MIXTURE RAG-KD: CONTROLLING RETRIEVAL-INDUCED SHIFTS FOR ROBUST KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

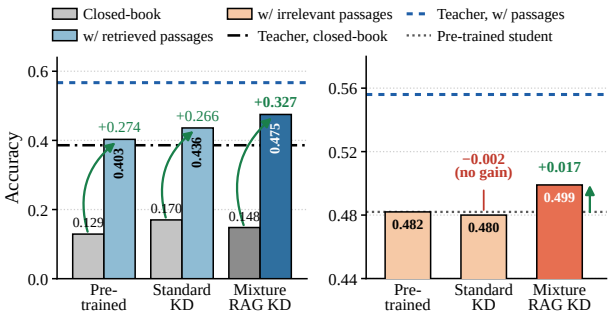
ABSTRACT

Knowledge distillation (KD) transfers the capabilities of large language models (LLMs) to smaller student models, but the limited parametric capacity of these students can constrain achievable performance. We argue that students should be distilled not only to imitate a teacher's parametric behavior, but also to effectively utilize external knowledge through retrieval-augmented generation (RAG). To this end, we formulate RAG-aware Knowledge Distillation, a framework that integrates retrieval into KD so that student models can effectively utilize information from retrieved contexts. We first identify a limitation of Vanilla RAG-KD, which directly distills the teacher's retrieval-conditioned distribution: under imperfect retrieval, its effectiveness is sensitive to retrieval quality. To address this issue, we propose Mixture RAG-KD, which combines the complementary strengths of the teacher's parametric and retrieval-conditioned distributions. We theoretically show why mixture supervision can be preferable to relying on either teacher alone by characterizing conditions under which an intermediate mixture achieves lower target risk than both endpoints. Building on this analysis, we develop global and instance-wise strategies for selecting the mixing coefficient α by minimizing forward-KL or squared-L_2 target risk. Experiments across general QA and diverse noisy-retrieval settings show that Mixture RAG-KD consistently improves over Vanilla RAG-KD and remains more robust as training-time retrieval noise increases.

1 INTRODUCTION

Large language models (LLMs) have achieved strong generative performance across diverse tasks but incur substantial computational and memory costs that limit practical deployment (Achiam et al., 2023; Yang et al., 2025; Team et al., 2023; Liu et al., 2024). As resource-constrained applications such as robotics and edge or on-device systems gain prominence (Xu et al., 2024b; Zheng et al., 2025), knowledge distillation (KD) (Hinton et al., 2015) has emerged as a key approach for transferring task-specific capabilities to compact models (Datta et al., 2025; Xu et al., 2025).

Despite this potential, scaling law studies (Kaplan et al., 2020) imply an intrinsic upper bound on the performance achievable by smaller models. To overcome this bottleneck without increasing model capacity, it becomes necessary to complement the student with external, non-parametric knowledge sources. Retrieval-augmented generation (RAG) (Lewis et al., 2020) provides a natural solution by enabling lightweight models to access and utilize external knowledge at inference time, effectively compensating for their limited memorization capacity. As shown in Figure 1a, standard KD leaves a per-



(a) Effect of retrieved passages (b) Irrelevant passages

Figure 1: Motivation for RAG-aware KD. Retrieval lifts every student, but standard KD gains nothing under irrelevant passages while Mixture RAG KD does.

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

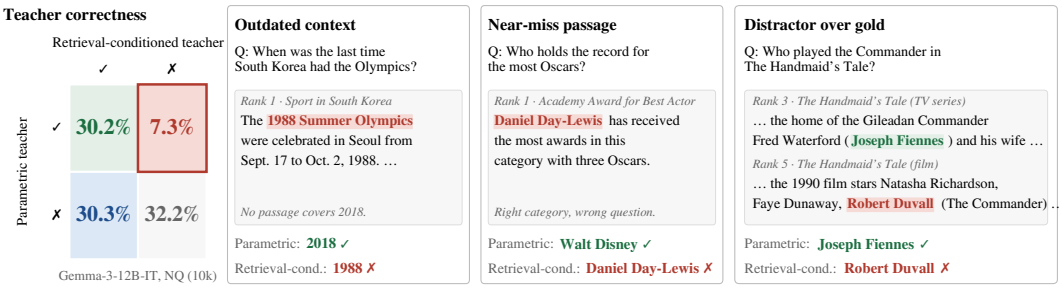


Figure 2: **Retrieval-induced errors of the teacher. Left:** Teacher correctness on 10K NQ questions. Retrieval corrects 30.3% of parametric errors, but causes errors in 7.3% of cases where the parametric teacher was originally correct. **Right:** Examples of noisy retrieval that mislead the teacher; red highlights copied incorrect evidence and green the retrieved gold answer.

sistent parametric gap between the teacher and student, whereas incorporating external retrieval substantially narrows this gap. This observation suggests that retrieval can serve as an effective non-parametric complement to student models. Although RAG can complement the limited parametric knowledge of compact students, its effectiveness depends on retrieval quality, as retrieved contexts may be informative, irrelevant, or misleading (Yoran et al., 2024; Shi et al., 2023; Cuconasu et al., 2024). Thus, an effective RAG student should exploit useful evidence while remaining robust to imperfect retrieval; however, standard KD does not directly address this retrieval-conditioned behavior and exhibits limited robustness under noisy contexts (Figure 1b).

These observations motivate a RAG-aware knowledge distillation framework that enables compact students to exploit useful retrieval while remaining robust to imperfect retrieval. The most direct approach is *Vanilla RAG-KD* (Eq. 2), which distills the teacher distribution conditioned on retrieved context. However, as shown in Figure 2, retrieval can both correct and corrupt the teacher's parametric prediction, causing erroneous retrieval-conditioned supervision to be transferred to the student.

We therefore view the teacher's parametric and retrieval-conditioned predictions as complementary sources of supervision, neither of which is uniformly preferable under imperfect retrieval. Based on this observation, we propose *Mixture RAG-KD* (Eq. 3), which balances the two sources through a mixing coefficient. We characterize conditions under which an intermediate mixture achieves lower target risk than either endpoint and develop practical coefficient selection strategies based on forward-KL and squared-L_2 risks. We then evaluate whether the resulting students can better exploit informative retrieval while remaining robust to noisy retrieval at inference time.

Our contributions are summarized as follows:

- **RAG-aware Knowledge Distillation.** We formulate a RAG-aware knowledge distillation framework that incorporates retrieval-conditioned teacher supervision and explicitly considers the effect of imperfect retrieval on the distillation target.
- **Mixture RAG-KD and theoretical analysis.** We propose Mixture RAG-KD, which combines parametric and retrieval-conditioned teacher supervision, and characterize conditions under which an intermediate mixture improves upon both endpoint targets.
- **Risk-based coefficient selection and evaluation.** We develop global and instance-level mixing strategies based on forward-KL and squared-L_2 target risks, and evaluate their effectiveness across diverse general-QA and noisy-retrieval settings.

2 Preliminary

2.1 Knowledge Distillation

We consider a dataset of input-output pairs $\mathcal{D} \coloneqq \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$. For each sample $(x, y) \in \mathcal{D}$, the output $y = (y_1, \ldots, y_L)$ is a token sequence of length L over a fixed vocabulary $\mathcal{V}$, where $y_l \in \mathcal{V}$ and $y_{<l} = (y_1, \ldots, y_{l-1})$. Given an input x and the preceding output tokens $y_{<l}$, a teacher model p and a student model q_θ define next-token distributions $p(\cdot \mid x, y_{<l})$ and $q_\theta(\cdot \mid x, y_{<l})$ over the vocabulary $\mathcal{V}$, respectively.

Design on Metric and Divergence: Knowledge distillation (KD) aligns the teacher and student next-token distributions. The conventional formulation, referred to as **Standard KD**, minimizes

$$\mathcal{L}_{\mathrm{KD}}(\theta) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L} D\left(p_\phi(\cdot \mid x, y_{<l}) \parallel q_\theta(\cdot \mid x, y_{<l})\right)\right], \tag{1}$$

where D denotes a divergence between two next-token distributions over $\mathcal{V}$. The choice of divergence D is one of the key design components of KD. For notational simplicity, at each token position l, we write $p \coloneqq p_\phi(\cdot \mid x, y_{<l})$ and $q_\theta \coloneqq q_\theta(\cdot \mid x, y_{<l})$ for the teacher and student next-token distributions, respectively. Prior works have predominantly employed forward KL (FKL) (Hinton et al., 2015; Kim & Rush, 2016), defined as $D_{\mathrm{FKL}}(p, q_\theta) \coloneqq \sum_{w\in\mathcal{V}} p(w) \log \frac{p(w)}{q_\theta(w)}$. Reverse KL (RKL) (Gu et al., 2024) reverses the direction of the divergence and is defined as $D_{\mathrm{RKL}}(p, q_\theta) \coloneqq D_{\mathrm{FKL}}(q_\theta, p)$. FKL and RKL are commonly associated with mode-covering and mode-seeking behaviors, respectively. Recent works have also explored generalized Jensen–Shannon divergence (JSD) (Agarwal et al., 2024), which introduces a weighted mixture of p and q_θ as an intermediate distribution.

Design on Intermediate Distribution: To mitigate the teacher–student distribution gap, recent approaches introduce intermediate targets that provide smoother supervision than directly matching the teacher distribution. DistiLLM (Ko et al., 2024) adopts skew KL and skew reverse KL divergences that incorporate mixtures of p and q_θ. TAID (Shing et al., 2025) constructs an adaptive intermediate target that gradually shifts from the student distribution toward the teacher distribution during training. Recently, AMiD (Shin et al., 2026) formalizes such intermediate distributions as *assistant distributions* and generalizes these approaches into an assistant-based KD framework.

