Yet existing token-selection strategies generally rely on one or two signals and apply a fixed selection policy throughout training, without adapting the difficulty of supervision as the student matures. Building on these advances, RAPS-DA routes training samples to regime-matched peer specialists via hard assignment, applies a dual-layer token selector combining three complementary signals (inter-teacher conflict, student–teacher gap, and student entropy), and introduces difficulty annealing that progressively focuses the token budget on harder positions as training proceeds.

3 Method

3.1 Overview

Given a question x and a retrieved context c, a language model π_θ generates an answer $y \sim \pi_\theta(\cdot \mid x, c)$. We categorize the reliability of c relative to the model's parametric knowledge into three regimes $r \in \mathcal{R} = \{G, A, R\}$ along a single axis, *how much could model trust context. Grounding* (G): the context is factually correct and relevant, and the model should integrate it faithfully. *Arbitration* (A): the context mixes reliable and unreliable evidence (including internally contradictory or irrelevant passages), and the model should selectively trust parts of it. *Resistance* (R): the context is adversarial or counterfactual, and the model should reject it and rely on parametric knowledge. This trichotomy is defined by the *logical relationship* between context and ground truth, independent of any particular benchmark. In Section 4.1.1 we show how existing conflict scenarios map onto these regimes. Regime labels are used only during training to assign teachers, and the deployed student infers appropriate behavior from (x, c) alone. Our goal is to train a single student π_θ that handles all three regimes, using three same-scale peer teachers $\{\pi_{T_r}\}_{r \in \mathcal{R}}$ initialized from a shared base model π_{base}.

To this end, RAPS-DA addresses the two challenges identified in Section 1 with two corresponding modules, plus a third that stabilizes their combination. **(i) Regime-routed peer supervision** (§3.2). Because a single jointly-trained teacher imposes contradictory gradient updates across regimes, we train peer teachers specialized per regime and hard-route each sample to its matched peer for on-policy reverse-KL supervision. **(ii) Conflict-aware token selection** (§3.3–3.3.3). Pivotal tokens carry disproportionate influence yet are the most susceptible to inter-teacher disagreement and student uncertainty. We thus introduce a dual-layer selector based on three diagnostic signals: inter-teacher conflict ($\mathcal{C}$), student–teacher gap ($\mathcal{G}$), and student entropy ($\mathcal{H}$). The hard mask filters uninformative or unstable tokens, while the soft weight upweights confidently misaligned ones. **(iii) Difficulty annealing** (§3.4). In early training, the student needs broad distributional coverage to build regime-consistent foundations. Restricting supervision to high-conflict tokens too early can hinder

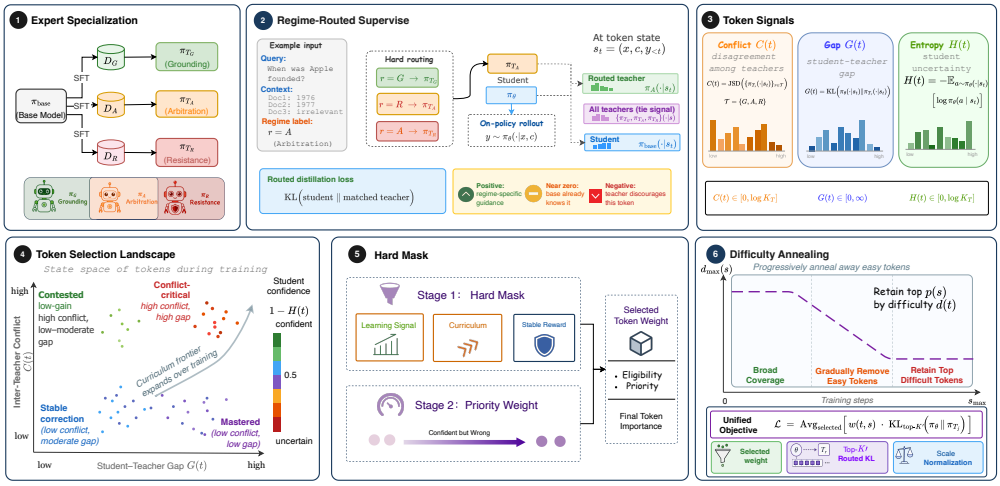

Figure 2: Overview of the RAPS-DA framework. **Panel 1:** A shared base model is fine-tuned into three regime-specialized peer teachers via SFT on the Grounding, Arbitration, and Resistance subsets. **Panel 2:** Each training sample is hard-routed by its regime label to the matched teacher; the student generates on-policy rollouts and receives routed reverse-KL supervision. **Panel 3:** Three token-level diagnostic signals—inter-teacher conflict $\mathcal{C}(t)$, student–teacher gap $\mathcal{G}(t)$, and student entropy $\mathcal{H}(t)$—are computed at each position. **Panel 4:** The token selection landscape, spanned by $\mathcal{G}(t)$ and $\mathcal{C}(t)$, partitions tokens into four regions; a curriculum frontier expands over training to cover increasingly difficult tokens. **Panel 5:** A dual-layer selector first applies a hard mask (information, curriculum, and stability filters) for eligibility, then a soft priority weight that upweights confident-but-misaligned tokens. **Panel 6:** Difficulty annealing progressively narrows the retained token set from broad coverage to the top-$p(s)$ most difficult tokens, and the unified objective aggregates the selected, weighted reverse-KL loss.

this process. An annealing schedule starts with the full token set and progressively focuses on the most difficult tokens as the student matures. Figure 2 provides an overview of the complete pipeline.

3.2 Regime-Routed Peer Supervision

Regime-specialized peer teachers. We partition the training data by the three regimes defined above and train three same-scale peer teachers from π_{base} via supervised fine-tuning on each subset $\mathcal{D}_r$:

$$\pi_{T_r} = \arg\min_{\pi} \; \mathbb{E}_{(x,c,y^*)\sim\mathcal{D}_r}\big[-\log \pi(y^* \mid x, c)\big], \tag{1}$$

where y^* is the regime-appropriate gold response (Section 4.1.1). Training a single teacher across all regimes leads to cross-regime interference (Section 4.4). Same-scale per-regime teachers isolate specialization from capacity scaling.

On-policy reverse-KL distillation. Each training sample is hard-routed to its regime-matched teacher. Unlike standard distillation that minimizes forward KL on gold trajectories, our goal is *on-policy correction*: the student generates from its own policy and the teacher corrects behavior under the prefixes the student will actually produce at deployment. We accordingly use on-policy rollouts so that the student is exposed to its own mistakes, and adopt reverse KL to encourage mode-seeking behavior that concentrates mass on tokens the routed teacher considers likely while suppressing those it deems unlikely. The per-sample on-policy reverse-KL objective is

$$\mathcal{L}_r(\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x,c)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\text{KL}}\big(\pi_\theta(\cdot \mid \mathbf{s}_t) \,\big\|\, \pi_{T_r}(\cdot \mid \mathbf{s}_t)\big)\right], \tag{2}$$

where $\mathbf{s}_t = (x, c, y_{<t})$ denotes the prefix state at position t.

