SEEK WHAT YOU IGNORANT: BOUNDARY-AWARE BEHAVIOURAL CALIBRATION FOR SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable agentic search requires large language models to calibrate trust in parametric memory and retrieved evidence, deciding when to use either source, integrate both, or abstain. Yet outcome rewards, whether binary or graded, provide only sparse feedback on completed trajectories and limited direct guidance for these intermediate decisions. Moreover, appropriate reasoning patterns vary across knowledge states. We propose **K**nowledge **b**oundary **S**elf-**D**istillation (**KbSD**), a quadrant-adaptive self-distillation framework that combines dense token-level supervision with GRPO-based reinforcement learning. KbSD estimates parametric certainty and response stability from repeated retrieval-free samples, and retrieval quality from the relevance of retrieved passages. These signals, together with quadrant labels and ground-truth answers, form privileged hints for a *teacher* that shares the *student*'s model parameters. Conditioned on these hints and the student's observed history, the teacher provides token-level guidance to the hint-free student. This information asymmetry enables dense supervision without a larger or separately trained teacher, and the student requires no privileged hints at inference. To accommodate differences in reasoning diversity across the four knowledge-boundary quadrants, KbSD uses mode-seeking reverse KL for concentrated evidence integration, mass-covering forward KL for diverse refusal patterns, and Pareto-optimal bidirectional KL for states that require both precise source selection and broad coverage. Experiments across nine benchmarks show lower unreliability than the evaluated baselines. We report task F1 alongside refusal rates to characterize the accuracy–refusal trade-off and interpret reliability gains in the context of answer coverage.

1 INTRODUCTION

Agentic search enables large language models (LLMs) to tackle knowledge-intensive reasoning tasks by gathering external evidence during inference (Yao et al., 2022; Jin et al., 2025; Nakano et al., 2021; Singh et al., 2025). These agents formulate queries, assess retrieved passages, and refine their reasoning through multi-turn interactions with external knowledge sources (Dong et al., 2025b;a; Li et al., 2025). Their reliability depends on ***knowledge boundary calibration***: determining when to trust parametric memory, when to rely on retrieved evidence, and how to act when the two sources disagree or neither provides sufficient support (Ren et al., 2025). Existing agentic search methods, such as Search-R1 (Jin et al., 2025) and TC-RAG (Jiang et al., 2024), emphasize *when to retrieve*, encouraging agents to search when their internal knowledge is insufficient. Yet initiating a search does not ensure that the resulting evidence is reliable: irrelevant, outdated, or noisy passages may leave uncertainty unresolved or even introduce new errors. Calibration requires the agent to make two independent assessments: whether its internal knowledge is reliable for the question at hand, and whether the retrieved evidence provides trustworthy support. And their combinations give rise to four qualitatively different situations: both sources may be reliable, only retrieval may help, only internal knowledge may suffice, or neither source may deserve trust.

Recent work has increasingly modeled retrieval as a decision that depends on both internal knowledge and external evidence. IKEA (Huang et al., 2025), CRaFT (Zhu et al., 2025), and GRAIT (Zhu* et al., 2025) use the distinction between known and unknown queries to determine when external retrieval is warranted. Building on this perspective, KbPO (Feng et al., 2026) introduces a four-quadrant taxonomy that jointly considers internal knowledge reliability and retrieval quality (Figure 1). ❶ In

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

Q1 (Integrated), reliable internal knowledge and consistent retrieved evidence support integrating both sources. ❷ In Q2 (External), reliable retrieval compensates for insufficient internal knowledge. ❸ In Q3 (Refusal), neither source provides reliable support, calling for explicit abstention (see Case 3 in Appendix C). ❹ In Q4 (Internal), reliable parametric memory supports an answer despite unhelpful retrieval. By incorporating these distinctions into post-training losses or rewards, such frameworks encourage agents to choose behaviors appropriate to their knowledge state.

These frameworks define *what* the agent should do in each knowledge state, and use outcome-level rewards to reinforce the corresponding behavior. However, identifying the desired behavior does not directly teach the reasoning needed to reach it. A reward delivered only after a complete trajectory tells the agent *whether* it succeeded, not *how* to reason toward that outcome. In particular, intermediate steps such as assessing retrieval relevance, gauging internal uncertainty, and weighing conflicting evidence receive no direct supervision. This gap gives rise to two challenges:

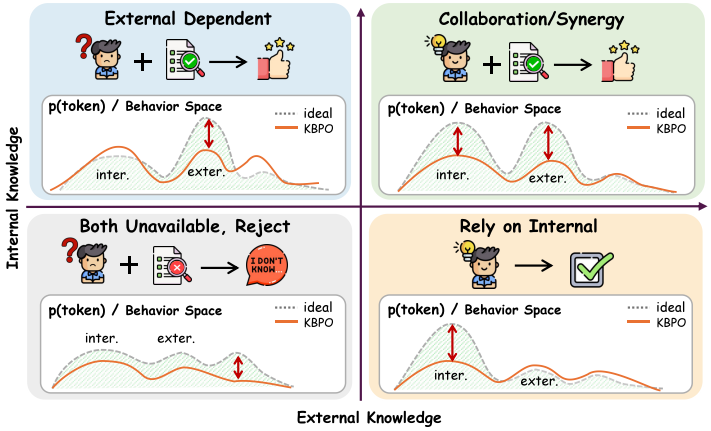

Figure 1: Four knowledge-boundary quadrants and their distillation objectives: reverse KL for Q1 (Integrated), bidirectional KL for Q2 (External) and Q4 (Internal), and forward KL for Q3 (Refusal).

❶ ***Challenge 1: From Sparse Rewards to Dense Supervision.*** Agents need guidance throughout the process of assessing retrieval relevance, gauging internal uncertainty, and deciding whether to trust memory, use retrieval, integrate both, or abstain. For example, in Q3, a terminal penalty can discourage an unsupported answer, but provides little direct guidance on how to recognize inadequate evidence and reason toward abstention. The challenge is to turn coarse feedback on completed trajectories into token-level supervision for these intermediate decisions.

❷ ***Challenge 2: Adapting to Reasoning Diversity Across Quadrants.*** Dense supervision must also account for differences in the reasoning patterns associated with each knowledge state. In Q1, agreement between reliable sources favors a relatively concentrated set of reasoning paths for integrating evidence. In Q3, appropriate abstention can follow diverse assessments of missing, noisy, or contradictory evidence, which a single refusal template may fail to capture. Q2 and Q4 require both precise source selection and robustness to varied forms of retrieval noise. A uniform distillation objective may therefore spread probability too broadly when focused reasoning is needed, or fail to cover valid alternatives when reasoning is more diverse.

To address these challenges, we propose **K**nowledge **b**oundary **S**elf-**D**istillation (**KbSD**), a quadrant-adaptive self-distillation framework for agentic search. We characterize each query using three probing signals: *parametric certainty*, *semantic stability of responses*, and *retrieval quality* (Section 3.1). Parametric certainty and retrieval quality determine the quadrant, while response stability is used to filter training data. These signals, the quadrant label, and the ground-truth answer form *knowledge boundary hints* for a privileged teacher that shares the student's model parameters. The teacher conditions on these hints and the student's observed history to provide dense token-level guidance; the student learns from the same history without access to the hints. This information asymmetry supplies supervision for intermediate reasoning decisions (Challenge 1), without requiring a larger or separately trained teacher. The privileged hints are used only during training.

To address Challenge 2, KbSD adapts the distillation objective to each quadrant. It uses mode-seeking reverse KL for Q1's concentrated integration behavior, mass-covering forward KL for Q3's diverse refusal patterns, and Pareto-optimal bidirectional KL to balance precision and coverage in Q2 and Q4. We combine this objective with GRPO policy optimization, so that dense guidance on the reasoning process complements sparse feedback on task outcomes. Our contributions are threefold:

- We introduce KbSD, a self-distillation framework that uses knowledge boundary hints available only to the teacher to provide dense supervision for intermediate reasoning decisions, without requiring a larger or separately trained model.
- We design a quadrant-adaptive distillation strategy that accounts for the reasoning diversity of different knowledge states: reverse KL for Q1, forward KL for Q3, and Pareto-optimal multi-objective optimization over both divergence directions for Q2 and Q4.
- Across nine benchmarks, KbSD achieves lower unreliability than the evaluated baselines. We report task F1 alongside refusal rates to make the accuracy–refusal trade-off more explicit.