2.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) augments an input x with top-K passages $\mathcal{C}_K(x)$ retrieved from a corpus $\mathcal{C}$, forming a context $c = z(\mathcal{C}_K(x))$ where $z(\cdot)$ denotes a context aggregation function. The model then predicts each token from the retrieval-conditioned distribution $p_\phi(\cdot \mid x, c, y_{<l})$.

We refer to the retrieval as *noisy* when the retrieved passages contain irrelevant, misleading, or conflicting information, and denote the resulting context by $\tilde{c}$. The corresponding teacher distribution becomes $p_\phi(\cdot \mid x, \tilde{c}, y_{<l})$. Such noisy contexts can distort the generation distribution and degrade generation performance (Yoran et al., 2024; Shi et al., 2023; Yu et al., 2024).

Prior work has improved robustness to noisy retrieval through self-reflection and explicit denoising (Asai et al., 2024; Wei et al., 2025), relevance-aware utilization of retrieved evidence (Wang et al., 2024; Kim & Lee, 2024; Yan et al., 2025), and filtering, reranking, or compression of retrieved contexts (Yan et al., 2024; Xu et al., 2024a; Wang et al., 2025). These approaches primarily improve robustness at the retrieval or generation stage of RAG. To the best of our knowledge, prior work has not explicitly considered the limitations of directly distilling from a teacher distribution conditioned on noisy retrieved contexts. We address this issue by designing a distillation objective that combines retrieval-conditioned and parametric teacher knowledge, providing robust supervision for the student under noisy retrieval.

3 RAG-aware Knowledge Distillation

Problem Formulation. We formulate RAG-aware Knowledge Distillation (RAG-KD) for transferring retrieval-conditioned knowledge from the teacher to the student. Following the notation introduced in Section 2, we refer to *parametric teacher distribution* p and $p_c \coloneqq p_\phi(\cdot \mid x, c, y_{<l})$ as the *retrieval-conditioned teacher distribution*, respectively, and to $q_\theta(\cdot \mid x, c, y_{<l})$ as the *retrieval-conditioned student distribution*. For each input-output pair $(x, y) \in \mathcal{D}$, retrieval augments the example with a context $c = z(\mathcal{C}_K(x))$. For simplicity, throughout the RAG-KD formulation, we continue to use $\mathcal{D}$ to denote the empirical distribution over the resulting triples (x, y, c).

3.1 A Naive Approach: Vanilla RAG-KD

The most direct formulation of RAG-aware knowledge distillation is to use the teacher's retrieval-conditioned distribution as the supervision target. Given the retrieval context c, Vanilla RAG-KD

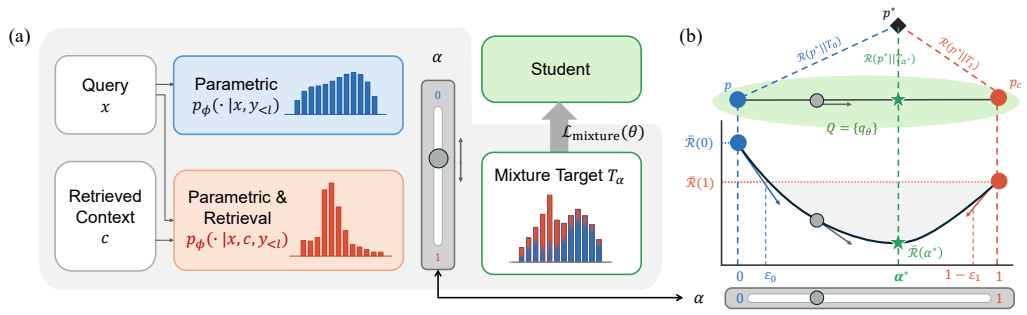


Figure 3: **Mixture RAG-KD and when it improves upon the endpoints.** (a) Overview. Mixture RAG-KD combines the parametric teacher p and retrieval-conditioned teacher p_c into $T_\alpha = (1-\alpha)p + \alpha p_c$, which is distilled into the retrieval-conditioned student via Eq. (3).(b) Geometry of Theorem 3.1. *Top:* as α varies, T_α traces the segment from p to p_c, and the target risk $\bar{\mathcal{R}}(\alpha) = \mathcal{R}(p^*, T_\alpha)$ measures its discrepancy from the optimal target p^*. *Bottom:* when retrieval is useful but imperfect, moving away from the parametric teacher initially reduces risk, while relying fully on the retrieval-conditioned teacher is also suboptimal. These two behaviors are captured by $\bar{\mathcal{R}}'_+(0) < 0$ and $\bar{\mathcal{R}}'_-(1) > 0$, respectively, yielding an interior optimum $\alpha^* \in (0, 1)$.

trains the student to match the *retrieval-conditioned teacher distribution* under the same context. Its objective is defined as

$$\mathcal{L}_{\text{vanilla}}(\theta) := \mathbb{E}_{(x,y,c)\sim\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L} D\left(p_\phi(\cdot \mid x, c, y_{<l}) \,\|\, q_\theta(\cdot \mid x, c, y_{<l})\right)\right], \tag{2}$$

However, Vanilla RAG-KD assumes the retrieval-conditioned teacher provides reliable supervision. Under imperfect retrieval, however, noisy contexts can distort the teacher distribution and transfer retrieval-induced errors to the student. Since Vanilla RAG-KD directly trains the student to match this distribution, such retrieval-induced errors can be transferred to the student. Therefore, under imperfect retrieval, exactly matching the retrieval-conditioned teacher distribution may not provide desirable supervision from a robustness perspective.

Limitation of Vanilla RAG-KD. Integrating RAG with KD requires careful consideration of retrieval quality. In the absence of perfect retrievers, retrieved contexts often contain irrelevant or erroneous information, which can induce hallucinations (Yoran et al., 2024; Cuconasu et al., 2024). As shown in Figure 2, we identify cases where even a strong teacher model, despite having correct parametric knowledge, is biased by faulty external evidence and produces incorrect outputs. In such cases, retrieval-induced errors in the teacher can be directly transferred to the student. Consistent with this observation, Figure 4 shows that the advantage of Vanilla RAG-KD over Standard KD decreases as training-time retrieval noise increases, revealing its dependence on retrieval quality. Additional experimental details and analyses are provided in Appendix C.

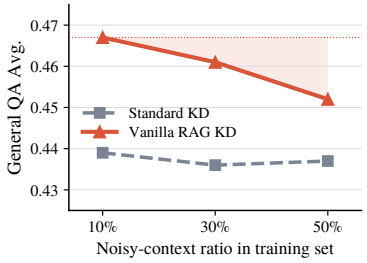


Figure 4: **Vanilla RAG-KD under noisy training retrieval.** General QA average as the fraction of training instances with a noisy retrieved context grows.

3.2 Proposed Method: Mixture RAG-KD

The parametric teacher p and the retrieval-conditioned teacher p_c provide two potentially complementary sources of supervision. While p reflects the teacher's parametric knowledge, p_c additionally incorporates information from retrieved passages. Under imperfect retrieval, their relative utility can vary across examples. We therefore propose *Mixture RAG-KD*, which combines the two teacher distributions through a mixing coefficient α. We define the mixture target as $T_\alpha := (1-\alpha)p + \alpha p_c$, where $\alpha \in [0, 1]$. The endpoints recover the parametric teacher at $\alpha = 0$ and Vanilla RAG-KD at

$\alpha = 1$, while intermediate values combine supervision from both teachers. Using this mixture target, we define the Mixture RAG-KD objective as

$$\mathcal{L}_{\text{mixture}}(\theta) \coloneqq \mathbb{E}_{(x,y,c)\sim\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} D\left(T_\alpha(\cdot \mid x, c, y_{<l}) \parallel q_\theta(\cdot \mid x, c, y_{<l})\right) \right], \tag{3}$$

Since the two teachers may provide better supervision in different situations, an appropriate mixture may improve upon relying on either source alone. This raises a central question: *under what conditions can an intermediate mixture provide better supervision than both endpoint targets?*

When can mixture supervision improve upon the endpoint targets? The mixture target interpolates between the parametric and retrieval-conditioned teachers, but interpolation alone does not guarantee better supervision. We characterize a sufficient condition under which an intermediate mixture achieves lower target risk than both endpoints.

Theorem 3.1. *Let p^* denote the optimal target distribution, and let $Q := \{q_\theta \mid \theta \in \Theta\}$ be a set of student models. We define the token-level risk functional between two distributions u and v as*

$$\mathcal{R}(u, v) \coloneqq \mathbb{E}_{(x,y,c)\sim\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} D\left(u(\cdot \mid x, c, y_{<l}) \parallel v(\cdot \mid x, c, y_{<l})\right) \right]. \tag{4}$$

For each $\alpha \in [0, 1]$, we define the optimal student model distilled from the mixture target T_α as

$$q^*_\alpha \in \arg\min_{q_\theta \in Q} \mathcal{R}(T_\alpha, q_\theta). \tag{5}$$

In particular, q^*_0 and q^*_1 are the optimal student models corresponding to the endpoint targets $T_0 = p$ and $T_1 = p_c$, respectively.

We define $\bar{\mathcal{R}}(\alpha) \coloneqq \mathcal{R}(p^*, T_\alpha)$. Suppose that (1) Q is expressive enough to include T_α for every $\alpha \in [0, 1]$; and (2) $\bar{\mathcal{R}}(\alpha)$ is finite and continuous on $[0, 1]$, and $\bar{\mathcal{R}}'_+(0)$ and $\bar{\mathcal{R}}'_-(1)$ exist.