From distribution KL to token-level implicit reward. To expose the per-token learning signal hidden inside Eq. equation 2, we follow the G-OPD framework [38] and rewrite the reverse KL by introducing the base model π_{base} as a reference distribution. Expanding the KL definition and adding-and-subtracting $\log \pi_{\text{base}}$ inside the logarithm gives

$$\begin{aligned} D_{\text{KL}}(\pi_\theta \| \pi_{T_r}) &= \mathbb{E}_{\pi_\theta}\left[\log \frac{\pi_\theta}{\pi_{T_r}}\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\log \frac{\pi_\theta}{\pi_{\text{base}}}\right] - \mathbb{E}_{\pi_\theta}\left[\log \frac{\pi_{T_r}}{\pi_{\text{base}}}\right] \\ &= D_{\text{KL}}(\pi_\theta \| \pi_{\text{base}}) - \mathbb{E}_{\pi_\theta}\left[r_t^{(r)}\right], \end{aligned} \tag{3}$$

where each sampled token a_t receives a dense *implicit reward*

$$r_t^{(r)} = \log \frac{\pi_{T_r}(a_t \mid \mathbf{s}_t)}{\pi_{\text{base}}(a_t \mid \mathbf{s}_t)}. \tag{4}$$

Substituting Eq. equation 3 into the per-sample loss yields the equivalent KL-constrained reward maximization

$$\min_\theta \ \mathcal{L}_r(\theta) \iff \max_\theta \left\{ \mathbb{E}_{\pi_\theta}\left[r_t^{(r)}\right] - D_{\text{KL}}(\pi_\theta \| \pi_{\text{base}}) \right\}. \tag{5}$$

Two properties follow. (i) π_{base} serves as an *implicit KL anchor*, providing regularization without an explicit reference term in the loss. (ii) Each token carries a dense, regime-specific learning signal $r_t^{(r)}$ that reflects how much the routed teacher agrees with the base at that position. Unlike RL's sparse outcome-level reward, this dense signal enables the dual-layer token selector in Section 3.3 to operate on a principled reward rather than on raw KL contributions. The full training objective aggregates Eq. equation 2 across regimes via hard routing, so that each sample activates exactly one teacher with no distribution mixing.

3.3 Token-Level Dual-Layer Selection

While regime routing assigns an appropriate teacher to each sample, supervision quality can still vary substantially across tokens. Some tokens provide little learning signal because the student is already aligned with the teacher [4, 18]. Others produce high-variance gradients that destabilize training [18]. Still others lie in high-conflict regions where teachers disagree. We therefore separate token selection into two stages: a *hard eligibility mask* $M(t)$ that removes tokens likely to inject harmful or premature supervision, and a *soft weight* $\omega(t, s)$ that assigns graded emphasis among the surviving tokens.

3.3.1 Token-Level Supervision Signals

No single signal can reliably distinguish informative tokens from uninformative or destabilizing ones, because inter-teacher consistency, student–teacher divergence, and student confidence jointly affect whether a token provides useful supervision. We therefore introduce three complementary diagnostic signals, all normalized to $[0, 1]$ before use.

Inter-teacher conflict $\mathcal{C}(t)$. This signal measures how much the K peer teachers disagree at a given token, defined as the Jensen–Shannon divergence:

$$\mathcal{C}(t) = H\left(\frac{1}{K}\sum_{k=1}^{K} \pi_{T_k}(\cdot|\mathbf{s}_t)\right) - \frac{1}{K}\sum_{k=1}^{K} H(\pi_{T_k}(\cdot|\mathbf{s}_t)), \tag{6}$$

where $H(\cdot)$ denotes entropy and the mixture is over all $K = 3$ teachers. $\mathcal{C}(t) = 0$ when all teachers agree; large $\mathcal{C}(t)$ indicates a token at the core of knowledge conflict. Note that while supervision comes exclusively from the routed teacher, $\mathcal{C}(t)$ queries all teachers to assess disagreement. For efficiency, $\mathcal{C}(t)$ is approximated on the union of each teacher's top-K' tokens and refreshed every N training steps.

Student–teacher gap $\mathcal{G}(t)$. This signal measures how far the student deviates from the routed teacher at each position, defined as the per-token reverse-KL:

$$\mathcal{G}(t) = D_{\mathrm{KL}}\big(\pi_\theta(\cdot|\mathbf{s}_t) \,\big\|\, \pi_{T_r}(\cdot|\mathbf{s}_t)\big)\,, \tag{7}$$

$\mathcal{G}(t) \approx 0$ means the student has already aligned with the teacher at this position. Large $\mathcal{G}(t)$ indicates that corrective supervision is still needed.

Student entropy $\mathcal{H}(t)$. Inspired by the "confident-but-misaligned" blind spot identified by Anonymous [4], we introduce a normalised entropy signal to capture the student's predictive uncertainty at position t:

$$\mathcal{H}(t) = \frac{H(\pi_\theta(\cdot|\mathbf{s}_t))}{\log|\mathcal{V}|} \in [0,1], \tag{8}$$

where $|\mathcal{V}|$ is the vocabulary size. Low $\mathcal{H}(t)$ indicates a confident student, which is desirable when correct but dangerous when misaligned with the teacher.

3.3.2 Hard Mask: Information and Stability Filters

Not all tokens within a regime-routed sample provide useful supervision. We use a hard eligibility mask to control two risks. First, tokens where the student has already aligned with the teacher carry negligible gradient signal and waste computation. Second, tokens with extreme negative implicit rewards under high student uncertainty arise from off-manifold rollouts where teacher supervision is unreliable and can destabilize gradients. A token is eligible only if both conditions hold:

$$M(t) = \mathbb{1}[\mathcal{G}(t) \geq \epsilon_{\mathcal{G}}(s)] \cdot \mathbb{1}\Big[\neg\big(r_t^{(r)} < r_{\min} \;\wedge\; \mathcal{H}(t) > h_{\max}\big)\Big]\,, \tag{9}$$

where the first factor is the *information filter* and the second is the *stability filter*, described below.