2 Related Work

Agentic Search and Knowledge Boundary Calibration. Retrieval-Augmented Generation grounds LLM outputs in external knowledge (Lewis et al., 2020; Guu et al., 2020; Jiang et al., 2023; Asai et al., 2024; Yan et al., 2024; Trivedi et al., 2023), and agentic search further treats retrieval as a sequential decision process: ReAct (Yao et al., 2023) interleaves reasoning with tool use, TC-RAG (Jiang et al., 2024) maintains stack memory across turns, and RL-based agents such as Search-R1 (Jin et al., 2025), ARPO (Dong et al., 2025b), AEPO (Dong et al., 2025a), and MEM1 (Zhou et al., 2025) optimize retrieval strategies with trajectory-level outcome rewards (Shao et al., 2024). Closer to our setting, IKEA (Huang et al., 2025) models a known–unknown distinction to suppress redundant retrieval, and KbPO (Feng et al., 2026) extends it to a four-quadrant taxonomy over internal certainty and retrieval quality with quadrant-specific alignment rewards. **However, these methods rely on sparse reward or loss signals that specify the desired behavioral mode without supervising the intermediate reasoning process that leads to it—the gap our self-distillation framework addresses.**

Knowledge Distillation. Knowledge distillation (Hinton et al., 2015) transfers capabilities from a teacher to a student; on-policy distillation removes the exposure bias of off-policy training by supervising student rollouts—MiniLLM (Gu et al., 2024) adopts reverse KL for mode-seeking behavior, GKD (Agarwal et al., 2024) unifies on-policy token-level matching, and DistiLLM-2 (Ko et al., 2025) treats teacher- and student-generated outputs asymmetrically. Recent work removes the separate teacher entirely, converting sparse rewards into dense process supervision via self-distillation (He et al., 2026; Yang et al., 2026a;b; Li et al., 2026) either under a homogeneous reasoning distribution assumption (He et al., 2026; Yang et al., 2026a), or with a uniform divergence objective regardless of the underlying knowledge state (Yang et al., 2026b; Li et al., 2026). **However, none of these approaches account for the heterogeneous reasoning distributions induced by distinct knowledge-boundary states.**

3 Method

KbSD combines knowledge-boundary estimation, privileged self-distillation, and reward-based policy optimization (Figure 2). Boundary labels specify whether an agent should integrate internal and external knowledge, rely on either source, or abstain. KbSD turns these behavioral targets into dense supervision through a teacher that receives privileged boundary hints, adapts the distillation objective to each knowledge state, and jointly optimizes the student with GRPO.

3.1 Preliminaries: Boundary Quadrants

Our training-data construction and boundary quantification follow KbPO (Feng et al., 2026). Starting from seed QA pairs, we construct boundary-probing queries through *fact injection*, which adds a reasoning hop, and *information fuzzing*, which replaces precise constraints with underspecified ones, followed by multi-stage validation (Gao et al., 2025). For each query q, we estimate three signals: parametric certainty $\mu(q)$, measured by the answer match rate over N retrieval-free samples; semantic stability $\sigma(q)$, measured by their mean pairwise similarity; and retrieval quality $\hat{\rho}(q)$. The binary

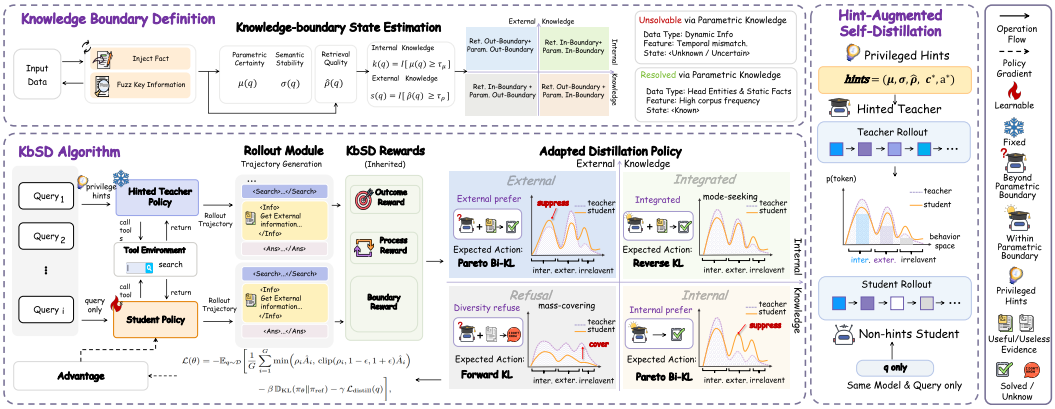

Figure 2: Overview of **KbSD**. We first synthesize boundary-challenging queries, then estimate each query's knowledge state using parametric certainty, semantic stability, and retrieval quality, and assign it to one of four behavioral quadrants. Based on these signals, a privileged hinted teacher generates calibrated reasoning demonstrations, which are distilled into a hint-free student with quadrant-adaptive objectives jointly optimized with GRPO. In this way, KbSD combines dense process-level supervision with sparse reward optimization.

indicators $k(q) := \mathbf{1}\{\mu(q) \geq \tau_\mu\}$ and $s(q) := \mathbf{1}\{\hat{\rho}(q) \geq \tau_\rho\}$ determine the target behavior:

$$c^*(q) = \begin{cases} \text{Integrated}, & (k(q), s(q)) = (1, 1), \\ \text{External}, & (k(q), s(q)) = (0, 1), \\ \text{Refusal}, & (k(q), s(q)) = (0, 0), \\ \text{Internal}, & (k(q), s(q)) = (1, 0). \end{cases} \tag{1}$$

These four states prescribe integrating both sources, relying on retrieval, abstaining, and trusting parametric memory, respectively. Response stability $\sigma(q)$ is used to filter training queries. Appendix B.1 specifies the estimators and probing settings; $\hat{\rho}(q)$ suppresses dependence on the observed retrieval context and current summary.

3.2 Hint-Augmented Self-Distillation

KbSD uses an information asymmetry between two branches of the same model to turn a coarse behavioral target into token-level supervision. Let $a^*(q)$ denote the annotated answer or answer aliases. The teacher receives a structured boundary hint, whereas the student receives the original query:

$$\begin{aligned} \boldsymbol{h}(q) &:= \big(\mu(q), \sigma(q), \hat{\rho}(q), c^*(q), a^*(q)\big), \\ \boldsymbol{p}^{\mathrm{T}}(q) &:= [\boldsymbol{h}(q); q], \qquad \boldsymbol{p}^{\mathrm{S}}(q) := q. \end{aligned} \tag{2}$$

The hint encodes internal certainty, response stability, retrieval reliability, the target behavior, and the annotated answer. It is available only to the teacher during training; the answer also enters the outcome reward through a separate path.

Let $\mathcal{H}_t := (\boldsymbol{o}_{<t}, O_{<t})$ collect the generated tokens and tool observations available before token t. The teacher evaluates the student's observed history under the privileged prompt:

$$\pi^{\mathrm{T}}(\cdot \mid q, \mathcal{H}_t) := \mathrm{sg}\big[\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{p}^{\mathrm{T}}(q), \mathcal{H}_t)\big], \tag{3}$$

where sg stops gradients through the teacher branch. Both branches share the parameter vector $\boldsymbol{\theta}$ and the tool observations from the actual student rollout; teacher evaluation does not invoke an independent search. Only the student branch contributes parameter gradients. For example, the hinted teacher can guide evidence assessment toward abstention in the Refusal quadrant or toward integration in the Integrated quadrant. The student learns these behaviors without access to the hints, including at inference time.

3.3 Quadrant-Adaptive Distillation

The four knowledge states favor different reasoning distributions: relatively concentrated integration in Q1, diverse abstention in Q3, and a balance between precise source selection and coverage in Q2

and Q4. We therefore adapt the direction of distillation to the target quadrant. Let $\boldsymbol{o}^{\mathrm{T}} \sim \pi^{\mathrm{T}}(\cdot \mid q)$ and $\boldsymbol{o} \sim \pi_{\boldsymbol{\theta}}(\cdot \mid q)$ denote teacher- and student-distributed token sequences, respectively. For compactness, the following objectives suppress tool observations in the conditioning context; these are retained as in Eq. (3).

Forward distillation: covering teacher behaviors. The forward objective maximizes the student's likelihood of teacher-distributed sequences:

$$\mathcal{L}_{\mathrm{fwd}}(q) \coloneqq -\mathbb{E}_{\boldsymbol{o}^{\mathrm{T}} \sim \pi^{\mathrm{T}}(\cdot|q)} \left[\sum_{t=1}^{|\boldsymbol{o}^{\mathrm{T}}|} \log \pi_{\boldsymbol{\theta}}(o_t^{\mathrm{T}} \mid \boldsymbol{o}_{<t}^{\mathrm{T}}, q) \right]. \tag{4}$$

With the teacher held fixed during the student update, this cross-entropy differs from forward KL only by the teacher's entropy. Its mass-covering behavior encourages the student to retain the range of responses supported by the teacher.

Reverse distillation: concentrating on supported behaviors. The reverse objective evaluates student-distributed sequences against the teacher:

$$\mathcal{L}_{\mathrm{rev}}(q) \coloneqq \mathbb{E}_{\boldsymbol{o} \sim \pi_{\boldsymbol{\theta}}(\cdot|q)} \left[\sum_{t=1}^{|\boldsymbol{o}|} \log \frac{\pi_{\boldsymbol{\theta}}(o_t \mid \boldsymbol{o}_{<t}, q)}{\pi^{\mathrm{T}}(o_t \mid \boldsymbol{o}_{<t}, q)} \right]. \tag{5}$$

This mode-seeking objective penalizes student behaviors that receive little support from the teacher, concentrating the policy on plausible reasoning paths.