If $\bar{\mathcal{R}}'_+(0) < 0$ and $\bar{\mathcal{R}}'_-(1) > 0$, then every $\alpha^* \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}(\alpha)$ satisfies $\alpha^* \in (0, 1)$, and

$$\mathcal{R}(p^*, q^*_{\alpha^*}) < \min\left\{\mathcal{R}(p^*, q^*_0), \mathcal{R}(p^*, q^*_1)\right\}. \tag{6}$$

The two endpoint conditions have complementary interpretations. The condition $\bar{\mathcal{R}}'_+(0) < 0$ indicates that introducing a retrieval-conditioned component improves upon the parametric target, whereas $\bar{\mathcal{R}}'_-(1) > 0$ indicates that partially restoring the parametric component improves upon the fully retrieval-conditioned target. Together, they characterize a useful-but-imperfect retrieval regime in which neither endpoint provides optimal supervision, as illustrated in Figure 2. Figure 3 (b) provides a geometric illustration of this regime, showing how the target risk decreases when moving inward from either endpoint and attains its minimum at an interior mixture. The proof is provided in Appendix A.1. We next derive an empirical characterization of these conditions and examine whether this regime arises in practice.

Interpreting the endpoint condition under imperfect retrieval. To assess the endpoint conditions empirically, we use the gold answers provided in the dataset as reference supervision for evaluating the mixture targets. For each answer token y_l, we define the empirical reference as $\widehat{p}^* = \delta_{y_l}$. Under forward-KL target risk, the endpoint conditions can then be expressed as follows; the derivation is provided in Appendix A.2.

$$\bar{\mathcal{R}}'_{\text{FKL},+}(0) < 0 \iff \mathbb{E}_{\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} \frac{p_c(y_l)}{p(y_l)} \right] > 1, \tag{7}$$

$$\bar{\mathcal{R}}'_{\text{FKL},-}(1) > 0 \iff \mathbb{E}_{\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} \frac{p(y_l)}{p_c(y_l)} \right] > 1. \tag{8}$$

The first condition captures the benefit of incorporating information conditioned on retrieval, while the second indicates that the complete transfer of p_c is suboptimal. These conditions are consistent with imperfect retrieval, where retrieved passages

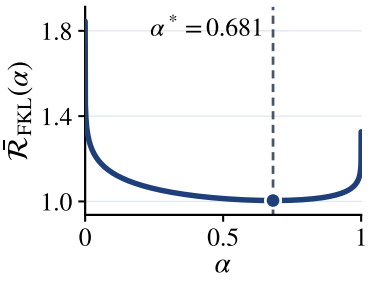


Figure 5: **Target-risk analysis**

provide useful information but may also introduce noise or misleading evidence. We examine whether this regime arises in practice by evaluating $\bar{\mathcal{R}}_{\mathrm{FKL}}(\alpha)$ over $\alpha \in [0,1]$ on a 10K-example subset of the Natural Questions training set(Figure 5). The observed risk curve attains its minimum at an interior mixture with lower risk than either endpoint, empirically instantiating the regime characterized by the Theorem and aligning with the imperfect-retrieval behavior illustrated in Figure 2.

3.3 Mixing Strategies for Mixture RAG-KD

The preceding analysis establishes conditions under which an intermediate mixture can improve upon both endpoint targets. We now consider how to select the mixing coefficient α. We study static mixing and two risk-based strategies using forward KL and squared L_2. For the risk-based strategies, we use the empirical reference $\widehat{p}^* = \delta_{y_l}$ in the risk functional defined above.

Static Mixing. Static mixing uses a fixed coefficient α for all instances and token positions. We treat $\alpha \in [0,1]$ as a hyperparameter, sample it at intervals of 0.2, and report the best-performing value. While simple and stable, this strategy assumes uniform reliability of retrieved contexts and may be suboptimal when retrieval quality varies across inputs.

Forward-KL Risk-Optimal Mixing. We select the global mixing coefficient by minimizing the empirical forward-KL target risk:

$$\alpha_{\mathrm{FKL}}^{\mathrm{global}} \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}_{\mathrm{FKL}}(\widehat{p}^*, T_\alpha). \tag{9}$$

Since the forward-KL risk is convex in α, we exploit this property to obtain the optimal coefficient using bisection. Let $g(\alpha) := \frac{d}{d\alpha}\bar{\mathcal{R}}_{\mathrm{FKL}}(\widehat{p}^*, T_\alpha)$. Convexity implies that $g(\alpha)$ is monotonically nondecreasing. We first examine the derivative at the two endpoints. If neither endpoint is optimal, the optimum lies in the interior and satisfies $g(\alpha) = 0$, which we solve by bisection. For instance-level mixing, we apply the same procedure independently to each training example, using a single coefficient shared across its answer positions. The explicit derivative and full derivation are provided in Appendix B.1.

Algorithm 1: FKL bisection

Require: $g(\alpha)$, iterations K
if $g(0) \geq 0$ **then return** 0
end if
if $g(1) \leq 0$ **then return** 1
end if
$a \leftarrow 0,\ b \leftarrow 1$
for $k = 1, \ldots, K$ **do**
  $m \leftarrow (a+b)/2$
  if $g(m) < 0$ **then**
    $a \leftarrow m$
  else
    $b \leftarrow m$
  end if
end for
return $(a+b)/2$

Squared-L_2 Risk-Optimal Mixing. We similarly select the global mixing coefficient by minimizing the empirical squared-L_2 target risk:

$$\alpha_{L_2}^{\mathrm{global}} \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}_{L_2}\left(\widehat{p}^*, T_\alpha\right). \tag{10}$$

Since the squared-L_2 risk is quadratic in α, the optimal mixing coefficient can be obtained in closed form, avoiding iterative optimization.

$$\alpha_{L_2}^{\mathrm{global}} = \mathrm{clip}\left(\frac{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\left\langle p_c - p,\, \widehat{p}^* - p\right\rangle\right]}{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\left\|p_c - p\right\|_2^2\right]}, 0, 1\right). \tag{11}$$

Geometrically, the unclipped coefficient corresponds to the projection of the displacement toward the empirical reference, $\widehat{p}^* - p$, onto the retrieval-induced direction, $p_c - p$, aggregated across examples and answer positions. Thus, α determines how far the target should move from p toward p_c to approach the empirical reference under squared-L_2 risk.

For instance-level mixing, we apply the same closed-form optimization independently to each training example, using a single coefficient shared across the entire answer span. The derivation and instance-level formulation are provided in Appendix B.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and Models. We conduct task-specific distillation on Natural Questions (NQ) (Kwiatkowski et al., 2019), a widely used open-domain QA benchmark. Prior work (Xu et al., 2025)[1] shows that effective distillation is possible even with small datasets ($\sim$1K), which we adopt as our default setting. We sample 1,000 training examples from NQ. For each question, we retrieve the top-5 passages from Wikipedia using DPR (Karpukhin et al., 2020) and construct training examples (x, y, c). To evaluate robustness to training-time retrieval noise, we additionally construct 1K training sets with 10%, 30%, and 50% noisy examples. Details of the noise construction are provided in Appendix C. We use Gemma-3-12B-IT (Team et al., 2025) as the teacher model and Gemma-3-1B-IT and Gemma-3-4B-IT as student models. To examine whether our KD methods remain effective when starting from an already retrieval-adapted student, we additionally evaluate them on students first fine-tuned with supervised learning on (x, y, c) triples. The training details of these SFT-initialized students are provided in the Appendix C.

Baselines. We evaluate five variants of Mixture RAG-KD: *Static*, *FKL-Global*, *FKL-Instance*, L_2*-Global*, and L_2*-Instance*. We compare them against **Standard KD** and the two endpoint baselines of Mixture RAG-KD: **Vanilla RAG-KD** ($T_{\alpha=1} = p_c$), which directly distills the retrieval-conditioned teacher, and the **Parametric-Teacher-KD** baseline ($T_{\alpha=0} = p$), where the student is conditioned on the retrieved context but supervised by the parametric teacher. We additionally include direct distillation from the empirical reference $\widehat{p}^*$ constructed from the gold answer, since our risk-based strategies select T_α by minimizing its target risk with respect to $\widehat{p}^*$.

Evaluation. We evaluate general QA performance on three open-domain QA benchmarks: NQ, PopQA (Mallen et al., 2023), and TriviaQA(Joshi et al., 2017). Robustness to retrieval noise is also critical for RAG systems. To evaluate this aspect, we use RAG-Bench, the retrieval-robustness benchmark introduced with RAAT (Fang et al., 2024), which provides different retrieval conditions for the same underlying questions, including gold, relevant, irrelevant, and counterfactual passages. In particular, it allows us to assess whether a model can exploit the gold evidence while remaining robust when noisy passages are presented alongside it. Specifically, we use the cached evaluation split released in the official RAAT repository,[2] which fixes one retrieval context per query and noise type for reproducibility. We evaluate semantic answer correctness using an LLM-as-a-judge protocol with GPT-4o-mini (OpenAI, 2024). Detailed evaluation settings are provided in Appendix C.

Table 2: **Results with an SFT-initialized student.** We report distillation performance from Gemma-3-12B-IT to an SFT-ed Gemma-3-1B-IT student on three open-domain QA benchmarks and RAG-Bench with different types of retrieved passages.