Information filter. Tokens with $\mathcal{G}(t) < \epsilon_{\mathcal{G}}(s)$ have already been learned and carry negligible gradient signal. The threshold is set as an adaptive percentile computed per regime within each batch, since gap distributions differ substantially across regimes (e.g., Grounding tokens tend to have smaller gaps than Resistance tokens):

$$\epsilon_{\mathcal{G}}(s) = Q_\beta(\{\mathcal{G}(t) : r_i = r\}_{t\in\mathrm{batch}})\,,$$

where $Q_\beta(\cdot)$ denotes the β-th quantile. This filters the bottom $\beta\%$ of tokens by learning gap within each regime in each batch, preventing systematic over-filtering of regimes with inherently smaller gaps while adapting automatically as the student improves..

Stability filter. Not all tokens with negative implicit rewards should be discarded. A token where the student is confident yet strongly disfavored by the regime teacher ($r_t^{(r)} \ll 0$, low $\mathcal{H}$) is precisely a confident-but-misaligned prediction that the soft weight should later prioritize for correction. Removing it would eliminate the most valuable corrective signal. We therefore filter only tokens that are both extremely disfavored and associated with high student uncertainty ($\mathcal{H}(t) > h_{\max}$), which typically arise from off-manifold student rollouts where teacher supervision is unreliable. Following the mixture-based lower-bound derivation of Ko et al. [18], the reward threshold is set as $r_{\min} = \log\lambda_{\mathrm{clip}}/(1-\lambda_{\mathrm{clip}})$, where λ_{clip} is the mixture coefficient that controls the clipping aggressiveness 18, $h_{\max}$ is a student-entropy ceiling. When the student is confident (low $\mathcal{H}$), the stability condition is automatically satisfied regardless of the reward, ensuring that confident mismatches are retained for correction.

3.3.3 Soft Weight for Correction Prioritisation

Among the tokens that pass the hard mask, not all surviving tokens deserve equal attention. The most valuable corrective signal comes from tokens where the student is confident yet far from the routed teacher, i.e., the "confident-but-misaligned" blind spot identified by Anonymous [4]. These tokens represent cases where the student has committed to an incorrect direction and ordinary gradient updates may be insufficient to reverse the commitment. In contrast, tokens where the student is already uncertain or close to the teacher need only standard-weight supervision.

We therefore introduce a soft weight that assigns graded priority among eligible tokens:

$$\omega(t,s) = 1 + \eta \cdot (1 - \mathcal{H}(t)) \cdot \mathcal{G}(t), \tag{10}$$

where η controls the boost magnitude. When the student is uncertain, i.e. $\mathcal{H}(t)$ is large, or already close to the teacher, i.e. $\mathcal{G}(t)$ is small, the weight reduces to $\omega \approx 1$ and standard supervision applies.

The hard mask eliminates tokens that would waste computation or destabilize training, while the soft weight fine-tunes priority among the remaining informative tokens.

3.4 Difficulty Annealing

Routing and token selection together concentrate supervision on high-signal, high-conflict tokens. In early training, however, the student has not yet acquired regime-consistent foundations, so restricting supervision to only high-conflict tokens from the start would deprive it of the broad distributional coverage needed to build basic competence. A natural remedy is to begin with the full token set and then progressively anneal away easy tokens, gradually focusing the training budget on the most informative, high-conflict positions as the student matures.

The difficulty score combines inter-teacher conflict with a confidence interaction term:

$$d(t) = \mathcal{C}(t) + \gamma \cdot \mathcal{C}(t) \cdot (1 - \mathcal{H}(t)), \tag{11}$$

where γ controls the interaction strength. The first term captures raw conflict intensity. The interaction term adds difficulty when the student is both in a high-conflict region and highly confident there, identifying tokens where the student may have committed to an incorrect direction. The annealing schedule controls how many tokens survive via a retention ratio $p(s)$ that decreases linearly from 1 to a final value p_{final} over an annealing horizon S_{a}, and remains constant thereafter:

$$p(s) = \begin{cases} 1 - (1 - p_{\text{final}}) \, \dfrac{s}{S_{\text{a}}}, & s < S_{\text{a}}, \\ p_{\text{final}}, & s \geq S_{\text{a}}. \end{cases} \tag{12}$$

A token is retained by the annealing mask if its difficulty ranks in the top $p(s)$ fraction within the sequence:

$$A(t,s) = \mathbb{1}\big[d(t) \geq Q_{1-p(s)}\big(\{d(t')\}_{t'}\big)\big] \, . \tag{13}$$

At the start of training $p(s)$=1 and all tokens pass the annealing filter. By step S_{a}, only the top p_{final} fraction of difficult tokens is retained. This transitions from broad coverage to focused, high-conflict supervision at the token level, rather than a coarser regime-level ordering.

Note that the annealing and soft weighting interact by design. The annealing gradually removes easy tokens whose learning signal has diminished, while the soft weight prioritizes the surviving high-conflict, confident-but-misaligned tokens for correction. The annealing parameters (p_{final}, S_{a}) are set once and shared across all regimes.

3.5 Unified Training Objective

Combining all components, we first define the effective per-token weight:

$$w(t,s) = M(t) \cdot A(t,s) \cdot \omega(t,s), \tag{14}$$

which folds together the hard mask, annealing mask and the soft correction-priority weight. The full RAPS-DA training objective is then

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,c,r)\sim\mathcal{D}} \, \mathbb{E}_{y\sim\pi_\theta} \left[\frac{1}{Z} \sum_{t=1}^{|y|} w(t,s) \cdot D_{\text{KL}}^{K'}\big(\pi_\theta(\cdot|\mathbf{s}_t) \,\big\|\, \pi_{T_r}(\cdot|\mathbf{s}_t)\big) \right], \tag{15}$$

where $Z = \max(\sum_t w(t,s),\ Z_{\min})$ normalises by the effective token count and is clipped from below to avoid high-variance gradients when few tokens survive. $D_{\text{KL}}^{K'}$ denotes the top-K' truncated reverse KL [10], which restricts the computation to the union of the teacher's and student's top-K' tokens with renormalization, preserving the reverse-KL penalty on out-of-support student assignments via a residual bucket. Algorithm 1 summarizes the complete training procedure.

Algorithm 1 RAPS-DA training (one step).

Input: student π_θ, regime teachers $\{\pi_{T_r}\}$, base π_{base}, batch $\mathcal{B}$ with regime labels, step s.

`// Stage 1: Annealing schedule`

Compute retention ratio $p(s)$ from annealing schedule.

`// Stage 2: On-policy rollout and signal computation`

for each $(x_i, c_i, r_i) \in \mathcal{B}$ **do**

  Generate rollout $y_i \sim \pi_\theta(\cdot \mid x_i, c_i)$.