Bidirectional distillation: balancing the two objectives. The External and Internal quadrants require both selective reliance on the appropriate source and robustness to varied retrieval noise. We balance forward and reverse distillation using their local gradient geometry. Let $\boldsymbol{g}_{\mathrm{f}} \coloneqq \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{fwd}}(q)$ and $\boldsymbol{g}_{\mathrm{r}} \coloneqq \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{rev}}(q)$. The mixing weight solves the two-objective minimum-norm problem:

$$\alpha^* \in \arg\min_{\alpha \in [0,1]} \left\| \alpha \boldsymbol{g}_{\mathrm{f}} + (1-\alpha) \boldsymbol{g}_{\mathrm{r}} \right\|_2^2. \tag{6}$$

The closed form follows by projecting the unconstrained minimizer onto $[0, 1]$ (Sener & Koltun, 2018). With $\boldsymbol{d} \coloneqq \boldsymbol{g}_{\mathrm{f}} - \boldsymbol{g}_{\mathrm{r}}$ and projection operator $\Pi_{[0,1]}$, it is

$$\alpha^* = \begin{cases} \Pi_{[0,1]}\left(-\dfrac{\langle \boldsymbol{g}_{\mathrm{r}}, \boldsymbol{d} \rangle}{\|\boldsymbol{d}\|_2^2} \right), & \|\boldsymbol{d}\|_2 > 0, \\ \frac{1}{2}, & \|\boldsymbol{d}\|_2 = 0. \end{cases} \tag{7}$$

When the gradients coincide, every weight is optimal; the symmetric choice above resolves the otherwise undefined quotient. Appendix A gives the derivation and the numerical tolerance used near this case. Holding the selected weight fixed for the student update, we set $\bar{\alpha} \coloneqq \mathrm{sg}[\alpha^*]$ and obtain

$$\mathcal{L}_{\mathrm{pareto}}(q) \coloneqq \bar{\alpha}\, \mathcal{L}_{\mathrm{fwd}}(q) + (1 - \bar{\alpha})\, \mathcal{L}_{\mathrm{rev}}(q). \tag{8}$$

The resulting gradient is the minimum-norm convex combination of $\boldsymbol{g}_{\mathrm{f}}$ and $\boldsymbol{g}_{\mathrm{r}}$. Its negative is a descent direction for both objectives whenever this combined gradient is nonzero.

Assignment by knowledge state. We select the distillation objective according to the target behavior:

$$\mathcal{L}_{\mathrm{distill}}(q) = \begin{cases} \mathcal{L}_{\mathrm{rev}}(q), & c^*(q) = \text{Integrated}, \\ \mathcal{L}_{\mathrm{fwd}}(q), & c^*(q) = \text{Refusal}, \\ \mathcal{L}_{\mathrm{pareto}}(q), & c^*(q) \in \{\text{External}, \text{Internal}\}. \end{cases} \tag{9}$$

Reverse KL favors focused integration, forward KL accommodates diverse refusal patterns, and the bidirectional objective balances precision and coverage when only one knowledge source is reliable.

3.4 Joint RL and Distillation Training

Distillation supervises intermediate reasoning decisions, while reinforcement learning evaluates their task outcomes. We combine the two using the boundary-aware reward design of KbPO and a clipped GRPO objective.

Composite rewards. For rollout $\boldsymbol{o}_i$ of length T_i, let r_i^{out} denote the final-answer reward, $r_{i,t}^{\text{proc}}$ the reward for valid tool use, and $r_{i,t}^{\text{align}}$ the reward for behavior aligned with $c^*(q)$. The stepwise reward and trajectory return are

$$\begin{aligned} r_{i,t} &\coloneqq \mathbf{1}\{t = T_i\}\big(r_i^{\text{out}} + r_i^{\text{pen}}\big) + \lambda r_{i,t}^{\text{proc}} + r_{i,t}^{\text{align}}, \\ R_i &\coloneqq \sum_{t=1}^{T_i} r_{i,t}. \end{aligned} \tag{10}$$

The outcome reward may be graded, such as F1. The terminal indicator ensures that the outcome reward and refusal penalty are each applied once. Define $\mathcal{C}_{\text{ans}} \coloneqq \{\text{Integrated}, \text{External}, \text{Internal}\}$, and let $b(\boldsymbol{o}_i) \in \{0, 1\}$ indicate whether the final response abstains. The penalty is

$$r_i^{\text{pen}} \coloneqq -\phi\, b(\boldsymbol{o}_i)\, \mathbf{1}\{c^*(q_i) \in \mathcal{C}_{\text{ans}}\}\,. \tag{11}$$

Thus, abstention is penalized only in knowledge states for which an answer is expected.

Joint objective. For each q, we collect G rollouts from the behavior policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$ and form the return vector $\boldsymbol{R} \coloneqq (R_1, \ldots, R_G)^\top$. Define the group-relative advantage, likelihood ratio, and clipped ratio as

$$\widehat{A}_i \coloneqq \frac{R_i - \text{mean}(\boldsymbol{R})}{\text{std}(\boldsymbol{R})}, \qquad w_i(\boldsymbol{\theta}) \coloneqq \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{o}_i \mid q)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{o}_i \mid q)},$$
$$\bar{w}_i(\boldsymbol{\theta}) \coloneqq \text{clip}\,\big(w_i(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon\big).$$

The joint minimization objective is

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q\sim\mathcal{D}}\left[-\frac{1}{G}\sum_{i=1}^{G} \min\Big\{ w_i(\boldsymbol{\theta})\widehat{A}_i, \bar{w}_i(\boldsymbol{\theta})\widehat{A}_i \Big\} + \beta\, D_{\text{KL}}(\pi_{\boldsymbol{\theta}} \,\|\, \pi_{\text{ref}}) + \gamma\, \mathcal{L}_{\text{distill}}(q) \right]. \tag{12}$$

Here, π_{ref} is the reference policy, ϵ controls policy clipping, and β and γ weight reference regularization and distillation, respectively. The first term promotes trajectories with higher group-relative returns; the remaining terms constrain policy drift and provide dense guidance for calibrating trust between internal and external knowledge.

4 Experiment

4.1 Experimental Setup

❶ Datasets and Metrics. Training queries follow KbPO's boundary-oriented synthesis pipeline (Feng et al., 2026), seeded by HotpotQA and 2WikiMultiHopQA. We evaluate on nine benchmarks in three categories:

- **Multi-hop QA:** HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023).
- **Open-domain QA:** NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023).
- **Agentic reasoning:** FRAMES (Krishna et al., 2025) and GAIA (Mialon et al., 2023).

HotpotQA and 2WikiMultiHopQA are in-domain; the remaining seven benchmarks are held out. We report **F1**, **refusal rate**, and **unreliability rate**, all as percentages. F1 is averaged over all queries, with refused responses assigned zero; explicit abstentions (e.g., "I don't know") count as refusals. We report the Unreliability Rate, defined on the $[0, 1]$ scale as $1 - (\text{F1} + \text{Refusal Rate})$, where F1 is a graded measure of answer overlap rather than a binary correctness judgment; accordingly, this quantity reflects the portion of graded answer quality that is neither captured by F1 nor accounted for by explicit refusal, rather than a strict count of incorrect answers. Lower is better. Table 1 reports unreliability, while Table 2 and Figure 4 report F1 alongside refusal rates.

❷ Baselines. Table 1 compares four baaselines families:

- **No RAG:** direct answering (Base) and chain-of-thought prompting (CoT).
- **Naive RAG:** few-shot RAG (FS-RAG) (Lewis et al., 2020) and forward-looking active retrieval (FL-RAG) (Jiang et al., 2023).
- **Agentic RAG:** ReAct (Yao et al., 2023), IRCoT (Trivedi et al., 2023), and TC-RAG (Jiang et al., 2024).
- **RL-based agentic RAG:** ReSearch (Chen et al., 2025), Search-R1 (Jin et al., 2025), AEPO (Dong et al., 2025a), ARPO (Dong et al., 2025b), MEM1 (Zhou et al., 2025), and KbPO (Feng et al., 2026).

KbPO shares our quadrant taxonomy and rewards but lacks dense self-distillation. Table 2 additionally includes IKEA (Huang et al., 2025) and GRAIT (Zhu* et al., 2025).

All baselines are reimplemented and retuned under a unified evaluation protocol. We use E5-Base-V2 (Wang et al., 2022) over the 2018 Wikipedia dump (Karpukhin et al., 2020), a top-3 passage budget, at most five retrieval iterations, shared prompt templates adapted to each method's original interface, and the same answer/refusal parsing script.

❸ Implementation and Training. We use Qwen2.5-3B and Qwen2.5-7B (Yang et al., 2024) with the shared retrieval protocol above. KbSD is implemented in VERL and trained with GRPO for 300 steps on eight NVIDIA A100 80GB GPUs using FSDP, with AdamW at a learning rate of 10^{-6}. Each batch contains 128 queries with five rollouts each. We set $\beta = 0.001$, $\lambda = 0.2$, $\gamma = 0.1$, $\tau_\mu = 0.6$, and $\tau_\rho = 0.5$ and use greedy decoding for evaluation. Appendix B provides the full training, generation, and retrieval configuration.

4.2 Main Results

Table 1 reports the Unreliability Rate of all methods at both scales. We highlight four findings.

Overall Performance. KbSD attains the lowest unreliability in every column at both scales. Relative to KbPO—which shares our quadrant taxonomy and reward design and differs only in the absence of dense self-distillation—the average drops from 64.32 to 47.63 on Qwen2.5-3B and from 57.99 to 41.15 on Qwen2.5-7B, absolute reductions of 16.69 and 16.84 points, respectively. Because the two methods are otherwise matched, this margin isolates the contribution of dense process-level supervision.

Calibration and Out-of-Domain Robustness. The gains hold across all seven held-out sets despite training only on HotpotQA and 2WikiMultiHopQA. On the long-tail PopQA, unreliability drops from 73.10 (KbPO) to 45.55 (3B), and on TriviaQA from 39.90 to 34.85, the lowest in the table; on FRAMES and GAIA, where Wikipedia retrieval often cannot supply the required evidence, it falls from 62.60 to 44.00 and from 77.20 to 58.07. A substantial part of these gains comes from calibrated abstention: when neither parametric knowledge nor retrieval is reliable, the agent abstains rather than producing confident errors—the behavior the Refusal quadrant is designed to teach.