Method		General QA				RAG-Bench				
		NQ	PopQA	TriviaQA	Avg.	Gold	Relevant	Irrelevant	Counter.	Avg.
Teacher		0.474	0.586	0.642	0.567	0.534	0.528	0.556	0.484	0.526
SFT-ed Student		0.368	0.514	0.488	0.457	0.470	0.390	0.482	0.386	0.432
`Gemma3-12B-IT` $\rightarrow$ `SFT-ed Gemma3-1B-IT`										
Standard KD		0.383$_{\pm 0.003}$	0.508$_{\pm 0.005}$	0.502$_{\pm 0.002}$	0.464	0.475$_{\pm 0.005}$	0.426$_{\pm 0.009}$	0.492$_{\pm 0.007}$	0.433$_{\pm 0.008}$	0.457
Parametric-Teacher-KD ($\alpha = 0$)		0.308$_{\pm 0.002}$	0.455$_{\pm 0.002}$	0.461$_{\pm 0.012}$	0.408	0.410$_{\pm 0.007}$	0.371$_{\pm 0.018}$	0.401$_{\pm 0.015}$	0.374$_{\pm 0.013}$	0.389
Vanilla RAG-KD ($\alpha = 1$)		0.385$_{\pm 0.005}$	0.475$_{\pm 0.006}$	0.495$_{\pm 0.006}$	0.452	0.485$_{\pm 0.007}$	0.403$_{\pm 0.012}$	0.498$_{\pm 0.005}$	0.400$_{\pm 0.014}$	0.447
Empirical-Reference Distill ($\widehat{p}^*$)		0.375$_{\pm 0.001}$	0.511$_{\pm 0.002}$	0.482$_{\pm 0.007}$	0.456	0.473$_{\pm 0.008}$	0.389$_{\pm 0.007}$	0.477$_{\pm 0.005}$	0.388$_{\pm 0.004}$	0.432
Mixture RAG-KD (Ours)										
	Static ($\alpha = 0.8$)	0.388$_{\pm 0.007}$	0.503$_{\pm 0.011}$	0.505$_{\pm 0.010}$	0.465	0.496$_{\pm 0.009}$	0.427$_{\pm 0.012}$	0.503$_{\pm 0.006}$	0.412$_{\pm 0.013}$	0.460
Global	FKL	0.399$_{\pm 0.004}$	0.495$_{\pm 0.016}$	0.509$_{\pm 0.006}$	0.468	0.484$_{\pm 0.007}$	0.424$_{\pm 0.010}$	0.494$_{\pm 0.019}$	0.425$_{\pm 0.005}$	0.457
	L_2	0.395$_{\pm 0.005}$	0.496$_{\pm 0.009}$	0.503$_{\pm 0.006}$	0.465	0.482$_{\pm 0.007}$	0.423$_{\pm 0.013}$	0.493$_{\pm 0.013}$	0.421$_{\pm 0.010}$	0.455
Instance	FKL	0.401$_{\pm 0.011}$	0.514$_{\pm 0.005}$	0.518$_{\pm 0.004}$	**0.478**	0.496$_{\pm 0.005}$	0.441$_{\pm 0.002}$	0.509$_{\pm 0.010}$	0.445$_{\pm 0.005}$	**0.473**
	L_2	0.397$_{\pm 0.006}$	0.520$_{\pm 0.007}$	0.514$_{\pm 0.005}$	<u>0.477</u>	0.491$_{\pm 0.006}$	0.440$_{\pm 0.005}$	0.503$_{\pm 0.005}$	0.433$_{\pm 0.008}$	<u>0.467</u>

4.2 EXPERIMENTAL RESULTS

Main Results. Table 1 shows that risk-based Mixture RAG-KD consistently outperforms both endpoint baselines across student scales. All FKL- and L_2-based variants achieve higher aggregate

[1] `https://github.com/google-research/google-research/tree/master/speculative_kd`

[2] `https://github.com/calubkk/RAAT`

Table 1: **Main results across general QA and diverse retrieval conditions.** We report distillation performance from pretrained Gemma-3-12B-IT to Gemma-3-1B-IT and Gemma-3-4B-IT students on three open-domain QA benchmarks and RAG-Bench with different types of retrieved passages.

Method		General QA				RAG-Bench				
		NQ	PopQA	TriviaQA	Avg.	Gold	Relevant	Irrelevant	Counter.	Avg.
Teacher		0.474	0.586	0.642	0.567	0.534	0.528	0.556	0.484	0.526
`Gemma3-12B-IT` → `Gemma3-1B-IT`										
Standard KD		0.336$_{\pm 0.017}$	0.471$_{\pm 0.012}$	0.502$_{\pm 0.008}$	0.436	0.463$_{\pm 0.005}$	0.440$_{\pm 0.009}$	0.480$_{\pm 0.005}$	0.420$_{\pm 0.008}$	0.451
Parametric-Teacher-KD($\alpha = 0$)		0.291$_{\pm 0.008}$	0.464$_{\pm 0.004}$	0.465$_{\pm 0.010}$	0.406	0.421$_{\pm 0.006}$	0.383$_{\pm 0.001}$	0.432$_{\pm 0.017}$	0.412$_{\pm 0.009}$	0.412
Vanilla RAG-KD ($\alpha = 1$)		0.379$_{\pm 0.001}$	0.485$_{\pm 0.002}$	0.500$_{\pm 0.013}$	0.455	0.464$_{\pm 0.009}$	0.401$_{\pm 0.013}$	0.493$_{\pm 0.008}$	0.391$_{\pm 0.002}$	0.437
Empirical-Reference Distill ($\widehat{p}^*$)		0.375$_{\pm 0.002}$	0.505$_{\pm 0.006}$	0.523$_{\pm 0.006}$	0.468	0.469$_{\pm 0.008}$	0.412$_{\pm 0.007}$	0.487$_{\pm 0.024}$	0.415$_{\pm 0.020}$	0.446
Mixture RAG-KD (Ours)										
Global	Static ($\alpha = 0.8$)	0.408$_{\pm 0.008}$	0.490$_{\pm 0.009}$	0.508$_{\pm 0.005}$	0.469	0.502$_{\pm 0.002}$	0.423$_{\pm 0.021}$	0.506$_{\pm 0.014}$	0.422$_{\pm 0.008}$	**0.463**
	FKL	0.402$_{\pm 0.009}$	0.491$_{\pm 0.008}$	0.506$_{\pm 0.009}$	0.466	0.492$_{\pm 0.005}$	0.424$_{\pm 0.005}$	0.495$_{\pm 0.021}$	0.414$_{\pm 0.013}$	0.456
	L_2	0.403$_{\pm 0.007}$	0.492$_{\pm 0.009}$	0.505$_{\pm 0.006}$	0.466	0.496$_{\pm 0.007}$	0.427$_{\pm 0.013}$	0.493$_{\pm 0.009}$	0.416$_{\pm 0.016}$	0.458
Instance	FKL	0.408$_{\pm 0.010}$	0.498$_{\pm 0.021}$	0.520$_{\pm 0.010}$	**0.475**	0.495$_{\pm 0.012}$	0.437$_{\pm 0.015}$	0.499$_{\pm 0.013}$	0.413$_{\pm 0.017}$	0.461
	L_2	0.405$_{\pm 0.004}$	0.503$_{\pm 0.017}$	0.514$_{\pm 0.007}$	0.474	0.489$_{\pm 0.006}$	0.428$_{\pm 0.007}$	0.498$_{\pm 0.013}$	0.421$_{\pm 0.008}$	0.459
`Gemma3-12B-IT` → `Gemma3-4B-IT`										
Standard KD		0.464$_{\pm 0.004}$	0.559$_{\pm 0.015}$	0.623$_{\pm 0.015}$	0.548	0.529$_{\pm 0.005}$	0.495$_{\pm 0.010}$	0.549$_{\pm 0.012}$	0.469$_{\pm 0.012}$	0.511
Parametric-Teacher-KD($\alpha = 0$)		0.355$_{\pm 0.005}$	0.469$_{\pm 0.015}$	0.597$_{\pm 0.013}$	0.473	0.473$_{\pm 0.004}$	0.481$_{\pm 0.007}$	0.490$_{\pm 0.009}$	0.469$_{\pm 0.001}$	0.478
Vanilla RAG-KD ($\alpha = 1$)		0.450$_{\pm 0.009}$	0.566$_{\pm 0.011}$	0.611$_{\pm 0.006}$	0.542	0.539$_{\pm 0.008}$	0.489$_{\pm 0.001}$	0.559$_{\pm 0.013}$	0.453$_{\pm 0.013}$	0.510
Empirical-Reference Distill ($\widehat{p}^*$)		0.477$_{\pm 0.003}$	0.570$_{\pm 0.006}$	0.599$_{\pm 0.007}$	0.549	0.525$_{\pm 0.011}$	0.471$_{\pm 0.006}$	0.544$_{\pm 0.004}$	0.461$_{\pm 0.002}$	0.500
Mixture RAG-KD (Ours)										
Global	Static ($\alpha = 0.8$)	0.479$_{\pm 0.006}$	0.563$_{\pm 0.005}$	0.626$_{\pm 0.002}$	0.556	0.547$_{\pm 0.012}$	0.508$_{\pm 0.004}$	0.559$_{\pm 0.009}$	0.495$_{\pm 0.009}$	**0.527**
	FKL	0.473$_{\pm 0.009}$	0.568$_{\pm 0.007}$	0.629$_{\pm 0.006}$	0.557	0.541$_{\pm 0.001}$	0.510$_{\pm 0.007}$	0.563$_{\pm 0.009}$	0.495$_{\pm 0.012}$	**0.527**
	L_2	0.472$_{\pm 0.004}$	0.564$_{\pm 0.004}$	0.633$_{\pm 0.005}$	0.556	0.543$_{\pm 0.006}$	0.509$_{\pm 0.008}$	0.561$_{\pm 0.003}$	0.494$_{\pm 0.007}$	**0.527**
Instance	FKL	0.473$_{\pm 0.003}$	0.569$_{\pm 0.004}$	0.631$_{\pm 0.009}$	0.558	0.537$_{\pm 0.008}$	0.508$_{\pm 0.002}$	0.558$_{\pm 0.012}$	0.493$_{\pm 0.006}$	0.524
	L_2	0.476$_{\pm 0.013}$	0.578$_{\pm 0.006}$	0.638$_{\pm 0.012}$	**0.564**	0.544$_{\pm 0.002}$	0.512$_{\pm 0.011}$	0.559$_{\pm 0.008}$	0.489$_{\pm 0.014}$	0.526