  Compute per-token signals $\mathcal{G}(t)$, $\mathcal{H}(t)$, and $r_t^{(r)}$.

  Refresh inter-teacher conflict $\mathcal{C}(t)$ every N steps.

end for

`// Stage 3: Token selection`

Batch-normalize $\mathcal{C}$, $\mathcal{G}$ to $[0, 1]$.

Form hard mask $M(t)$ via information and stability filters.

Compute difficulty $d(t)$ and retain top-$p(s)$ fraction as annealing mask $A(t, s)$.

Form soft weight $\omega(t, s) \leftarrow 1 + \eta \cdot (1 - \mathcal{H}(t)) \cdot \mathcal{G}(t)$.

`// Stage 4: Gradient update`

Compute $\mathcal{L}(\theta)$ with effective weight $w(t, s) = M(t) \cdot A(t, s) \cdot \omega(t, s)$.

Update $\theta \leftarrow \theta - \text{lr} \cdot \nabla_\theta \mathcal{L}(\theta)$.

4 Experiments

We evaluate RAPS-DA across multiple knowledge-conflict scenarios, examining whether regime-aware peer specialization yields a single student that appropriately leverages or disregards retrieved context depending on its reliability, without relying on a stronger or larger teacher.

4.1 Experimental Setup

4.1.1 Datasets

In-distribution. We adopt the five knowledge-conflict scenarios from the Knowledgeable-R1 (KR1) benchmark [22], each requiring a distinct behavior from the model: S1 (correct context) provides accurate supporting evidence that should be faithfully integrated; S2 (adversarial context) contains deliberately counterfactual passages that should be rejected; S3 (self-conflicting context) presents internally contradictory claims requiring arbitration; S4 (irrelevant context) includes topically unrelated passages that should be ignored; and S5 (partially relevant context) mixes valid evidence with distractors, demanding selective extraction. For our three-regime formulation ($\mathcal{R} = \{G, A, R\}$), we group these scenarios by the dominant behavior required. S1 and S5 map to Grounding (G), where the model should leverage reliable or partially relevant context. S3 and S4 map to Arbitration (A), where the model must arbitrate among conflicting or irrelevant evidence. S2 maps to Resistance (R), where the model must reject misleading evidence.

Out-of-distribution. To test whether the learned conflict-resolution policy generalizes beyond the training distribution, we evaluate on two held-out benchmarks not used in any training stage: **(1) ConflictQA** [28] (PopQA subset, 7,947 samples) pairs each factual question with a parametric-consistent (PC) instance whose context supports the correct answer and a non-consistent (NC) instance whose context contains counterfactual evidence. Comparing PC vs. NC accuracy directly measures conflict robustness. **(2) TruthfulQA** [21] (817 questions) probes whether the model retains truthful parametric knowledge. We evaluate in a no-context setting without retrieved passages to test whether training preserves the base model's parametric accuracy.

4.1.2 Evaluation Metrics

We adopt exact match (EM) as the primary metric, consistent with prior work [16, 22]. For each scenario, EM is computed on the respective test or validation split; we additionally report unweighted

regime-level averages ($G/A/R$) and overall averages. For the OOD benchmarks, we report EM on ConflictQA under the RAG-augmented setting and EM on TruthfulQA under the no-context setting (testing parametric-knowledge preservation).

4.1.3 Baselines

We compare against the following methods, spanning prompting, decoding, fine-tuning, and distillation approaches:

Inference-time methods. **Query-only prompting** uses only parametric knowledge without retrieved context, serving as a lower bound for context-dependent scenarios and an upper bound when context is misleading. **RAG prompting** prepends the top-5 retrieved passages to the query without additional training. **Astute-RAG** [31] consolidates and clusters conflicting sources before answering. **CK-PLUG** [6] is a plug-and-play decoding strategy that adjusts token probabilities at detected conflict spans.

Training-based methods. **Mixed SFT** performs supervised fine-tuning on the union of all regime-specific data, using the same total number of training samples as RAPS-DA.

RL methods. **GRPO w/ RAG** [12] applies group relative policy optimization with outcome-level reward on RAG rollouts. **Knowledgeable-R1** [22] extends GRPO with joint PK/CK/RPK sampling and adaptive advantage modulation, representing the state-of-the-art on the benchmark.

OPD methods. **OPCD** [39] conditions the teacher on context while the student generates without it, transferring contextual reasoning into parametric knowledge. Since OPCD uses a single jointly-trained teacher with the same OPD hyperparameters as RAPS-DA, it also serves as our single-teacher OPD baseline in the ablation study. **TIP** [4] profiles token importance and up-weights informative tokens via a learned scorer.

4.1.4 Implementation Details

We adopt `Qwen2.5-7B-Instruct`[1] [25] as the shared base model for both the student and all peer teachers. This same-scale (7B→7B) setting isolates the effect of regime-aware specialization from any teacher capacity advantage. All three peer teachers are initialized from the same checkpoint and fine-tuned independently on their regime-specific subsets (Section 3.2). For the Llama baseline in Table 1, we use `Llama3.1-8B-Instruct` from `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`.

For OPD training, we use AdamW with a learning rate of 1×10^{-6} and linear warmup over the first 10% of steps. Each batch consists of 64 prompts with 4 rollouts each via top-p sampling (p=0.9, τ=1.0). The token-level reverse KL uses the teacher's top-K'=32 tokens with renormalization [10]: concretely, let $\mathcal{S}_t = \mathrm{TopK}'(\pi_\theta(\cdot|\mathbf{s}_t)) \cup \mathrm{TopK}'(\pi_{T_r}(\cdot|\mathbf{s}_t))$ and define $\tilde{\pi}(v) = \pi(v)$ for $v \in \mathcal{S}_t$ and $\tilde{\pi}(\textsc{rest}) = \sum_{u \notin \mathcal{S}_t} \pi(u)$; the KL is computed over $\mathcal{S}_t \cup \{\textsc{rest}\}$. The inter-teacher conflict signal $\mathcal{C}(t)$ is refreshed every N=50 training steps; intermediate steps reuse the cached estimate. Special tokens are excluded from all signal computations and from the loss. The base model π_{base} serves as the implicit KL anchor. Training proceeds for 2 epochs on 8 NVIDIA H100 80GB GPUs in 2 hours. All results are averaged over three random schedules.

4.2 Main Results

Table 1 reports per-scenario exact-match accuracy and regime-level averages across both backbone models. We compare RAPS-DA against prompting, decoding, fine-tuning, and RL baselines to assess whether regime-aware peer specialization yields consistent improvements, and include the oracle peer specialist as a non-deployable upper bound that requires ground-truth regime labels at inference.