Comparison with RL-based Methods. KbSD also improves substantially over boundary-agnostic RL agents. Several behave erratically in our unified setting—e.g., AEPO's unreliability rises from 77.02 (3B) to 88.88 (7B)—reflecting that outcome-only rewards give no guidance on allocating trust between memory and retrieval, a failure mode that boundary-aware supervision avoids.

4.3 Ablation Study

We ablate three components on Qwen2.5-3B across five benchmarks. Removing dense distillation raises average unreliability from 46.46 to 61.92 (KbPO); replacing quadrant-adaptive selection with fixed forward or reverse KL yields 49.13 or 64.32 (Table 5, Appendix F). We next isolate the teacher's answer hint.

Table 1: **Unreliability Rate (%, ↓)**, defined as $1 - (\text{F1} + \text{Refusal Rate})$: a graded residual, not a binary error count. HotpotQA and 2Wiki are in-domain; the other benchmarks are held out. Bold marks the best values at each model scale; blue highlights the best average.

	In-Domain		Out-of-Domain							
Method	**2Wiki**	**HotpotQA**	**Bamboogle**	**FRAMES**	**GAIA**	**MuSiQue**	**NQ**	**PopQA**	**TriviaQA**	**Avg**
Qwen2.5-3B										
No RAG										
Base	76.02	75.92	90.55	91.99	94.30	90.30	85.73	87.89	61.16	83.76
CoT	81.10	76.18	79.20	92.84	92.98	89.53	83.47	88.63	60.38	82.70
Naive RAG										
FS-RAG	84.53	74.15	89.52	89.58	95.70	92.36	80.16	69.59	54.62	81.13
FL-RAG	83.20	73.22	88.95	90.81	98.42	92.71	78.07	66.53	51.50	80.38
Agentic RAG										
ReAct	74.91	65.63	75.14	89.47	95.22	86.08	72.81	63.53	53.96	75.19
IRCoT	84.11	75.50	74.73	93.21	97.23	87.57	72.14	64.98	50.81	77.81
TC-RAG	71.53	78.06	82.41	92.31	94.47	91.01	79.26	82.31	49.54	80.10
RL-based Agentic RAG										
ReSearch	72.77	66.04	84.91	90.00	95.52	90.53	65.39	64.03	46.07	75.03
Search-R1	70.10	62.76	70.10	89.24	95.32	86.47	65.27	65.76	44.92	72.22
AEPO	76.99	71.29	77.91	87.74	94.99	88.30	73.24	70.52	52.22	77.02
ARPO	70.45	63.52	72.68	86.51	92.29	86.62	66.71	66.56	46.34	72.41
MEM1	81.94	79.85	94.81	95.01	97.53	95.53	80.82	83.29	66.91	86.19
KbPO	70.50	57.60	68.10	62.60	77.20	61.40	68.50	73.10	39.90	64.32
KbSD (Ours)	**45.89**	**41.57**	**52.20**	**44.00**	**58.07**	**42.09**	**64.43**	**45.55**	**34.85**	**47.63**
Qwen2.5-7B										
No RAG										
Base	74.59	73.37	82.14	87.48	95.71	87.85	80.28	85.21	50.92	79.73
CoT	76.45	70.90	62.44	82.40	93.83	85.65	77.53	84.93	50.67	76.09
Naive RAG										
FS-RAG	82.29	70.79	83.14	87.48	94.98	89.26	83.18	81.06	64.98	81.91
FL-RAG	80.22	65.58	75.90	87.90	95.40	87.54	80.28	74.72	57.34	78.32
Agentic RAG										
ReAct	72.49	57.19	72.37	84.71	93.63	80.66	69.99	63.71	45.45	71.13
IRCoT	63.55	73.71	78.10	93.22	94.50	91.61	80.37	65.86	50.57	76.83
TC-RAG	70.30	59.17	74.87	83.54	91.79	82.44	70.99	62.83	45.22	71.24
RL-based Agentic RAG										
ReSearch	69.97	69.61	69.58	84.39	95.57	87.42	76.31	82.09	51.75	76.30
Search-R1	64.97	61.11	57.96	81.99	89.92	80.92	70.41	72.53	44.09	69.32
AEPO	80.12	86.15	86.76	92.76	97.35	94.15	90.07	90.05	82.47	88.88
ARPO	69.29	74.80	67.06	87.82	91.44	87.29	82.20	86.38	59.84	78.46
MEM1	74.71	70.02	63.50	85.85	93.50	85.87	73.62	74.20	48.96	74.47
KbPO	65.20	51.80	61.50	56.30	71.90	54.70	62.30	67.80	32.60	57.99
KbSD (Ours)	**44.27**	**38.71**	**33.80**	**41.11**	**41.82**	**38.73**	**59.01**	**44.06**	**28.86**	**41.15**

Contribution of the Answer Hint. We test whether gains depend on the teacher's access to ground-truth answers by removing $a^*(q)$ while retaining $(\mu(q), \sigma(q), \hat{\rho}(q), c^*(q))$. Average unreliability rises by 2.78 points (46.46 to 49.24), yet remains 12.68 points below KbPO (Figure 3). Thus, the answer hint helps, but substantial gains persist with boundary signals alone in the teacher hint. The no-answer variant retains approximately 82% of the full model's unreliability reduction over KbPO (12.68/15.46). This supports gains beyond answer disclosure, but does not distinguish improved answer quality from increased refusal.

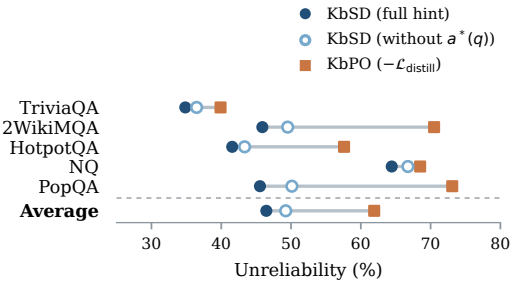

Figure 3: Answer-hint ablation on Qwen2.5-3B (unreliability, %). Average covers five datasets.

4.4 Analysis

We further analyze how KbSD balances answer quality and abstention. We compare F1 and refusal rates across methods and individual benchmarks to examine whether reduced unreliability is accompanied by preserved or improved answer quality.

Accuracy–Refusal Trade-off. Table 2 reports higher F1 for KbSD than KbPO (22.10 vs. 20.50), alongside higher refusal (30.30% vs. 15.18%). KbSD also has higher F1 and lower refusal than GRAIT.

Per-Dataset F1 and Refusal. Figure 4 compares KbPO and KbSD on Qwen2.5-3B. KbSD has higher refusal rates on all nine benchmarks, while F1 increases on seven: for example, 2Wiki improves from 17.13 to 23.85, HotpotQA from 28.54 to 32.60, and PopQA from 5.85 to 9.24. The exceptions are Bamboogle (16.66 to 15.63) and GAIA (3.32 to 0.91). On GAIA, refusal rises from 19.48% to 41.02% while F1 falls by 2.41 points, indicating an accuracy cost despite low F1 for both methods. Overall, higher refusal rates coexist with higher F1 on most benchmarks, which is consistent with gains beyond simply sacrificing task performance for abstention. KbSD also leads on eight of nine benchmarks in both natural-coverage Conditional F1 and a consistency-ranked, matched-coverage comparison with KbPO (Appendix G; Figures 6–7). GAIA remains the exception; matching coverage controls the retained fraction, but not which queries are retained.

Table 2: Qwen2.5-3B baseline comparison on all benchmarks. The highest F1 is shown in bold.

Methods	F1-Score	Refusal-Ratio
KbSD	**22.10**	**30.30**
KbPO	20.50	15.18
IKEA	21.47	N/A
GRAIT	18.90	32.70
Raw	16.24	N/A

Behavioral Constraint and Model Scale. Q1 exhibits greater student–teacher KL divergence than Q4 in both directions ($p < 0.01$), consistent with different degrees of behavioral constraint. KbSD also retains comparable unreliability reductions over KbPO at 3B and 7B (16.69 and 16.84 points). Appendix E provides the divergence analysis (Figure 5) and model-scale comparison; Appendix F reports the full fixed-direction ablations.

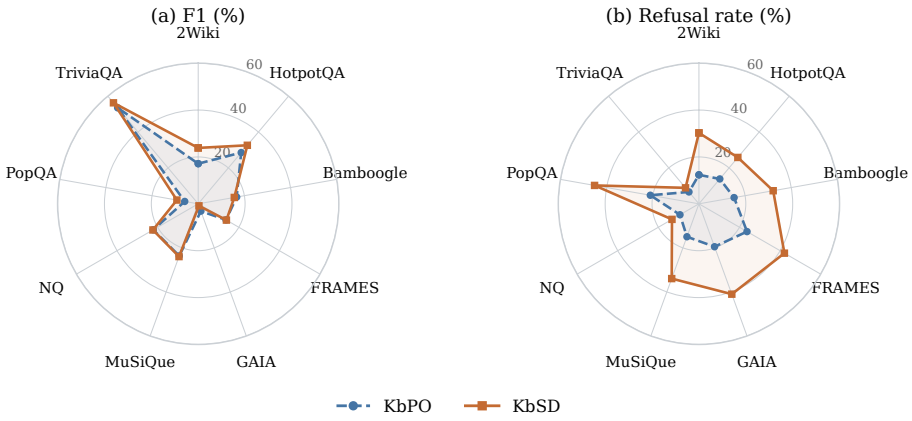

Figure 4: Per-dataset comparison of KbPO and KbSD on Qwen2.5-3B: (a) F1 and (b) refusal rate, both in percent on the same 0–60 radial scale. KbSD has higher refusal rates on all nine benchmarks and higher F1 on seven, with declines on Bamboogle and GAIA. Results use each method's operating point, without matching refusal rates or coverage.