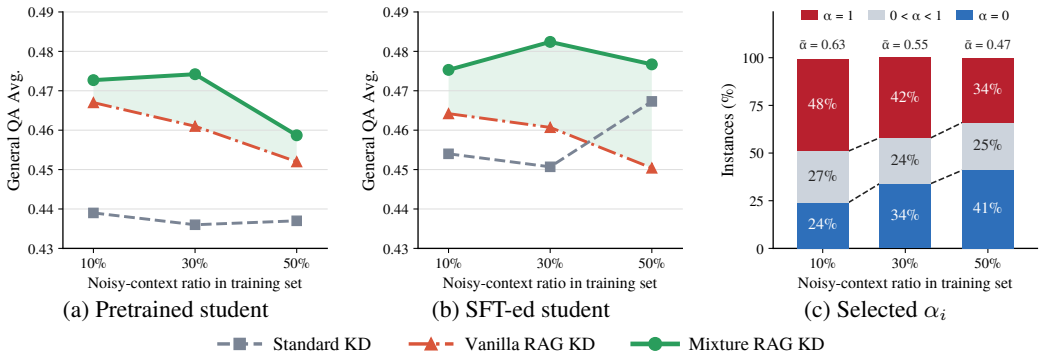


Figure 6: **Robustness to noisy retrieval and instance-wise coefficient behavior.** (a), (b) General QA performance as the fraction of noisy training instances increases from 10% to 50% for pretrained and SFT-ed students. Standard KD is shown as a retrieval-independent reference, and shaded regions indicate the gain of Mixture RAG-KD over Vanilla RAG-KD. (c) Distribution of instance-wise coefficients α_i selected by the forward-KL rule, grouped into $\alpha_i = 0$, $0 < \alpha_i < 1$, and $\alpha_i = 1$; the mean $\bar{\alpha}$ is shown above each bar.

performance on both general QA and RAG-Bench than the Parametric-Teacher KD($\alpha = 0$) and Vanilla RAG-KD ($\alpha = 1$) baselines, suggesting that neither endpoint alone provides uniformly effective supervision under imperfect retrieval. Compared with Vanilla RAG-KD, Mixture RAG-KD improves general QA by up to 0.02 and 0.022 and RAG-Bench by up to 0.026 and 0.017 for the 1B and 4B students, respectively. These consistent gains across both student scales demonstrate that mixture supervision remains effective across different student capacities.

We further observe that risk-based mixture targets can outperform direct distillation from the empirical reference $\widehat{p}^*$. Although $\widehat{p}^*$ is constructed from the observed gold answer, its one-hot form does not retain the richer distributional information contained in soft teacher predictions. This suggests that $\widehat{p}^*$ can serve effectively as a reference for selecting α, while the resulting soft mixture T_α provides a stronger distillation target.

Results with an SFT-initialized student. Table 2 evaluates Mixture RAG-KD on top of a student already trained with retrieval-conditioned SFT on (x, c, y). Despite having already learned from the

gold answers and retrieved contexts, the SFT-ed student further improves from 0.457 to 0.478 on General QA and from 0.432 to 0.473 on RAG-Bench with instance-wise FKL, outperforming both Standard KD (0.464/0.457) and Vanilla RAG-KD (0.452/0.447). These results show that Mixture RAG-KD can be applied orthogonally to SFT, providing additional supervision on top of an already SFT-trained student. While SFT provides hard supervision from the gold answer, Mixture RAG-KD additionally exploits the teachers' soft predictive distributions and controls the relative contribution of parametric and retrieval-conditioned knowledge.

Robustness to noisy retrieval. Figure 6 varies the fraction of training instances with noisy retrieved contexts from 10% to 50%. Vanilla RAG-KD degrades monotonically as noise increases, whereas Mixture RAG-KD consistently outperforms it and remains more stable, particularly for the SFT-ed student. At 50% noise, Mixture RAG-KD still outperforms Standard KD, while Vanilla RAG-KD falls below it, highlighting the vulnerability of directly distilling retrieval-conditioned supervision under noisy retrieval. Figure 6(c) further shows how the instance-wise coefficients adapt to increasing retrieval noise. As the noise ratio increases, the fraction of instances assigned $\alpha_i = 0$ rises from 24% to 41%, while the mean coefficient $\bar{\alpha}$ decreases from 0.63 to 0.47, indicating an increasing reliance on the parametric teacher. Meanwhile, the selected coefficients remain heterogeneous across instances, suggesting that the appropriate balance between parametric and retrieval-conditioned supervision depends on the individual example and motivating instance-wise α selection.

Why instance-wise coefficients. Figure 7 examines the instance-wise coefficients from two sides. On the performance side (top), sweeping a single fixed α over 13 values reveals a clear interior optimum, peaking at $\alpha = 0.85$ and falling toward both endpoints, so neither teacher alone is the best target. Our instance-wise coefficients match or exceed this best fixed value (0.475/0.474 vs. 0.472 for FKL/L_2) without any search over α, and the two selection rules perform alike, indicating that the gain comes from adapting the coefficient per instance rather than from a particular target-risk metric. On the behavioral side (bottom), the selected α_i align with which teacher is actually right, even though correctness is never used in the selection: when only the retrieval-conditioned teacher is correct the mean coefficient is $\bar{\alpha} \approx 0.91$, and when only the parametric teacher is correct it drops to $\bar{\alpha} \approx 0.27$, consistently under both rules.

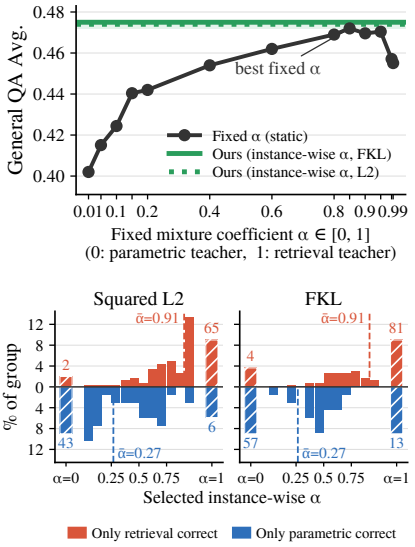


Figure 7: **Instance-wise coefficients.** *Top:* fixed α sweep (black) vs. instance-wise coefficients (green). *Bottom:* selected α_i for correct cases of single teacher, grouped by teacher with mean $\bar{\alpha}$. Full distribution in Appendix D.4.

5 Conclusion

We studied RAG-aware knowledge distillation under imperfect retrieval, where retrieval-conditioned teacher supervision can be beneficial when retrieved evidence is informative but can also propagate retrieval-induced errors when the context is noisy or misleading. Rather than treating the retrieval-conditioned teacher as uniformly preferable, we viewed the teacher's parametric and retrieval-conditioned predictions as complementary sources of supervision and introduced *Mixture RAG-KD* to balance them through a mixing coefficient. We characterized conditions under which an intermediate mixture achieves lower target risk than either endpoint and developed practical global and instance-wise coefficient selection strategies based on forward-KL and squared-L_2 target risks.

Across student scales and retrieval conditions, Mixture RAG-KD improves over directly distilling either endpoint and better maintains performance as retrieval noise increases. Instance-wise mixing further adapts the reliance on retrieval across examples, while forward-KL and squared-L_2 selection provide comparable improvements despite their different optimization procedures. The method also remains effective when applied on top of a retrieval-conditioned SFT student, suggesting that soft teacher supervision provides information complementary to hard gold-answer training. Overall, our results highlight that RAG distillation should explicitly account for the varying reliability of retrieved information, enabling compact students to exploit useful retrieval without indiscriminately inheriting retrieval-conditioned errors.

AI USE STATEMENT

We used generative AI tools in this work for the four purposes below, following the ICLR policy on LLM use. Writing. We used generative AI tools to aid and polish writing, including translation, drafting and editing parts of the manuscript to improve readability, limited suggestions on manuscript organization and presentation, LaTeX formatting, table preparation, and creating and stylistically refining scientific figures. We also used generative AI tools to improve the presentation of mathematical derivations and proofs after they had been fully developed by the authors.
Retrieval and discovery. We used generative AI tools for sourcing information and identifying potentially relevant prior work. All suggested references and citations were manually checked against the original sources by the authors. Research ideation and execution. We used generative AI tools to provide feedback on research methodology and experimental design, to implement and debug methods, to perform minor coding tasks, to clean and reformat datasets, and to support qualitative analysis and interpretation of experimental results, in particular examining and categorizing failure cases. We have not used generative AI tools to propose or refine research hypotheses, to develop the theoretical models or conceptual frameworks, to formulate theorems or other mathematical claims, or to provide critical ingredients for proving them; the research hypotheses, theoretical statements, assumptions, derivations, and proofs were developed and verified by the authors, and all conclusions drawn from the experiments were made by the authors.
Other purposes. We used the GPT API to generate evaluation data in the form of LLM-as-a-judge outputs used to assess model responses, as described in the experimental setup. These judgments were used only for evaluation and never for training the student models.
We have reviewed all AI-assisted work. AI-assisted code was reviewed and tested for correctness; AI-generated failure-case analyses were reviewed against the original model outputs and confirmed by the authors; and all AI-assisted text, figures, tables, references, evaluation outputs, and other artifacts were manually verified by the authors. We take full responsibility for the final content of this work, including all text, claims, experimental results, code, and artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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 *The twelfth international conference on learning representations*, 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*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 719–729, 2024.

Joyeeta Datta, Niclas Doll, Qusai Ramadan, and Zeyd Boukhers. Exploring the limits of model compression in LLMs: A knowledge distillation study on QA tasks. In Frédéric Béchet, Fabrice Lefèvre, Nicholas Asher, Seokhwan Kim, and Teva Merlin (eds.), *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 477–483, Avignon, France, August 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.sigdial-1.39. URL https://aclanthology.org/2025.sigdial-1.39/.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10028–10039, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.540. URL https://aclanthology.org/2024.acl-long.540/.