[1] `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`

Table 1: Main in-distribution results (EM, %) following the KR1 protocol [22]. Best in **bold**, second-best underlined; Oracle PS is excluded from ranking. Shaded rows are our methods.

	S1 Correct (G)			S2 Wrong (R)			S3 (A)	S4 (A)	S5 Partly-Irr. (G)			Regime Avg.		
Method	PC-MR	PC-MC	PC-QA	NC-MR	NC-MC	NC-QA	SC	ExpPE	HotPot	2Wiki	Musique	G	A	R
`Qwen2.5-7B-Instruct`														
Query-only	27.7	24.7	31.7	25.9	25.8	32.3	29.7	64.5	20.9	25.5	4.4	22.5	47.1	28.0
RAG prompting	65.7	66.4	74.3	13.5	8.1	11.3	59.5	62.2	20.4	22.5	6.4	42.6	60.9	10.9
CK-PLUG [6]	64.7	66.5	78.7	11.6	8.1	7.9	55.0	55.0	22.7	24.8	6.2	43.9	55.0	9.2
Astute [31]	65.5	66.0	77.6	12.8	7.1	10.3	54.2	56.7	17.9	20.4	6.3	42.3	55.5	10.1
SFT	72.0	77.7	74.7	24.9	21.1	22.0	68.5	66.6	30.1	32.2	11.8	49.7	67.5	22.6
GRPO w/ RAG [12]	77.6	77.4	80.0	26.9	19.7	26.0	75.3	66.5	27.9	34.0	11.8	51.4	70.9	24.2
KR1 [22]	75.1	75.5	**80.9**	43.9	37.3	29.4	76.3	67.6	31.4	37.5	12.0	52.1	72.0	36.9
OPCD [39]	64.7	64.4	62.8	40.2	37.0	15.8	56.5	18.1	24.6	34.5	8.8	43.3	37.3	31.0
TIP [4]	81.8	79.1	76.8	33.3	36.5	41.0	76.2	52.9	31.7	44.5	15.2	54.9	64.5	36.9
RAPS-DA (ours)	**89.3**	**87.0**	**80.9**	**51.5**	**48.2**	**44.6**	**87.3**	**69.9**	**33.9**	**45.5**	**20.1**	**59.5**	**78.6**	**48.1**
`Llama3.1-8B-Instruct`														
Query-only	29.4	26.2	39.9	27.1	27.6	42.6	32.1	43.3	20.7	21.0	6.2	23.9	37.7	32.5
RAG prompting	64.8	62.0	76.4	22.9	16.3	24.9	61.2	39.2	24.4	23.5	8.2	43.2	50.2	21.4
CK-PLUG	54.8	58.5	69.7	12.1	9.1	17.3	42.0	31.5	22.4	24.6	5.2	39.2	36.8	12.8
Astute	65.8	64.9	78.0	17.0	9.1	17.3	59.8	40.1	1.6	30.3	9.6	41.7	50.0	14.4
SFT	72.9	79.2	73.8	42.6	35.5	35.9	70.1	47.2	33.6	38.2	13.4	51.9	58.6	38.0
GRPO w/ RAG	78.0	79.7	82.6	41.6	35.7	39.3	76.6	47.6	34.8	41.2	16.6	55.5	62.1	38.8
KR1	73.8	80.2	80.0	55.4	41.1	44.6	73.7	49.6	37.1	45.4	14.7	55.2	61.6	47.0
OPCD [39]	61.8	59.5	70.2	46.7	40.4	32.9	54.2	25.6	27.8	38.7	8.5	44.4	39.9	40.0
TIP [4]	75.4	78.1	45.8	62.5	**58.9**	41.3	65.5	36.4	35.2	43.7	15.2	48.9	51.0	54.2
RAPS-DA (ours)	**79.5**	**83.7**	**84.4**	**67.1**	52.9	**55.7**	**77.8**	**52.6**	**40.2**	**49.5**	**18.3**	**59.2**	**65.2**	**58.6**

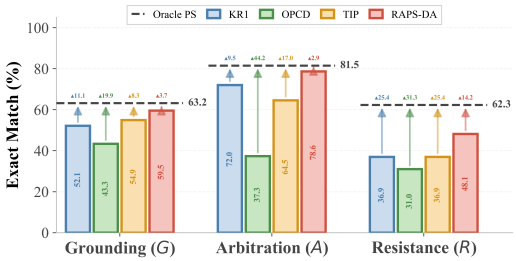

Figure 3: Regime-level performance comparison (Qwen-7B). Bars show the regime-averaged EM of four representative methods; the dashed line marks the oracle peer specialist upper bound. ▲ values indicate the remaining gap to the oracle. RAPS-DA achieves the smallest gap across all three regimes.

RAPS-DA outperforms all baseline categories without trading off across regimes. Prompting and decoding methods (Astute-RAG, CK-PLUG) lack training signal for conflict resolution and fail on adversarial context, with Resistance averages below 11% on Qwen. SFT and RL methods (Mixed SFT, GRPO, KR1) improve Resistance substantially through conflict-aware training objectives but share a single policy for all regimes, resulting in cross-regime interference that limits their Grounding average to around 50%. OPD methods (OPCD, TIP) learn from teacher demonstrations and achieve competitive Grounding performance, but a single generalist teacher cannot specialize to all three regimes simultaneously, limiting their Resistance average. RAPS-DA addresses these limitations by combining regime-specialized peer teachers with token-level selection. On the Qwen backbone, RAPS-DA achieves regime averages of 59.5% (G), 78.6% (A), and 48.1% (R), improving over KR1 by 7.4, 6.6, and 11.2 points and over TIP by 4.6, 14.0, and 11.2 points, respectively. The largest margin in both comparisons appears on Resistance, where the model must reject adversarial context entirely, confirming that regime-specialized teachers provide the greatest benefit when the required behavior diverges most from standard RAG. Crucially, unlike baselines that trade Grounding accuracy for Resistance robustness, RAPS-DA improves both ends of the reliability spectrum simultaneously. On S1 (correct context), accuracy reaches 89.3% and 87.0% on PC-MR and PC-

Table 2: Out-of-distribution evaluation on ConflictQA (EM, %) and TruthfulQA (EM, %, no-context). Neither benchmark participates in any training stage. Best in **bold**, second best underlined.

Method	ConflictQA	TruthfulQA
RAG prompting	34.3	4.4
Astute-RAG [31]	16.9	**5.4**
CK-PLUG [6]	7.9	0.0
Mixed SFT	39.7	1.8
GRPO w/ RAG	36.3	4.2
Knowledgeable-R1 [22]	36.5	4.2
OPCD [39]	40.0	3.7
TIP [4]	38.4	4.2
RAPS-DA (ours)	**42.7**	4.4

MC, while on S2 (adversarial context) it improves to 51.5%, 48.2%, and 44.6%, demonstrating that learning to reject misleading context does not reduce the ability to leverage correct context.