5 Conclusion and Future Work

We presented **KbSD**, a knowledge-boundary-aware self-distillation framework that teaches search agents when to trust parametric memory, rely on retrieved evidence, integrate both sources, or abstain. An information-asymmetric same-model teacher provides dense token-level supervision during training. Quadrant-adaptive distillation complements GRPO's sparse outcome feedback to guide reasoning and source selection, without requiring privileged hints at inference time. Across nine benchmarks, KbSD lowers average unreliability from 64.32 to 47.63 on 3B models and from 57.99 to 41.15 on 7B models relative to KbPO, with improvements extending to all seven held-out benchmarks. These results motivate broader evaluation across model families and retrieval systems, including risk–coverage curves and explicit measures of over-refusal, to assess when lower unreliability translates into more useful answers.

AI USE STATEMENT

Large language models (LLMs) were used to assist with manuscript writing, language polishing, figure presentation, and LaTeX formatting. Specifically, LLMs helped refine clarity, grammar, and readability, suggest alternative phrasings, and reduce repetitive expressions. They also assisted with visualization layouts and plotting and document-preparation code. The experimental values used in these figures were supplied and verified by the authors; generative AI was not used to create these values. The authors retain responsibility for reviewing AI-assisted text, code, and figures, for the interpretation of the results, and for all final scientific and editorial decisions.

ETHICS STATEMENT

This work studies automated search agents using publicly available language models, question-answering benchmarks, and a Wikipedia retrieval corpus. The experiments evaluate model behavior rather than human participants. The study aims to reduce unsupported answers by improving how agents use internal knowledge, retrieved evidence, and explicit abstention. Nevertheless, retrieved passages may contain inaccurate, outdated, or biased information, and lower unreliability does not by itself establish factual correctness or suitability for high-stakes use. We therefore report answer quality alongside refusal behavior and discuss the limitations of aggregate calibration metrics. Use of the models, datasets, and retrieval corpus should respect their respective licenses and terms of use.

REPRODUCIBILITY STATEMENT

Experiments use Qwen2.5-3B and Qwen2.5-7B with E5-Base-V2 retrieval over the 2018 Wikipedia dump. We evaluate on HotpotQA, 2WikiMultiHopQA, Bamboogle, FRAMES, GAIA, MuSiQue, NQ, PopQA, and TriviaQA. Section 4.1 describes the data sources, metrics, baseline reproduction protocol, and training setup. All baselines are reimplemented and retuned under the shared evaluation protocol, with top-3 retrieval, at most five retrieval iterations, method-adapted prompt templates, and a common answer/refusal parsing script. Section 3 specifies the boundary signals, teacher construction, distillation objectives, and joint optimization procedure; Appendix A provides the bidirectional-distillation derivation. Appendix B records the model configuration, software environment, hardware, generation settings, and training hyperparameters.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. Research: Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5–6):313–318, 2012.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, et al. Agentic entropy-balanced policy optimization. *arXiv preprint arXiv:2510.14545*, 2025a.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025b.

Tao Feng, Xinke Jiang, Xinyan Hu, Yonggang Zhang, Zhen Tao, Wentao Zhang, Boyang Liu, Wenhao Jiang, and Chao Wu. Trust within? seek beyond? knowledge boundary aware policy optimization for agentic search. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27664–27682, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. URL https://aclanthology.org/2026.acl-long.1276/.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. *arXiv preprint arXiv:2508.07976*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

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

Xanh Ho, Anh-Khoa Duong, Quoc-Huy Nguyen, and Suong Nguyen. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *COLING*, 2020.

Ziyang Huang, Xiaowei Yuan, Yiming Ju, Jun Zhao, and Kang Liu. Reinforced internal-external knowledge synergistic reasoning for efficient adaptive search agent. *arXiv preprint arXiv:2505.07596*, 2025.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, et al. Tc-rag: Turing-complete rag's case study on medical llm systems. *arXiv preprint arXiv:2408.09199*, 2024.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwyer, and Mohit Iyyer. Active retrieval augmented generation. *EMNLP*, 2023. (FLARE/FL-RAG).

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.

Mandar Joshi et al. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781, 2020.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025.

Tom Kwiatkowski et al. Natural questions: A benchmark for question answering research. *TACL*, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Alex Mallen, Akari Asai, et al. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL*, 2023.

Grégoire Mialon et al. Gaia: A benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2023.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3697–3715, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.250/.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

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

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V. Vasilakos. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. ♫ MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin 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, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL `https://api.semanticscholar.org/CorpusID:274859421`.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Zhilin Yang, Peng Qi, Saizheng Zhang, et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Shunyu Yao et al. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025.

Runchuan Zhu*, Xinke Jiang*, Jiang Wu*, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. Grait: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. *NAACL 2025*, 2025.

Runchuan Zhu, Zhipeng Ma, Jiang Wu, Junyuan Gao, Jiaqi Wang, Dahua Lin, and Conghui He. Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26157–26165, 2025.

A Derivation of the Closed-Form Solution for Pareto Bidirectional Distillation

This appendix provides the complete derivation of the closed-form solution α^* stated in the main text. The result is a two-objective specialization of the Multiple-Gradient Descent Algorithm (MGDA), originally proposed by Désidéri (Désidéri, 2012) and later applied to multi-task deep learning by Sener and Koltun (Sener & Koltun, 2018).

Problem statement. Let $\mathbf{g}_f = \nabla_\theta \mathcal{L}_{\text{fwd}}(q)$ and $\mathbf{g}_r = \nabla_\theta \mathcal{L}_{\text{rev}}(q)$ be the gradients of the forward and reverse distillation losses with respect to the model parameters θ, respectively. We seek the convex combination weight $\alpha \in [0, 1]$ that minimizes the squared norm of the combined gradient:

$$\begin{aligned} \alpha^* &= \arg\min_{\alpha \in [0,1]} f(\alpha), \\ \text{where} \quad f(\alpha) &= \left\| \alpha\, \mathbf{g}_f + (1-\alpha)\, \mathbf{g}_r \right\|^2 . \end{aligned} \tag{13}$$

Geometrically, as α varies over $[0, 1]$, the point $\alpha\, \mathbf{g}_f + (1-\alpha)\, \mathbf{g}_r$ traces the line segment from $\mathbf{g}_r$ (at $\alpha = 0$) to $\mathbf{g}_f$ (at $\alpha = 1$), and we seek the point on this segment closest to the origin. The direction $-[\alpha^* \mathbf{g}_f + (1-\alpha^*)\mathbf{g}_r]$ is then guaranteed to be a descent direction for both objectives simultaneously (Désidéri, 2012).

Step 1: Expand the objective. Define $\mathbf{d} = \mathbf{g}_f - \mathbf{g}_r$ for notational convenience. Then

$$\alpha\, \mathbf{g}_f + (1-\alpha)\, \mathbf{g}_r = \mathbf{g}_r + \alpha\, \mathbf{d}. \tag{14}$$

Substituting into $f(\alpha)$ and expanding:

$$\begin{aligned} f(\alpha) &= \left\| \mathbf{g}_r + \alpha\, \mathbf{d} \right\|^2 \\ &= \left(\mathbf{g}_r + \alpha\, \mathbf{d} \right)^\top \left(\mathbf{g}_r + \alpha\, \mathbf{d} \right) \\ &= \mathbf{g}_r^\top \mathbf{g}_r + 2\alpha\, \mathbf{d}^\top \mathbf{g}_r + \alpha^2\, \mathbf{d}^\top \mathbf{d} \\ &= \|\mathbf{g}_r\|^2 + 2\alpha\, \mathbf{d}^\top \mathbf{g}_r + \alpha^2 \|\mathbf{d}\|^2 . \end{aligned} \tag{15}$$

This is a scalar quadratic function in α with a positive leading coefficient $\|\mathbf{d}\|^2 \geq 0$, so it is convex.

Step 2: Compute the unconstrained minimizer. Setting the derivative to zero:

$$\frac{\mathrm{d}f}{\mathrm{d}\alpha} = 2\, \mathbf{d}^\top \mathbf{g}_r + 2\alpha\, \|\mathbf{d}\|^2 = 0. \tag{16}$$

Solving for α gives the unconstrained minimizer:

$$\hat{\alpha} = -\frac{\mathbf{d}^\top \mathbf{g}_r}{\|\mathbf{d}\|^2} = \frac{-(\mathbf{g}_f - \mathbf{g}_r)^\top \mathbf{g}_r}{\|\mathbf{g}_f - \mathbf{g}_r\|^2} = \frac{(\mathbf{g}_r - \mathbf{g}_f)^\top \mathbf{g}_r}{\|\mathbf{g}_f - \mathbf{g}_r\|^2}. \tag{17}$$

Step 3: Enforce the box constraint. Since we require $\alpha \in [0, 1]$, and $f(\alpha)$ is convex, the constrained minimizer is obtained by projecting $\hat{\alpha}$ onto the interval $[0, 1]$:

$$\alpha^* = \text{clip}(\hat{\alpha},\ 0,\ 1) = \max\Big(0,\ \min\big(1,\ \hat{\alpha}\big)\Big). \tag{18}$$

Substituting Eq. (17) into Eq. (18) yields the final expression:

$$\alpha^* = \text{clip}\left(\frac{(\mathbf{g}_r - \mathbf{g}_f)^\top \mathbf{g}_r}{\|\mathbf{g}_f - \mathbf{g}_r\|^2},\ 0,\ 1 \right). \tag{19}$$

Geometric interpretation of the boundary cases. The three cases of the clipping operation correspond to distinct geometric configurations of the two gradient vectors (see also Figure 1 of Désidéri (Désidéri, 2012)):

- $\hat{\alpha} \leq 0$ **(clipped to** $\alpha^* = 0$**):** This occurs when $(\mathbf{g}_r - \mathbf{g}_f)^\top \mathbf{g}_r \leq 0$, i.e., $\mathbf{g}_f^\top \mathbf{g}_r \geq \|\mathbf{g}_r\|^2$. The reverse gradient $\mathbf{g}_r$ alone is already the minimum-norm point on the segment; the combined direction reduces to $\mathbf{g}_r$, meaning reverse distillation dominates.