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. URL https://openreview.net/forum?id=5h0qf7IBZZ.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Kiseung Kim and Jay-Yoon Lee. RE-RAG: Improving open-domain QA performance and interpretability with relevance estimator in retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22149–22161, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1236. URL https://aclanthology.org/2024.emnlp-main.1236/.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=lsHZNNoC7r.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 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. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. URL https://aclanthology.org/2023.acl-long.546.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. Accessed: 2025-01-05.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International conference on machine learning*, pp. 31210–31227. PMLR, 2023.

Donghyeok Shin, Yeongmin Kim, Suhyeon Jo, Byeonghu Na, and Il chul Moon. AMid: Knowledge distillation for LLMs with α-mixture assistant distribution. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7WPJ0EgPdW.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cqsw28DuMW.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5613–5626, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.321. URL https://aclanthology.org/2024.emnlp-main.321/.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. Infogain-rag: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7201–7215, 2025.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P1qhkp8gQT.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=mlJLVigNHp.

Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088*, 2024b.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation, 2024. URL https://openreview.net/forum?id=JnWJbrnaUE.

Shi-Qi Yan, Quan Liu, and Zhen-Hua Ling. RPO: Retrieval preference optimization for robust retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5228–5240, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.261. URL https://aclanthology.org/2025.acl-long.261/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, volume 2024, pp. 29862–29883, 2024.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.813. URL https://aclanthology.org/2024.emnlp-main.813/.

Yue Zheng, Yuhao Chen, Bin Qian, Xiufang Shi, Yuanchao Shu, and Jiming Chen. A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8):1–35, 2025.

A PROOFS AND DERIVATIONS

A.1 PROOFS OF THEOREM 3.1

Theorem 3.1. *Let p^* denote the optimal target distribution, and let $Q := \{q_\theta \mid \theta \in \Theta\}$ be a set of student models. We define the token-level risk functional between two distributions u and v as*

$$\mathcal{R}(u, v) \coloneqq \mathbb{E}_{(x,y,c)\sim\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} D\left(u(\cdot \mid x, c, y_{<l}) \,\|\, v(\cdot \mid x, c, y_{<l})\right) \right]. \tag{4}$$

For each $\alpha \in [0, 1]$, we define the optimal student model distilled from the mixture target T_α as

$$q_\alpha^* \in \arg\min_{q_\theta \in Q} \mathcal{R}(T_\alpha, q_\theta). \tag{5}$$

In particular, q_0^* and q_1^* are the optimal student models corresponding to the endpoint targets $T_0 = p$ and $T_1 = p_c$, respectively.
We define $\bar{\mathcal{R}}(\alpha) \coloneqq \mathcal{R}(p^*, T_\alpha)$. Suppose that (1) Q is expressive enough to include T_α for every $\alpha \in [0, 1]$; and (2) $\bar{\mathcal{R}}(\alpha)$ is finite and continuous on $[0, 1]$, and $\bar{\mathcal{R}}'_+(0)$ and $\bar{\mathcal{R}}'_-(1)$ exist.
If $\bar{\mathcal{R}}'_+(0) < 0$ and $\bar{\mathcal{R}}'_-(1) > 0$, then every $\alpha^* \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}(\alpha)$ satisfies $\alpha^* \in (0, 1)$, and

$$\mathcal{R}(p^*, q_{\alpha^*}^*) < \min\left\{\mathcal{R}(p^*, q_0^*), \mathcal{R}(p^*, q_1^*)\right\}. \tag{6}$$

Proof. We first focus on the strict improvement of the mixed target supervision. Since $\bar{\mathcal{R}}(\alpha)$ is finite and continuous on the compact interval $[0, 1]$, by the extreme value theorem, there exists $\alpha^* \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}(\alpha)$ on $[0, 1]$. Since $\bar{\mathcal{R}}'_+(0) < 0$, by the definition of the right derivative, there exists sufficiently small $\varepsilon_0 > 0$ such that:

$$\bar{\mathcal{R}}(\varepsilon_0) < \bar{\mathcal{R}}(0) \tag{12}$$

Hence, $\alpha = 0$ cannot be a global minimizer. Similarly, since $\bar{\mathcal{R}}'_-(1) > 0$, there exists sufficiently small $\varepsilon_1 > 0$ such that:

$$\bar{\mathcal{R}}(1 - \varepsilon_1) < \bar{\mathcal{R}}(1) \tag{13}$$

Hence, $\alpha = 1$ also cannot be a global minimizer. Therefore, every $\alpha^* \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}(\alpha)$ satisfies $\alpha^* \in (0, 1)$. Furthermore, note that $\bar{\mathcal{R}}(\alpha^*) \le \bar{\mathcal{R}}(\varepsilon_0) < \bar{\mathcal{R}}(0)$ and $\bar{\mathcal{R}}(\alpha^*) \le \bar{\mathcal{R}}(1 - \varepsilon_1) < \bar{\mathcal{R}}(1)$ by the optimality of α^*. Therefore, we have:

$$\bar{\mathcal{R}}(\alpha^*) < \min\{\bar{\mathcal{R}}(0), \bar{\mathcal{R}}(1)\}. \tag{14}$$

Now, we consider the distillation part. Since D is a proper statistical divergence and assumption (1), we have $q_\alpha^* = T_\alpha$ for any $\alpha \in [0, 1]$. Therefore,

$$\mathcal{R}(p^*, q_\alpha^*) = \mathcal{R}(p^*, T_\alpha) = \bar{\mathcal{R}}(\alpha) \quad \text{for every } \alpha \in [0, 1]. \tag{15}$$

Applying Eq. (15) with $\alpha \in \{0, \alpha^*, 1\}$ to Eq. (14), we obtain the final result:

$$\mathcal{R}(p^*, q_{\alpha^*}^*) = \bar{\mathcal{R}}(\alpha^*) < \min\{\bar{\mathcal{R}}(0), \bar{\mathcal{R}}(1)\} = \min\big\{\mathcal{R}(p^*, q_0^*),\, \mathcal{R}(p^*, q_1^*)\big\}. \tag{16}$$

□

A.2 FORWARD-KL ENDPOINT CONDITIONS

We derive the endpoint conditions for the forward-KL target risk. For each answer token y_l, we use the empirical reference $\widehat{p}^* = \delta_{y_l}$ and define the mixture target

$$T_\alpha = (1 - \alpha)p + \alpha p_c, \qquad \alpha \in [0, 1]. \tag{17}$$

Since the reference distribution is one-hot, the forward-KL target risk reduces to

$$\bar{\mathcal{R}}_{\mathrm{FKL}}(\alpha) = \mathbb{E}_{\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} D_{\mathrm{KL}}\left(\delta_{y_l} \,\|\, T_\alpha\right) \right] \tag{18}$$

$$= -\mathbb{E}_{\mathcal{D}} \left[\frac{1}{L} \sum_{l=1}^{L} \log\left((1 - \alpha)p(y_l) + \alpha p_c(y_l)\right) \right]. \tag{19}$$

Differentiating with respect to α gives

$$\bar{\mathcal{R}}'_{\mathrm{FKL}}(\alpha) = -\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\frac{p_c(y_l) - p(y_l)}{(1-\alpha)p(y_l) + \alpha p_c(y_l)}\right]. \tag{20}$$

At the parametric endpoint $\alpha = 0$,

$$\bar{\mathcal{R}}'_{\mathrm{FKL},+}(0) = -\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\left(\frac{p_c(y_l)}{p(y_l)} - 1\right)\right]. \tag{21}$$

Therefore,

$$\bar{\mathcal{R}}'_{\mathrm{FKL},+}(0) < 0 \iff \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\left(\frac{p_c(y_l)}{p(y_l)} - 1\right)\right] > 0 \tag{22}$$

$$\iff \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\frac{p_c(y_l)}{p(y_l)}\right] > 1. \tag{23}$$

Similarly, at the retrieval-conditioned endpoint $\alpha = 1$,

$$\bar{\mathcal{R}}'_{\mathrm{FKL},-}(1) = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\left(\frac{p(y_l)}{p_c(y_l)} - 1\right)\right], \tag{24}$$

and hence

$$\bar{\mathcal{R}}'_{\mathrm{FKL},-}(1) > 0 \iff \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\left(\frac{p(y_l)}{p_c(y_l)} - 1\right)\right] > 0 \tag{25}$$

$$\iff \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\frac{p(y_l)}{p_c(y_l)}\right] > 1. \tag{26}$$

The curvature of the forward-KL risk is

$$\bar{\mathcal{R}}''_{\mathrm{FKL}}(\alpha) = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\frac{(p_c(y_l) - p(y_l))^2}{((1-\alpha)p(y_l) + \alpha p_c(y_l))^2}\right] \geq 0, \tag{27}$$

so $\bar{\mathcal{R}}_{\mathrm{FKL}}(\alpha)$ is convex along the mixture path. Consequently, when $\bar{\mathcal{R}}'_{\mathrm{FKL},+}(0) < 0$ and $\bar{\mathcal{R}}'_{\mathrm{FKL},-}(1) > 0$, its minimum is attained at an interior mixture $\alpha^* \in (0,1)$, corresponding to the regime characterized in Theorem 3.1.

B Mixing Strategies Details

In this section, we describe how the global and instance-wise mixing coefficients are selected by minimizing the corresponding target risk. For coefficient selection, we use the empirical reference $\widehat{p}^* = \delta_{y_l}$, i.e., a one-hot distribution on the gold answer token at each answer position l. We compute the target risk over the gold answer span, excluding the trailing newline token automatically appended by the Gemma-3 chat template. This exclusion is applied consistently to all coefficient-selection strategies.