Regime-specialized OPD narrows the gap to the oracle upper bound. Figure 3 visualizes the regime-level comparison against the oracle peer specialist, a non-deployable upper bound that requires ground-truth regime labels at inference. RAPS-DA closes 69.6% of the gap between KR1 and the oracle on Arbitration, and even surpasses the oracle on Musique within the Grounding regime, suggesting that cross-regime knowledge transfer can compensate for the oracle's lack of inter-regime information sharing. Among OPD baselines, TIP achieves competitive Grounding through importance-weighted token selection but remains limited on Resistance due to its single generalist teacher. OPCD, which also serves as our single-teacher OPD ablation (Table 3), underperforms substantially on Arbitration and Resistance, confirming that a jointly-trained teacher systematically underfits every regime. These results indicate that the gains originate from regime-specialized supervision quality rather than from model scale or ensemble effects.

4.3 Out-of-Distribution Generalization

We next evaluate whether the conflict-resolution strategy learned by RAPS-DA transfers to held-out benchmarks not seen during teacher training or student training. ConflictQA tests robustness under parametric–contextual conflict; TruthfulQA tests parametric-knowledge preservation in a no-context setting, verifying that training does not erode the base model's truthfulness. Table 2 reports results on both benchmarks.

RAPS-DA achieves the highest ConflictQA EM of 42.7%, outperforming all baselines including OPCD (40.0%) and TIP (38.4%), while maintaining 4.4% on TruthfulQA, matching the base model and confirming that regime-specialized training does not erode parametric knowledge. Notably, OPCD attains comparable ConflictQA performance but drops to 3.7% on TruthfulQA, suggesting that its context-conditioned teacher transfers conflict-handling ability at the cost of parametric knowledge retention. Two edge cases merit discussion. CK-PLUG degenerates to 0.0% on TruthfulQA because the no-context setting leaves its confidence-gain signal undefined. Astute-RAG achieves the highest TruthfulQA score of 5.4% but the lowest ConflictQA score among training-based methods at 16.9%, as its prompting-based consolidation lacks the learned conflict resolution needed for parametric–contextual conflicts.

4.4 Component and Motivation-Validating Ablations

The ablations in this section serve two complementary purposes. First, they verify that each component of RAPS-DA is necessary by removing it and measuring the resulting performance drop. Second, they validate the underlying motivations that justify the framework's design, including the presence of cross-regime interference under a single teacher, the complementary roles of sample-level routing and token-level selection, the necessity of difficulty annealing over static masking, and the individual contributions of the three diagnostic signals. Table 3 summarizes all motivation-validating variants and component removals; subsequent subsections analyze each finding in detail.

Table 3: Motivation-validating ablations on Qwen2.5-7B-Instruct. Each row removes or replaces a single design choice while keeping all other components and hyperparameters fixed. The upper block validates our core design motivations; the lower block isolates individual selector components. Best results are in **bold**; shaded row denotes the full method.

	S1 Correct (G)			**S2 Wrong** (R)			**S3** (A)	**S4** (A)	**S5 Partly-Irr.** (G)			**Regime Avg.**		
Variant	PC-MR	PC-MC	PC-QA	NC-MR	NC-MC	NC-QA	SC	ExpPE	HotPot	2Wiki	Musique	G	A	R
RAPS-DA (full)	**89.3**	**87.0**	**80.9**	**51.5**	**48.2**	**44.6**	**87.3**	**69.9**	**33.9**	**45.5**	**20.1**	**59.5**	**78.6**	**48.1**
Motivation-validating variants														
Single-teacher OPD (= OPCD)	64.7	64.4	62.8	40.2	37.0	15.8	56.5	18.1	24.6	34.5	8.8	43.3	37.3	31.0
Multi-teacher uniform pooling	80.7	78.4	84.7	15.3	10.4	19.2	67.4	55.3	33.5	44.4	11.1	55.5	61.4	15.0
Routing only	83.3	80.7	85.2	24.2	19.1	34.9	75.1	65.9	33.5	45.9	13.8	57.1	70.5	26.1
Token selection only	77.4	75.5	79.0	21.0	14.1	25.8	65.2	64.1	30.9	35.0	13.1	51.8	64.7	20.3
Static masking	85.5	84.3	83.0	40.7	40.5	41.2	82.2	69.9	32.9	43.9	15.1	57.4	76.1	40.8
Component removals														
− Soft weight ω	88.3	87.5	84.3	41.8	39.0	44.6	87.2	69.8	33.0	44.8	19.4	59.6	78.5	41.8
− Information filter	89.1	86.8	84.2	45.6	41.0	42.0	86.1	72.4	33.6	45.5	18.2	59.6	79.2	42.9
− Stability filter	88.4	87.7	81.1	46.8	40.0	43.6	87.1	69.1	32.8	43.6	18.1	58.6	78.1	43.5

Table 4: Cross-regime evaluation matrix. Each row corresponds to a teacher trained exclusively on one regime; columns indicate the evaluation regime. Diagonal entries (bold) reflect specialization, while off-diagonal drops quantify cross-regime interference.

Trained on \ Eval on	G	A	R	Avg
Specialist on G	**60.5**	59.7	11.0	43.7
Specialist on A	57.7	**77.7**	30.5	55.3
Specialist on R	35.8	53.2	**47.2**	45.4

Several observations emerge from Table 3. *Single-teacher OPD*, equivalent to OPCD in Table 1, removes both regime routing and token selection. It suffers the largest overall degradation, with Arbitration dropping from 78.6% to 37.3% and Resistance from 48.1% to 31.0%. This confirms that a single jointly-trained teacher cannot provide effective supervision across heterogeneous conflict regimes. *Multi-teacher uniform pooling* recovers Grounding performance to 55.5% by averaging multiple teacher signals, but catastrophically collapses on Resistance to 15.0%, indicating that uniformly mixing teacher distributions re-introduces cross-regime interference at the token level. Comparing *Routing only* and *Token selection only* reveals that the two mechanisms are complementary rather than substitutable. Routing alone improves Arbitration to 70.5% but leaves Resistance at 26.1%, while token selection alone achieves 64.7% on Arbitration but only 20.3% on Resistance. Neither mechanism in isolation approaches the full model. *Static masking*, which retains a fixed set of tokens without annealing, achieves 40.8% on Resistance compared to 48.1% for the full model. The 7.3-point gap confirms that the progressive easy-to-hard schedule contributes meaningfully beyond a fixed selection budget. Among the component removals, the soft weight ω has the largest impact on Resistance, reducing it by 6.3 points. This is consistent with its role in prioritizing confidently misaligned tokens, which are most prevalent in the adversarial regime.