- $\hat{\alpha} \geq 1$ **(clipped to $\alpha^* = 1$):** This occurs when $(\mathbf{g}_r - \mathbf{g}_f)^\top \mathbf{g}_r \geq \|\mathbf{g}_f - \mathbf{g}_r\|^2$, which simplifies to $\mathbf{g}_f^\top \mathbf{g}_r \geq \|\mathbf{g}_f\|^2$. The forward gradient $\mathbf{g}_f$ alone is the minimum-norm point; forward distillation dominates.
- $0 < \hat{\alpha} < 1$ **(no clipping):** The minimum-norm point lies in the interior of the segment. In this case, the optimal combined direction $\alpha^* \mathbf{g}_f + (1-\alpha^*)\mathbf{g}_r$ is perpendicular to $\mathbf{d} = \mathbf{g}_f - \mathbf{g}_r$, i.e., it is the foot of the perpendicular from the origin to the line through $\mathbf{g}_f$ and $\mathbf{g}_r$.

Special case: $\mathbf{g}_f = \mathbf{g}_r$. When the two gradients are identical, $\mathbf{d} = \mathbf{0}$ and the denominator $\|\mathbf{d}\|^2 = 0$. In this degenerate case, every $\alpha \in [0, 1]$ yields the same combined gradient, so any choice is optimal. In practice, we set $\alpha^* = 0.5$ whenever $\|\mathbf{d}\|^2$ falls below a small numerical threshold ϵ (we use $\epsilon = 10^{-8}$).

Connection to MGDA. The optimization problem in Eq. (13) is the two-objective ($T = 2$) instance of the general MGDA subproblem: finding the minimum-norm point in the convex hull of T gradient vectors. For $T = 2$, the convex hull is a line segment and the solution admits the analytical form derived above, as noted in Eq. (4) of Sener and Koltun (Sener & Koltun, 2018) and Algorithm 1 therein. For $T > 2$, the problem becomes a quadratic program over the simplex and is typically solved with the Frank–Wolfe algorithm (Sener & Koltun, 2018). Since our setting involves exactly two objectives (forward and reverse KL divergence), the closed-form solution applies directly, incurring negligible computational overhead.

B Reproducibility Details

We implement KbSD using the PyTorch-based VERL framework on a high-performance computing cluster. The following configuration is used for all reported experiments unless otherwise specified.

B.1 Knowledge-Boundary Estimators

We make the estimators in Section 3.1 explicit below. The probing implementation draws $N = 10$ retrieval-free responses per query with frozen model parameters, temperature 0.7, top-$p = 0.9$, and at most 64 newly generated tokens per response. This sampling count is distinct from the five rollouts per query used during GRPO training. The certainty and stability estimators follow the probing framework of KbPO (Feng et al., 2026), with the concrete preprocessing and encoder specified here.

Answer extraction and normalization. Let $E(y)$ extract the first `<answer>...</answer>` span, ignoring tag case; if the span is absent, it returns the full response. Let $\nu(x)$ lowercase and trim the text, replace markup tags with spaces, replace each run of characters outside ASCII letters, digits, and the CJK range `U+4E00–U+9FFF` with a space, then collapse whitespace and trim again. Articles are not removed. For the set of annotated answer aliases $A^*(q)$, the implementation uses normalized substring matching:

$$M(y, A^*) = \mathbf{I}[\exists a \in A^* : \ \nu(a) \neq \emptyset, \ \nu(E(y)) \neq \emptyset, \ \nu(a) \subseteq_{\text{str}} \nu(E(y))], \tag{20}$$

where $\subseteq_{\text{str}}$ denotes contiguous substring containment, including exact equality. Thus, this probing match is not strict normalized exact match and is separate from the graded F1 metric used for evaluation. Parametric certainty is

$$\mu(q) = \frac{1}{N} \sum_{j=1}^{N} M(y_j, A^*(q)), \qquad N = 10. \tag{21}$$

Semantic stability. We encode the extracted answer strings $E(y_j)$ with `sentence-transformers/all-MiniLM-L6-v2`. Writing $e_j = \text{Enc}(E(y_j))$, stability is the mean cosine similarity over all distinct response pairs:

$$\sigma(q) = \frac{2}{N(N-1)} \sum_{1 \leq j < k \leq N} \frac{e_j^\top e_k}{\|e_j\|_2 \|e_k\|_2}. \tag{22}$$

The encoder receives extracted text before the string normalization used for answer matching. Cosine similarities are averaged directly, without remapping them to $[0, 1]$.

Retrieval quality. Let $C_{<t}$ contain the retrieved passages observed along the student's actual trajectory and let u_t be the current summary. The local implementation uses a single LLM judge (default `gpt-4o`, temperature 0), rather than an ensemble of judges. Its prompt jointly requests a retrieval-quality score and a behavioral category. The retrieval rubric assigns anchors of 0 to irrelevant context, 0.5 to partially relevant context missing key information, and 1 to highly relevant, useful context; intermediate scores are permitted. If $J(q, C_{<t}, u_t)$ is the number parsed from the judge's `<retrieval_score>` field, then

$$\hat{\rho}(q; C_{<t}, u_t) = \begin{cases} \min\{1, \max\{0, J(q, C_{<t}, u_t)\}\}, & \text{if a valid score is returned,} \\ 0, & \text{if scoring or parsing fails.} \end{cases} \tag{23}$$

The contexts are concatenated in observed order and truncated to 8,000 characters, with a truncation marker appended; the summary is truncated to 2,000 characters. The score assesses the combined context, rather than averaging passage-level similarities or retriever scores. The notation $\hat{\rho}(q)$ in the main text suppresses this dependence on the observed context and summary. The thresholds used for quadrant assignment are $\tau_\mu = 0.6$ and $\tau_\rho = 0.5$.

B.2 Model Configuration

We evaluate the Qwen2.5 model family at the 3B and 7B scales. Both models are initialized from the official HuggingFace checkpoints and loaded in `bfloat16` precision.

B.3 Inference and Retrieval Settings

We use greedy decoding for deterministic evaluation. Retrieval uses the E5-Base-V2 dense retriever indexed on the 2018 Wikipedia dump. Table 3 lists the complete generation and retrieval configuration.

Table 3: Inference and retrieval configuration used for KbSD.

Category	Parameter	Value
Generation	Max prompt length	4096 tokens
	Max response length	3000 tokens
	Temperature (evaluation)	0.0 (greedy)
	Temperature (training)	0.7
Retrieval	Retriever model	E5-Base-V2
	Knowledge source	Wikipedia (2018)
	Top-k passages	3
	Passage truncation	Middle (256 tokens)

B.4 Training Configuration

Training is conducted on 8×NVIDIA A100 (80GB) GPUs using Fully Sharded Data Parallel (FSDP) under Ubuntu 22.04.5 LTS. The software environment uses Python 3.11.11, Conda 23.5.2, PyTorch 2.6.0, HuggingFace Transformers 4.51.3, and spaCy 3.8.4. Table 4 lists the final hyperparameters. The boundary thresholds are selected on the validation set to maximize separation between known and unknown queries.

Table 4: Training configuration used for KbSD.

Category	Hyperparameter	Value
Optimization	Optimizer	AdamW
	Learning rate	1×10^{-6}
	LR scheduler	Cosine decay
	Warmup ratio	0.285
	Query batch size	128
	Training steps	300
GRPO	Group size (G)	5
	KL coefficient (β)	0.001
	Clip threshold (ϵ)	0.2
	Advantage normalization	Token-level
	Maximum retrieval iterations	5
KbSD	Process-reward weight (λ)	0.2
	Distillation weight (γ)	0.1
	Certainty threshold (τ_μ)	0.6
	Retrieval threshold (τ_ρ)	0.5

C Case Study

We organize the examples by the four behavioral regimes, followed by additional failure cases. Cases 1–2 are real evaluation traces from an archived Qwen2.5-7B-Instruct run (v19, checkpoint 80); they illustrate answer patterns relevant to the Q1/Q2 discussion, but their quadrant membership is not verified because the export lacks parametric probes and retrieval-quality scores. Cases 3–4 retain the original manuscript's Q3/Q4 examples and reported retrieval scores. These selected examples describe behaviors, not their population frequencies.

C.1 Q1: Integrated—Reconciling Internal and External Knowledge

The Integrated regime calls for agreement between reliable internal knowledge and useful retrieved evidence. Case 1 shows a correct comparison supported by the model's recorded summary. Verifying it as Q1 would additionally require a high parametric-certainty probe and a high retrieval-quality score; correctness alone is insufficient.