B.1 Forward-KL Coefficient Solver

Global and instance-wise objectives. For global mixing, we select a single coefficient shared across the entire training set by minimizing the empirical forward-KL target risk:

$$\alpha_{\mathrm{FKL}}^{\mathrm{global}} \in \operatorname*{arg\,min}_{\alpha\in[0,1]} \bar{\mathcal{R}}_{\mathrm{FKL}}(\alpha), \tag{28}$$

where

$$\bar{\mathcal{R}}_{\mathrm{FKL}}(\alpha) = \mathbb{E}_{(x,y,c)\sim\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L} D_{\mathrm{KL}}\left(\widehat{p}^* \,\|\, T_\alpha\right)\right]. \tag{29}$$

For instance-wise mixing, we instead optimize a separate coefficient for each training instance i:

$$\alpha^{\text{inst}}_{\text{FKL},i} \in \arg\min_{\alpha\in[0,1]} \mathcal{R}^{(i)}_{\text{FKL}}(\alpha), \tag{30}$$

where

$$\mathcal{R}^{(i)}_{\text{FKL}}(\alpha) = \frac{1}{L_i}\sum_{l=1}^{L_i} D_{\text{KL}}\left(\widehat{p}^{*} \,\|\, T_\alpha\right). \tag{31}$$

A single α_i is therefore shared across the entire answer span of instance i, rather than estimated separately for each token.

Bisection solver. As established in Appendix A.2, $\bar{\mathcal{R}}_{\text{FKL}}(\alpha)$ is convex in α. Let

$$g(\alpha) := \frac{d}{d\alpha}\bar{\mathcal{R}}_{\text{FKL}}(\alpha). \tag{32}$$

Convexity implies that $g(\alpha)$ is monotonically nondecreasing, so the sign of the derivative determines the location of an optimum. In particular,

$$\alpha^{\text{global}}_{\text{FKL}} = \begin{cases} 0, & g_+(0) \geq 0, \\ 1, & g_-(1) \leq 0, \\ \alpha \in (0,1) \text{ s.t. } g(\alpha) = 0, & g_+(0) < 0 < g_-(1). \end{cases} \tag{33}$$

When the optimum is interior, we initialize a bracketing interval

$$[\ell_0, h_0] = [0, 1], \tag{34}$$

where ℓ_k and h_k denote the lower and upper bounds after iteration k, respectively. The endpoint condition guarantees

$$g(\ell_0) < 0 < g(h_0). \tag{35}$$

At iteration $k = 1, \ldots, K$, we evaluate the derivative at the midpoint

$$m_k = \frac{\ell_{k-1} + h_{k-1}}{2}. \tag{36}$$

Since g is nondecreasing, we update the bracket according to

$$(\ell_k, h_k) = \begin{cases} (m_k, h_{k-1}), & g(m_k) < 0, \\ (\ell_{k-1}, m_k), & g(m_k) > 0. \end{cases} \tag{37}$$

If $g(m_k) = 0$, then m_k is an interior optimum and the procedure terminates.
Otherwise, the update preserves

$$g(\ell_k) < 0 < g(h_k), \qquad \alpha^{\text{global}}_{\text{FKL}} \in [\ell_k, h_k], \tag{38}$$

while halving the bracket width at every iteration:

$$h_k - \ell_k = 2^{-k}. \tag{39}$$

After K iterations, we return

$$\widehat{\alpha}_K = \frac{\ell_K + h_K}{2}, \tag{40}$$

which satisfies

$$\left|\widehat{\alpha}_K - \alpha^{\text{global}}_{\text{FKL}}\right| \leq \frac{h_K - \ell_K}{2} = 2^{-(K+1)}. \tag{41}$$

For instance-wise mixing, we analogously define

$$g_i(\alpha) = -\frac{1}{L_i}\sum_{l=1}^{L_i} \frac{p_c(y_{i,l}) - p(y_{i,l})}{(1-\alpha)p(y_{i,l}) + \alpha p_c(y_{i,l})}, \tag{42}$$

and apply the same endpoint tests and bisection updates independently to each training instance. We use $K = 10$ for instance-wise coefficients.

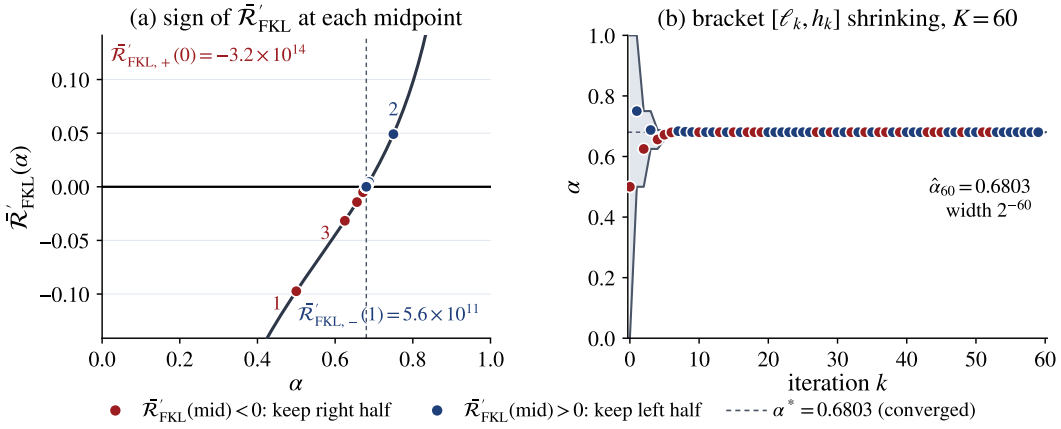


Figure 8: **Global forward-KL coefficient solved by bisection on NQ-RAG-1K** ($K = 60$)**.** (a) Derivative $g(\alpha) = \bar{\mathcal{R}}'_{\mathrm{FKL}}(\alpha)$ evaluated at the midpoint of each iteration. A negative derivative updates the lower bound to the midpoint, whereas a positive derivative updates the upper bound. (b) Evolution of the bracket $[\ell_k, h_k]$, whose width is halved at each iteration and converges to $\alpha^{\mathrm{global}}_{\mathrm{FKL}} = 0.68026$.

k	$[\ell_{k-1}, h_{k-1}]$	m_k	$g(m_k)$	Retained half
1	$[0, 1]$	0.50000	-9.7×10^{-2}	right
2	$[0.5, 1]$	0.75000	$+4.9 \times 10^{-2}$	left
3	$[0.5, 0.75]$	0.62500	-3.2×10^{-2}	right
4	$[0.625, 0.75]$	0.68750	$+4.5 \times 10^{-3}$	left
5	$[0.625, 0.6875]$	0.65625	-1.4×10^{-2}	right
6	$[0.65625, 0.6875]$	0.67188	-5.1×10^{-3}	right
7	$[0.67188, 0.6875]$	0.67969	-3.5×10^{-4}	right
8	$[0.67969, 0.6875]$	0.68359	$+2.1 \times 10^{-3}$	left
9	$[0.67969, 0.68359]$	0.68164	$+8.5 \times 10^{-4}$	left
10	$[0.67969, 0.68164]$	0.68066	$+2.5 \times 10^{-4}$	left
$\vdots$				
$\alpha^{\mathrm{global}}_{\mathrm{FKL}} = 0.68026$ after 60 iterations				

Table 3: First ten iterations of the global forward-KL bisection shown in Figure 8.

Example of the global solve. Figure 8 and Table 3 illustrate the global solve on NQ-RAG-1K, containing 1,000 training instances and 10,969 supervised tokens. The endpoint derivatives are $g_+(0) = -3.2 \times 10^{14}$ and $g_-(1) = 5.6 \times 10^{11}$, satisfying the condition for an interior solution. The bracket rapidly contracts around the root, and using $K = 60$ iterations yields

$$\alpha^{\mathrm{global}}_{\mathrm{FKL}} = 0.68026. \tag{43}$$

Sixty iterations are more than sufficient to reach numerical precision. The global solve is performed only once before student training using cached teacher statistics, and therefore its iteration depth does not contribute to the per-step training cost.

B.2 Squared-L_2 Risk-Optimal Mixing

For notational simplicity, we suppress the answer-position index l; all distributions inside the sum over l are understood to be conditioned on the corresponding prefix $y_{<l}$.

Global coefficient.

$$\alpha^{\mathrm{global}}_{L_2} \in \arg\min_{\alpha\in[0,1]} \bar{\mathcal{R}}_{L_2}(\alpha), \tag{44}$$

where

$$\bar{\mathcal{R}}_{L_2}(\alpha) = \mathbb{E}_{(x,y,c)\sim\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\|\widehat{p}^{*} - T_\alpha\|_2^2\right]. \tag{45}$$

Using

$$T_\alpha = p + \alpha(p_c - p), \tag{46}$$

we can write

$$\widehat{p}^{*} - T_\alpha = (\widehat{p}^{*} - p) - \alpha(p_c - p). \tag{47}$$

Substituting this into Eq. (45) and expanding the squared norm gives

$$\bar{\mathcal{R}}_{L_2}(\alpha) = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\Big(\|\widehat{p}^{*} - p\|_2^2 - 2\alpha\,\langle p_c - p,\, \widehat{p}^{*} - p\rangle + \alpha^2\|p_c - p\|_2^2\Big)\right]. \tag{48}$$

Differentiating with respect to α yields

$$\frac{d}{d\alpha}\bar{\mathcal{R}}_{L_2}(\alpha) = -2\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\langle p_c - p,\, \widehat{p}^{*} - p\rangle\right] + 2\alpha\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\|p_c - p\|_2^2\right]. \tag{49}$$

Therefore, when the denominator is nonzero, the unconstrained minimizer is

$$\widetilde{\alpha}_{L_2}^{\text{global}} = \frac{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\langle p_c - p,\, \widehat{p}^{*} - p\rangle\right]}{\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\|p_c - p\|_2^2\right]}. \tag{50}$$

Imposing the constraint $\alpha \in [0, 1]$ gives

$$\boxed{\alpha_{L_2}^{\text{global}} = \text{clip}\left(\widetilde{\alpha}_{L_2}^{\text{global}}, 0, 1\right)}. \tag{51}$$

The second derivative is

$$\frac{d^2}{d\alpha^2}\bar{\mathcal{R}}_{L_2}(\alpha) = 2\mathbb{E}_{\mathcal{D}}\left[\frac{1}{L}\sum_{l=1}^{L}\|p_c - p\|_2^2\right] \geq 0, \tag{52}$$

confirming that the objective is convex in α and that Eq. (51) gives a global minimizer. If the denominator in Eq. (50) is zero, then $p_c = p$ over all supervised positions contributing to the risk, so T_α is independent of α and every $\alpha \in [0, 1]$ is optimal.