4.4.1 Cross-Regime Interference and the Need for Multiple Teachers

A central design motivation for RAPS-DA is that a single model struggles to simultaneously integrate reliable context, arbitrate mixed evidence, and reject adversarial passages, because these behaviors impose partially contradictory optimization pressures. To make this interference empirically visible, we train three regime-specialized peer teachers in isolation and evaluate each teacher on all three regimes. Table 4 reports the resulting cross-regime evaluation matrix.

The matrix exhibits a pronounced diagonal structure. The Grounding specialist achieves 60.5% on its own regime but collapses to 11.0% on Resistance, confirming that optimizing for context integration directly conflicts with context rejection. Conversely, the Resistance specialist attains 47.2% on adversarial samples but drops to 35.8% on Grounding, where its rejection bias suppresses useful evidence. The Arbitration specialist reaches the highest single-regime score of 77.7% but still drops substantially on Resistance (30.5%), further illustrating that no single teacher can cover all regimes. These results validate the core premise that a single supervisory policy cannot simultaneously serve

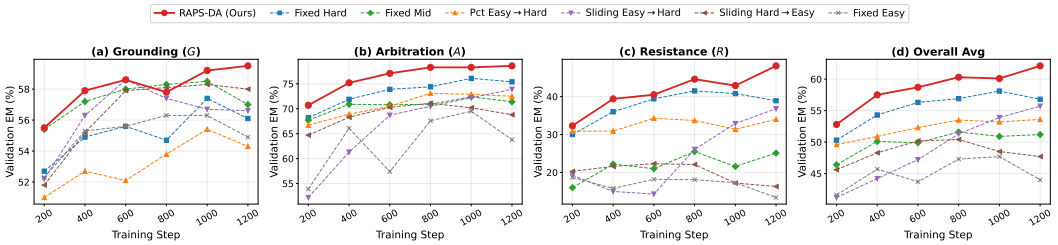

Figure 4: Training dynamics under different annealing schedules. Each panel reports validation EM over training steps for one regime or the overall average. RAPS-DA with linear annealing (solid orange) reaches the highest final plateau across all regimes. On Resistance, Sliding Hard→Easy collapses to 16.3% and Fixed Easy to 13.4%, confirming that the easy-to-hard ordering is essential for learning adversarial context rejection.

Table 5: Model-scale generalization on Qwen2.5-Instruct at 3B and 14B. EM (%) is reported for all five scenarios and three regime averages. KR1 baseline numbers are reproduced from Lin et al. [22]; "—" denotes results not reported in the original paper. The 7B results appear in Table 1. Shaded rows denote our method.

	S1 Correct (G)			**S2 Wrong** (R)			**S3** (A)	**S4** (A)	**S5 Partly-Irr.** (G)			**Regime Avg.**		
Method	PC-MR	PC-MC	PC-QA	NC-MR	NC-MC	NC-QA	SC	ExpPE	HotPot	2Wiki	Musique	G	A	R
`Qwen2.5-3B-Instruct`														
Query-only	16.8	18.4	23.9	16.0	18.3	22.6	20.8	53.3	12.6	11.3	2.1	14.2	37.0	19.0
RAG prompting	55.1	52.9	59.5	13.5	8.7	6.8	51.2	42.2	14.7	13.0	3.6	33.1	46.7	9.7
GRPO w/ RAG	71.0	70.6	80.0	19.9	15.1	19.4	66.7	53.6	24.8	35.6	9.1	48.5	60.1	18.1
KR1	66.2	59.8	78.7	28.8	28.6	22.0	—	54.7	—	—	—	—	—	26.5
RAPS-DA (ours)	74.8	73.5	83.2	35.6	33.3	28.7	69.8	57.4	28.4	39.9	12.1	52.0	63.6	32.5
`Qwen2.5-14B-Instruct`														
Query-only	30.0	26.5	35.8	30.6	28.9	39.4	30.9	70.8	25.4	26.7	5.7	25.0	50.9	33.0
RAG prompting	63.2	63.9	73.3	22.2	13.0	28.1	61.9	70.5	22.9	22.5	6.9	42.1	66.2	21.1
GRPO w/ RAG	73.6	74.7	78.9	38.9	31.7	36.5	74.0	—	34.7	39.0	14.7	—	—	35.7
KR1	70.6	75.2	78.8	47.8	33.1	38.1	72.5	—	37.0	42.3	14.6	—	—	39.7
RAPS-DA (ours)	84.4	85.7	88.2	60.3	55.8	42.6	78.5	75.3	38.5	47.8	17.6	60.4	76.9	52.9

all conflict regimes, and that regime-specific teachers are necessary to eliminate cross-regime interference.

4.4.2 Why Difficulty Annealing Matters

The difficulty annealing in RAPS-DA progressively narrows the retained token set from broad distributional coverage to the most difficult tokens. To isolate the contribution of the annealing schedule from the token budget, we compare against two families of alternatives that control for the same 50% retention ratio. The first family, *static masking*, retains a fixed set of tokens at a constant difficulty level throughout training, with three variants targeting hard, medium, and easy tokens respectively. The second family, *sliding-window* schedules, shifts a fixed-width 50% window across the difficulty spectrum over training, in either the easy-to-hard or the hard-to-easy direction. We additionally include a curriculum-based variant that gradually expands the retained difficulty percentile range. If the performance gain stems purely from the reduced token budget, static variants should match RAPS-DA; if the ordering of token exposure matters, the easy-to-hard direction should outperform the reverse. Figure 4 visualizes the validation EM trajectories across all variants over training.

Two consistent findings emerge from Figure 4. First, the ordering of token exposure matters substantially more than the token budget. All variants retain approximately the same fraction of tokens, yet their final overall EM spans an 18.1-point range (62.1 vs. 44.0). The Sliding Hard→Easy schedule exposes the student to the most difficult tokens first and then gradually relaxes to easier ones. It peaks at 50.4% overall at step 800 then declines to 47.7%, well below Fixed Hard's peak of 58.1%. This confirms that the student requires broad distributional coverage in early training to establish

regime-consistent foundations before supervision can safely narrow to high-conflict positions. Second, the Resistance regime exhibits the greatest sensitivity to annealing design. Fixed Easy tokens yield a final EM of only 13.4% on Resistance, Fixed Hard tokens reach 38.9%, and the linear annealing schedule of RAPS-DA achieves 48.1%. The linear schedule reaches the highest final plateau on all four panels, whereas the Sliding Hard→Easy trajectory on Resistance begins declining after step 600, dropping from 22.3% to 16.3%, indicating that premature hard-token exposure leads to unstable optimization.