Case 1: Comparison example for the Q1 discussion

Source: hotpotqa, record 913 (zero-based); recorded F1: 100%.
Question: Are Benjamin Britten and Béla Bartók from the same country?
Reference: no
Search query: nationality of Benjamin Britten and Béla Bartók
Recorded model summary: The search results indicate that Benjamin Britten was British and Béla Bartók was Hungarian.
Answer spans in order: `no` → `no`
Analysis: The final answer matches the reference, and the recorded summary supplies the two nationalities used in the comparison. This illustrates a successful observed answer, but does not identify whether the decisive facts came from retrieval or parametric memory.

C.2 Q2: External—Answering from Retrieved Evidence

The External regime calls for useful retrieval to compensate for unreliable internal knowledge. Case 2 shows the entity-to-attribute reasoning pattern relevant to this regime. Its successful search-assisted answer does not establish a low internal prior; a verified Q2 assignment requires the missing probe and retrieval-quality measurements.

Case 2: Entity-to-attribute example for the Q2 discussion

Source: 2wikimultihopqa, record 604 (zero-based); recorded F1: 100%.
Question: What nationality is the director of film The Beach (Film)?
Reference: British
Search query: nationality of the director of The Beach film
Recorded model summary: I found that the director of the film "The Beach" is Danny Boyle, who is British.
Answer spans in order: `British` → `British`
Analysis: The summary links the film to its director and then to the requested nationality. The final answer matches the reference. Without a retrieval-free probe for this exact rollout, this success cannot by itself establish an External or Integrated quadrant assignment.

C.3 Q3: Refusal—Abstaining When Evidence Is Insufficient

Case 3 is the original manuscript's refusal example: the agent checks the query's constraints and ultimately abstains. The reported retrieval score is below $\tau_\rho = 0.5$; the original Q3 attribution is retained here, rather than inferred from the additional evaluation export.

Case 3 Q3: Refusal (Low Retrieval + Low Prior)
Distillation Policy: Forward KL

Question: *Which conductor, famous for leading orchestras using historical instruments, led a Verdi opera with a Moorish protagonist, collaborating with the Mahler Chamber Orchestra in the capital of the country where Frederick, heir to the Danish throne, was born and later married, though not in the same year?* **GT:** Invalid question

Rollout Summary (Outcome 0.50 IDK reward, Retrieval 0.363): The model checked each constraint against retrieved evidence, which remained contradictory and incomplete. It briefly proposed a candidate, then reconsidered and answered *"I don't know."*

Analysis: The mutually incompatible constraints make the question unanswerable; KbSD inspects the evidence constraint by constraint, identifies the gap, and abstains rather than hallucinating—the principled abstention that the Refusal quadrant's forward objective is designed to preserve.

C.4 Q4: Internal—Avoiding Unnecessary Reliance on Retrieval

Case 4 is the original manuscript's Internal example. It reaches the reference answer despite low-scoring retrieval, while also illustrating unnecessary search. The original Q4 attribution is retained; the example does not demonstrate perfectly efficient source selection.

Case 4 Q4: Internal (Low Retrieval + High Prior)
Distillation Policy: Bidirectional KL

Question: *HMAS "Platypus" was a submarine depot ship operated by the Royal Australian Navy (RAN). HMAS "AE1" was an "E"-class submarine of which organization?*
GT: Royal Australian Navy (RAN)

Rollout Summary (Outcome 2.00, Retrieval 0.496): The answer is implicitly stated in the question itself, yet the model continued consulting retrieved evidence before converging. The retrieved passages initially contained tangential information (torpedo specifications of AE1); the model eventually inferred through chained reasoning that HMAS AE1 belongs to the Royal Australian Navy.

Analysis: This Q4 case (high parametric prior, low retrieval quality) reveals a residual challenge: KbSD still occasionally over-relies on retrieval even when the parametric answer is accessible. The bidirectional objective guides the model to the correct final answer, yet its continued reliance on retrieval indicates remaining headroom in calibrating retrieval *efficiency* under the Internal quadrant.

D More Case Studies

D.1 Additional Evaluation Traces: Failure Modes

Cases 5–8, like Cases 1–2, come from the independent archived 7B run (v19, checkpoint 80), evaluated with greedy decoding; they are not attributed to the main-table checkpoint. The exported

records contain the question, reference answer, model summary, and answer spans, but not the retrieved passages or per-query boundary scores. Model summaries are therefore reported claims, not independently verified retrieval evidence. Repeated and malformed answer spans are retained, and no quadrant labels are assigned to these additional cases.

Case 5: Over-refusal after identifying the reference answer

Source: triviaqa, record 923 (zero-based); recorded F1: 0%.
Question: Who wrote the book of verses 'Now We Are Six'?
Reference: A A Milne
Search query: author of Now We Are Six
Recorded model summary: I found that the book "Now We Are Six" was written by A.A. Milne.
Answer spans in order: `answer: A.A. Milne` → `I don't know`
Analysis: The summary and first answer span identify the reference author, but the final answer is an explicit refusal. This is a concrete over-refusal failure: a refusal can reduce the reported unreliability residual while losing an otherwise correct answer.

Case 6: Refusal following an incorrect intermediate candidate

Source: frames, record 648 (zero-based); recorded F1: 0%.
Question: How old was Joel McHale the first time the Philadelphia Waterdogs won the PLL Championship?
Reference: Joel McHale was 50 years old.
Search query: Joel McHale age when Philadelphia Waterdogs won PLL Championship
Recorded model summary: I found that the Philadelphia Waterdogs won the PLL Championship in 2017. Joel McHale was 33 years old in 2017.
Answer spans in order: `33` → `I don't know`
Analysis: The recorded trajectory proposes 33 before ending in refusal; the reference answer is 50. The final refusal avoids retaining that incorrect candidate, but the log does not explain the revision or prove principled uncertainty assessment. The query remains answerable under the benchmark reference.

Case 7: Incorrect answer retained after search

Source: nq, record 764 (zero-based); recorded F1: 0%.
Question: how many episodes are there in poldark season 3
Reference: 9
Search query: number of episodes in poldark season 3
Recorded model summary: Search results indicate that there are 6 episodes in Poldark season 3.
Answer spans in order: `6` → `6`
Analysis: The final answer repeats the value asserted in the summary and disagrees with the reference. A search call and an affirmative summary therefore do not ensure correctness. Because the retrieved passages are absent from this export, the error cannot be attributed specifically to retrieval or reasoning.

Case 8: Malformed final answer after a correct candidate

Source: bamboogle, record 117 (zero-based); recorded F1: 0%.
Question: What rocket was used for the mission that landed the first humans on the moon?
Reference: Saturn V
Search query: rocket used for the first moon landing
Recorded model summary: The search returned information that the Saturn V rocket was used for the Apollo 11 mission, which landed the first humans on the moon.
Answer spans in order: `answer: Saturn V` → `answer> >`
Analysis: The summary and first answer span contain the reference answer, but the final extracted span is malformed. Under final-answer parsing, the recorded F1 is zero. This illustrates an output-format failure rather than explicit, calibrated abstention.

E Additional Behavioral Analysis

Quadrant-Adaptive Routing Controls Behavioral Constraint. The per-quadrant choice is motivated by the *shape* of valid reasoning—concentrated for *Integrated* and diverse for *Refusal*—but this is difficult to verify directly. We instead test the resulting *behavioral constraint*: reverse-only Q1 is the loosest, penalizing only student mass outside the teacher's support, whereas bidirectional Q2/Q4 is the tightest. Consistent with this hierarchy, Figure 5 shows significantly greater divergence for Q1

than Q4 in both directions ($p < 0.01$), confirming that the assigned objective governs deviation from the teacher. This direct mechanistic evidence complements the task-level ablation and verifies the intended effect of quadrant-adaptive routing.

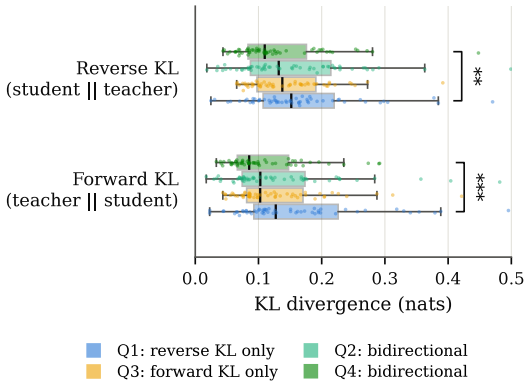

Figure 5: Student–teacher KL divergence on held-out rollouts. Q1 shows greater divergence than Q4 in both directions ($^{**}p < 0.01$, $^{***}p < 0.001$), consistent with weaker behavioral constraint.

Effect of Model Scale. At 7B, KbSD reaches a lower average (41.15 vs. 47.63 at 3B) while retaining a comparable gain over KbPO (16.84 vs. 16.69 points), showing that dense supervision remains effective across scales.

Adaptive Refusal Behavior. Inspecting rollouts, we observe difficulty-adaptive abstention: refusal is least frequent where at least one knowledge source is typically sufficient (e.g., NQ, TriviaQA) and most frequent where compositional reasoning over rare knowledge is required (e.g., MuSiQue, FRAMES). As answerability falls, abstention rises and unreliability drops accordingly—the qualitative signature of calibrated uncertainty, arising exactly where sparse rewards are least informative. This indicates the model learns to recognize its own knowledge boundary rather than applying a fixed abstention rate.

Balancing Helpfulness and Reliability. The reliability gains do not come from indiscriminate abstention. In rollouts, abstention on the in-domain benchmarks remains moderate, while their unreliability still drops sharply relative to KbPO (2WikiMQA: 70.50 → 45.89; HotpotQA: 57.60 → 41.57). This balance indicates that the quadrant-adaptive objective separates queries that warrant confident answering from those that warrant cautious abstention, rather than shifting all probability mass toward refusal.

F Distillation-Direction Ablation

Table 5 ablates the two design choices of KbSD on Qwen2.5-3B over five representative benchmarks.

Table 5: Ablation on Qwen2.5-3B (Unreliability Rate, %, lower is better). $-\mathcal{L}_{\text{distill}}$ removes the distillation term, which reduces objective to KbPO; Fixed-FKL and Fixed-RKL replace quadrant-adaptive routing with a single direction.

Dataset	KbSD	$-\mathcal{L}_{\text{distill}}$	Fixed-FKL	Fixed-RKL
TriviaQA	**34.85**	39.90	39.70	50.50
2WikiMQA	45.89	70.50	**44.00**	68.90
HotpotQA	**41.57**	57.60	45.50	59.90
NQ	**64.43**	68.50	66.47	72.10
PopQA	**45.55**	73.10	50.00	70.20
Average	**46.46**	61.92	49.13	64.32

Dense Distillation Drives Most of the Gain. Removing the distillation loss ($-\mathcal{L}_{\text{distill}}$), which reduces the objective exactly to KbPO, degrades the average from 46.46 to 61.92. Dense process-level supervision is therefore the dominant source of KbSD's improvement.

No Single Divergence Direction Fits All Quadrants. Replacing quadrant-adaptive routing with a fixed direction is consistently worse on average: Fixed-FKL reaches 49.13 and Fixed-RKL 64.32. The failure of Fixed-RKL is the most instructive: a purely mode-seeking objective collapses the diverse abstention behaviors required in the Refusal quadrant, degrading TriviaQA from 34.85 to 50.50. Fixed-FKL remains competitive and is even slightly better on 2WikiMQA (44.00 vs. 45.89), which is unsurprising given that mass-covering imitation suffices when concentrated integration patterns dominate; however, it trails KbSD on the other four benchmarks and by 2.7 points on average. Together, these results indicate that the distillation signal itself contributes the largest share of the gain, while quadrant-adaptive routing yields a consistent further improvement and, crucially, avoids the severe failure mode of a purely mode-seeking objective.

G Conditional F1 and Matched-Coverage Analysis

G.1 Conditional F1 at Natural Coverage

We complement the overall F1 and refusal rates in Section 4.4 with answer quality conditioned on non-refusal. For method m and dataset D, let $A_m = \{i \in D : m \text{ does not refuse query } i\}$, and let $f_{m,i} \in [0, 1]$ be the per-query answer F1. Conditional F1 is

$$\mathrm{CF1}_m(D) = \frac{100}{|A_m|} \sum_{i \in A_m} f_{m,i}. \tag{24}$$

If refused responses receive zero F1 under the evaluation protocol, this equals overall F1 divided by coverage, $\mathrm{F1}_m(D)/(1 - r_m(D))$, where overall F1 is expressed in percent and $r_m(D) \in [0, 1]$ is the refusal rate. Conditional scores are reported to two decimal places.

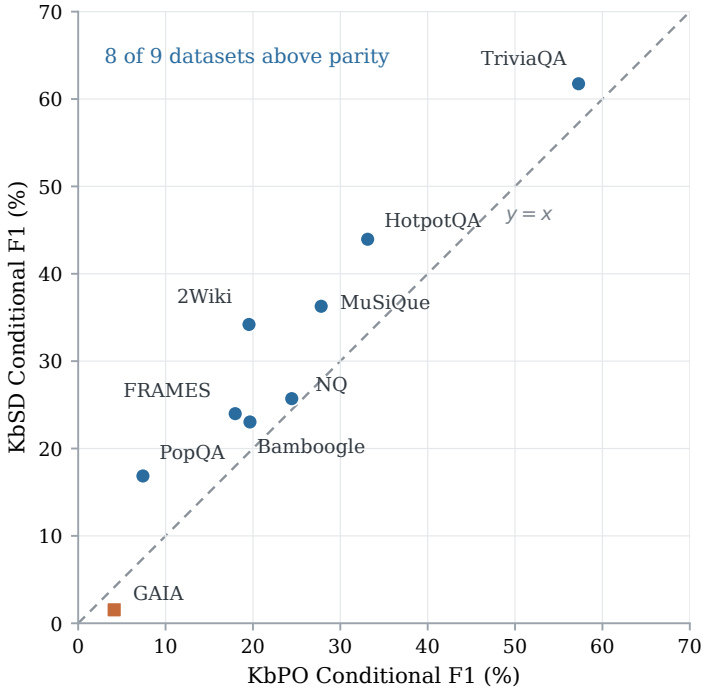

Figure 6: Conditional F1 on non-refused responses for Qwen2.5-3B: KbPO on the horizontal axis and KbSD on the vertical axis. Each point represents one dataset; equal axis scales and the $y = x$ line indicate parity. KbSD scores higher on eight datasets; GAIA, shown as an orange square, is the exception. Each method is evaluated on its own non-refused subset, without matching coverage.

Results. Figure 6 shows higher Conditional F1 for KbSD on eight of nine datasets. The largest absolute increase occurs on 2Wiki (19.55 to 34.20), followed by HotpotQA (33.13 to 43.95). Bamboogle also improves conditionally (19.66 to 23.04), despite its decrease in overall F1 (16.66 to 15.63): higher quality among the retained answers coexists with lower coverage. GAIA remains the exception, falling from 4.12 to 1.54, consistent with the accuracy cost observed in the main analysis.

Interpretation. These results show that KbSD's non-refused answers have higher average overlap with reference answers on most benchmarks. However, the two methods can answer different subsets, so Conditional F1 does not isolate improved reasoning from selective abstention or demonstrate superiority at equal coverage. The matched-coverage analysis below controls for the retained fraction, while paired response-level analysis is still needed to identify which answers are replaced by additional refusals.

G.2 Consistency-Ranked Matched-Coverage Comparison

To control for differing refusal rates, we compare both methods at KbSD's natural coverage on each dataset. KbPO predictions are ranked by answer consistency estimated from 10 samples per query, used as a confidence proxy, and the retained fraction is adjusted to match KbSD's coverage. We report mean answer F1 over the retained responses, rather than over all queries. KbSD's retained set and scores remain unchanged from Figure 6; only the KbPO comparator changes.

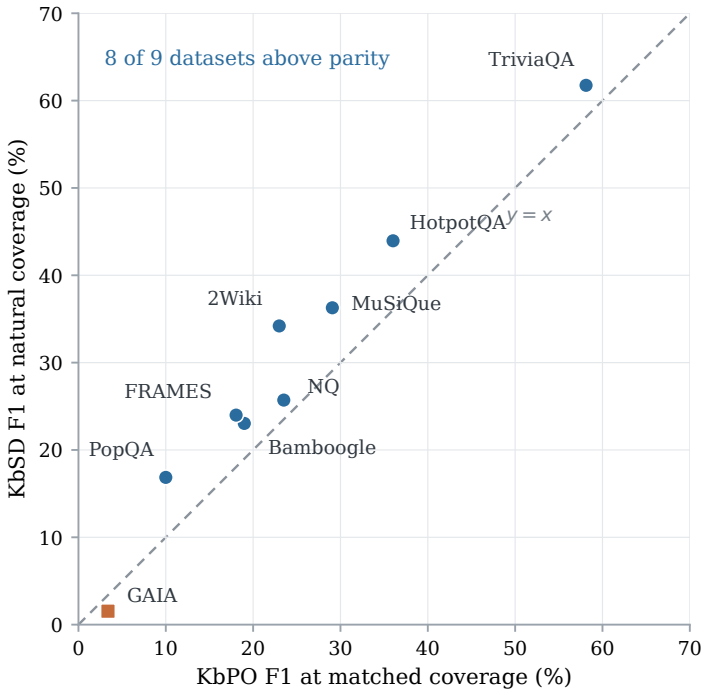

Figure 7: F1 on retained responses at matched coverage for Qwen2.5-3B. The horizontal axis shows consistency-ranked KbPO at KbSD's natural coverage; the vertical axis shows KbSD. KbSD values are identical to Figure 6. Eight datasets lie above $y = x$; GAIA (orange square) lies below. Equal coverage matches the retained fraction, not necessarily the identities of retained queries.

Results. KbSD outperforms consistency-ranked, coverage-matched KbPO on eight of nine benchmarks (Figure 7). For example, F1 increases from 22.99 to 34.20 on 2Wiki and from 10.00 to 16.86 on PopQA. GAIA remains the exception: matched-coverage KbPO reaches 3.38, compared with 1.54 for KbSD. The direction of the comparison agrees with the natural-coverage Conditional F1 results across all nine datasets. Because the KbSD scores are reused, these are complementary comparisons rather than independent replications.

Scope of the Comparison. Matching coverage removes the difference in the number of retained responses as an explanation for the observed F1 gap under this selection protocol. Answer consistency, however, is an imperfect confidence proxy: a model can consistently produce an incorrect answer, and the comparison depends on the quality of the ranking. These results therefore support KbSD over this consistency-ranked KbPO baseline at the evaluated coverage levels, rather than over every possible abstention policy. The methods may retain different queries; the analysis does not isolate reasoning gains from better selection. Without uncertainty estimates or repeated runs, the recurring GAIA exception also does not establish statistical significance.