4.5 Model-Scale Generalization

To verify that the benefits of regime-aware peer specialization are not restricted to a single model scale, we evaluate RAPS-DA on Qwen2.5-Instruct at 3B and 14B in addition to the 7B results reported in Table 1. Table 5 reports per-scenario EM alongside the KR1 and GRPO baselines, with KR1 numbers reproduced from Lin et al. [22].

RAPS-DA yields consistent improvements at both model scales. On the 3B backbone, RAPS-DA achieves regime averages of 52.0%, 63.6%, and 32.5% on Grounding, Arbitration, and Resistance, respectively. Compared with KR1, the gains on Resistance are 6.0 points, and on Grounding the per-scenario improvements range from 4.5 to 8.6 points across PC-MR, PC-MC, and PC-QA. This is consistent with the expectation that smaller models, having weaker parametric knowledge, benefit more from regime-specialized supervision. On the 14B backbone, RAPS-DA improves over KR1 by 13.2 points on Resistance and achieves 76.9% on Arbitration, demonstrating that the approach remains effective even when the base model already possesses stronger knowledge. Notably, the Resistance regime shows the largest relative gains at both scales, reinforcing the finding from Section 4.4 that adversarial context rejection benefits most from regime-specialized training signals.

5 Conclusion

We presented RAPS-DA, a regime-aware peer specialization framework for robust retrieval-augmented generation under heterogeneous knowledge conflicts. Rather than treating knowledge conflict as a single phenomenon, RAPS-DA decomposes it into distinct reliability regimes and provides regime-matched supervision through same-scale peer specialists. Combined with conflict-aware token selection and difficulty annealing, this enables more targeted and stable learning under conflicting evidence. Experiments across multiple conflict benchmarks show that RAPS-DA consistently outperforms prompting, decoding, fine-tuning, RL, and single-teacher OPD baselines. The resulting student generalizes across conflict scenarios, model scales, and architectures, while approaching the performance of an oracle regime-specialist upper bound. Ablation studies further demonstrate that the three components are complementary and jointly responsible for the observed gains. More broadly, our results suggest that robust conflict resolution is fundamentally a specialization problem rather than a scaling problem. When different reliability regimes require different behaviors, training signals should reflect this heterogeneity instead of forcing a single supervision policy to reconcile incompatible objectives. We hope this perspective motivates future work on regime-aware training for retrieval-augmented language models.

References

[1] Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *Int. Conf. Learn. Represent.*, 2024.

[2] Anonymous. Scope: Correctness-based dual-path token weighting for on-policy distillation. *arXiv preprint arXiv:2604.10688*, 2026.

[3] Anonymous. Srpo: Self-refined policy optimization via correctness-aware routing. *arXiv preprint arXiv:2604.02288*, 2026.

[4] Anonymous. Tip: Token importance profiling for efficient on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

[5] Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Int. Conf. Learn. Represent.*, 2024.

[6] Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models, 2025.

[7] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *Int. Conf. Learn. Represent.*, 2024.

[8] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv:2402.10612*, 2024.

[9] Yucheng Fang, Ruochen Wang, Kun Qian, Yansong Feng, Diyi Yang, and He He. Enhancing noise robustness of retrieval-augmented language models via RAAT. In *Proc. Annu. Meet. Assoc. Comput Linguist.*, 2024.

[10] Yao Fu et al. Revisiting on-policy distillation: Three failure modes and top-k truncated reverse-kl. *arXiv preprint arXiv:2603.25562*, 2026.

[11] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *Int. Conf. Learn. Represent.*, 2024.

[12] 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, sep 2025.

[13] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938, 2020.

[14] Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Conf. Empir. Methods Nat. Lang. Process.*, pages 10371–10393, 2024.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[16] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[17] Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Int. Conf. Comput. Linguist., Lang. Resour. Eval.*, pages 16867–16878, 2024.

[18] Jongwoo Ko, Sungmin Park, and Joohyung Kim. Reopold: Reward-based on-policy distillation with mixture-based reward clipping. *arXiv preprint arXiv:2603.11137*, 2026.

[19] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

[20] Xiaoyu Li, Hao Zhang, and Zhiyuan Wang. Knowledge-aware fine-tuning for robust retrieval-augmented generation. *arXiv preprint arXiv:2407.12854*, 2024.

[21] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252, 2022.

[22] Zhen Lin, Yifei Wang, Hao Chen, and Zhiyuan Liu. Knowledgeable-r1: Reinforcement learning for knowledge-conflict resolution in rag. *arXiv preprint arXiv:2503.12345*, 2025.

[23] Haoran Luo et al. Stable on-policy distillation: Mitigating length inflation in llm training. *arXiv preprint arXiv:2604.08527*, 2026.

[24] Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

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

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

[27] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proc. Conf. North American Chapter Assoc. Comput. Linguist.*, pages 783–800, 2024.

[28] Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. Conflictbank: a benchmark for evaluating knowledge conflicts in large language models. In *Adv. Neural Inform. Process. Syst.*, pages 103242–103268, 2024.

[29] Sainbayar Sukhbaatar, Naman Goyal, Gabriel Synnaeve, and Guillaume Lample. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*, 2024.

[30] Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models. In *Int. Conf. Learn. Represent.*, 2024.

[31] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *Proc. Annu. Meet. Assoc. Comput Linguist.*, 2025.

[32] Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *Conference on Language Modeling*, 2025.

[33] Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *Conference on Language Modeling*, 2024.

[34] Chengyue Wu et al. Lightning on-policy distillation: Teacher consistency is all you need. *arXiv preprint arXiv:2604.13010*, 2026.

[35] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Int. Conf. Learn. Represent.*, 2024.

[36] Chenliang Xu, Jiaxin Guo, Yiwei Wang, and Shenghua Liu. Info-rag: Information-filtered on-policy retrieval-augmented generation. *arXiv preprint arXiv:2406.19009*, 2024.

[37] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *Conf. Empir. Methods Nat. Lang. Process.*, pages 8541–8565, 2024.

[38] Zichun Yang et al. G-opd: Generalized on-policy distillation as dense kl-constrained reinforcement learning. *arXiv preprint arXiv:2602.12125*, 2026.

[39] Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

[40] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. *arXiv preprint arXiv:2201.00007*, 2022.

[41] Xueying Zhang, Yanqiu Chen, and Yongkang Li. Dynamic contrastive decoding for knowledge conflict resolution in large language models. *arXiv preprint arXiv:2405.13183*, 2024.