Instance-wise coefficient. For instance-wise mixing, we minimize the same squared-L_2 risk separately for each training instance i:

$$\alpha_{L_2,i}^{\text{inst}} \in \arg\min_{\alpha\in[0,1]} \mathcal{R}_{L_2}^{(i)}(\alpha), \tag{53}$$

where

$$\mathcal{R}_{L_2}^{(i)}(\alpha) = \frac{1}{L_i}\sum_{l=1}^{L_i}\|\widehat{p}^{*} - T_\alpha\|_2^2. \tag{54}$$

Applying the same quadratic minimization gives

$$\boxed{\alpha_{L_2,i}^{\text{inst}} = \text{clip}\left(\frac{\frac{1}{L_i}\sum_{l=1}^{L_i}\langle p_c - p,\, \widehat{p}^{*} - p\rangle}{\frac{1}{L_i}\sum_{l=1}^{L_i}\|p_c - p\|_2^2}, 0, 1\right)}. \tag{55}$$

The factor $1/L_i$ appears in both the numerator and denominator and may equivalently be canceled. As in the FKL formulation, a single α_i is estimated for each training instance and shared across its entire answer span.

Geometric interpretation. Before clipping, the numerator measures the alignment between the retrieval-induced direction $p_c - p$ and the direction from the parametric teacher toward the empirical reference $\widehat{p}^* - p$, while the denominator measures the squared magnitude of the retrieval-induced shift. Thus, the unconstrained coefficient is the least-squares projection coefficient that determines how far to move from p toward p_c in order to approach $\widehat{p}^*$ under squared-L_2 risk. Unlike the forward-KL coefficient, the squared-L_2 coefficient is available in closed form and requires no iterative solver.

C Experiments Details

Noise-Level Training Dataset. To control the level of retrieval noise in the training data, we identify instances from the original NQ training set for which the parametric teacher p produces the correct answer, whereas the retrieval-conditioned teacher p_c produces an incorrect answer given the retrieved context c. We treat these instances as retrieval-induced failure cases and construct 1K training sets in which they account for 10%, 30%, and 50% of the examples, respectively.
This construction does not artificially inject contradictory passages or irrelevant sentences. Instead, it captures naturally occurring cases in which retrieved evidence changes an otherwise correct parametric teacher prediction into an incorrect retrieval-conditioned prediction. By varying the proportion of such retrieval-induced failure cases, we obtain training sets with controlled levels of retrieval noise and evaluate how the distillation methods behave as training-time retrieval noise increases.

SFT-ed Student Training Details. To construct the SFT-initialized student, we fully fine-tune Gemma-3-1B-IT for one epoch on the same 1K NQ RAG training set used for distillation, using the gold answers as supervision. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01, and gradient clipping at a global ℓ_2 norm of 1.0. The learning rate is selected from $2 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}$ using a held-out validation set of 500 examples, with 2×10^{-5} selected as the best setting. We use a linear learning-rate decay to zero without warmup, an effective batch size of 8, and random seed 2025. With 1,000 training examples, this corresponds to 125 optimizer steps. Training is performed in bfloat16 with FlashAttention-2; input prompts are truncated to 1,280 tokens and the maximum prompt-response sequence length is 2,560 tokens.

Distillation Details. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay of 0.01, together with linear learning-rate decay to zero without warm-up and gradient clipping at a global norm of 1.0. Gradients are accumulated over 8 training examples, yielding an effective batch size of 8 and 125 optimizer steps per epoch on the 1,000 training pairs.Both the student temperature and teacher temperature are set to 1.0. Training is performed in bf16 with FlashAttention-2, while divergence computation and α estimation are performed in fp32 (fp64 for gold-column bisection). All experiments are run on a single H100 GPU.
For each distillation method, we train with seed 2025 using $\eta \in 2 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}$ and select the learning rate that achieves the lowest validation loss on a held-out set of 500 examples under the corresponding training objective. We then rerun the selected setting with seeds 2026 and 2027 and report the average over the three seeds. For the **pretrained student**, distillation starts from the released instruction-tuned checkpoint and runs for 3 epochs (375 optimizer steps). For the **SFT-initialized student**, distillation starts from a Gemma-3-1B-IT checkpoint that has been SFT-ed for one epoch on the same 1,000 training examples used for distillation. We then train the student with each distillation objective using the same optimization and model-selection procedure as in the pretrained-student setting. All optimization, batch, and sequence-length settings are kept identical to those used for the pretrained student. We additionally train the SFT-initialized student for two distillation epochs and report the corresponding performance in the main text, while results after three epochs are provided in the Appendix D.3.

Evaluation Datasets. To assess both in-domain performance and out-of-distribution (OOD) generalization, we evaluate the models on three standard Open-Domain QA benchmarks:

- Natural Questions (NQ) (Kwiatkowski et al., 2019): serves as the in-domain dataset, following the same distribution as the training set;
- PopQA (Mallen et al., 2023): serves as an OOD dataset focusing on long-tail knowledge entities;
- TriviaQA (Joshi et al., 2017): serves as an OOD dataset requiring complex compositional reasoning.

To evaluate robustness against specific types of retrieval noise, we use RAG-Bench, released with RAAT (Fang et al., 2024). The benchmark is built on top of open-domain QA questions and annotates the retrieved passages for each question into one golden category and three noise categories. Following the authors' description, the categorization is manual and defined as follows:

- **Golden**: a passage that is related to the question and contains the answer entity (verified by regular-expression matching).
- **Relevant Noise**: a passage that is highly related to the question but does *not* contain the answer entity.
- **Irrelevant Noise**: a passage that is unrelated to the question, taken directly from the retrieval results of a different question.
- **Counterfactual Noise**: a copy of a golden passage in which the answer entity has been replaced with a counterfactual entity generated by ChatGPT from the original answer.

Since RAG-Bench provides multiple candidate passages per noise type, we adopt the cached evaluation split released in the official repository (`benchmark_cache/{o,p,c,f}.json`),[3] which fixes a single passage per question and noise type to remove selection randomness and to keep results comparable to those reported in RAAT. From this cached split (3,000 questions), we randomly sample 500 questions and evaluate each under four retrieval conditions: (1) **Golden**: the golden passage only (`o`); (2) **Golden + Relevant Noise** (`p`); (3) **Golden + Irrelevant Noise** (`c`); and (4) **Golden + Counterfactual Noise** (`f`). In the three noisy conditions, the golden passage and the noise passage are concatenated in a 1:1 ratio, so that the model always has access to the correct evidence and the question is whether it can ignore or resist the accompanying noise. The same 500 questions are used across all four conditions, making the per-condition scores directly comparable.

Evaluation Metric. We adopt the LLM-as-a-judge paradigm for evaluation. Specifically, we use GPT-4o-mini (OpenAI, 2024) to assess whether the model-generated response matches the gold reference answer in terms of semantic accuracy. To ensure statistical reliability, all reported results are averaged over three independent evaluation runs.

LLM-as-a-Judge Evaluation Prompt Exact-match metrics penalize correct answers that are phrased differently from the gold string (e.g., aliases or full-sentence answers), and reward answers that merely contain the gold string while stating something wrong. We therefore score every test prediction with an LLM judge. We use `gpt-4o-mini` (OpenAI, 2024) with temperature 0 and JSON-mode output. The judge receives the question (including the retrieved passages for RAG tasks), the model's final predicted answer, and the gold answer, and returns a binary score with a short justification. Reported accuracy is the mean score over the test set. The full prompt is shown in Table 4, and two scored examples in Table 5.

Table 4: Prompt used for the LLM judge.

Compare the following predicted answer with the gold (correct) answer and determine if they match in meaning given the question. Be reasonable - phrasing differences are okay if the core meaning is the same. You should accept the aliases and the answer conveying the same conclusion. If the predicted answer is overly verbose and fails to capture the key information found in the gold answer (i.e., low recall), consider it a false answer. Question: {question} Predicted answer: {predicted_answer} Gold answer: {gold_answer}
Output your evaluation as a JSON object with two fields: - explanation: Brief explanation of your scoring decision - score: 0 or 1 indicating if answers match

Dataset Instruction The instruction consists of three components: a header specifying the answer-format rules and worked examples, a body containing the retrieved context and question, and a final one-line format reminder. We use this strict instruction because, under a simple "answer in `<answer>` tags" prompt, the teacher model (Gemma-3-12B-it) frequently generates explanatory

[3] `https://github.com/calubkk/RAAT`

Table 5: Examples of LLM-judge decisions on the NQ test set (retrieved passages omitted). In the second example the prediction contains the gold string, but the judge rejects it because the answer as a whole is wrong.

Question: who is the owner of the crowne plaza **Predicted answer:** The owner of the Crowne Plaza is the InterContinental Hotels Group. **Gold answer:** InterContinental Hotels Group **Judge:** *"The predicted answer correctly identifies the owner of Crowne Plaza as the InterContinental Hotels Group, which matches the gold answer. The phrasing is slightly different, but the core meaning is the same, thus it is an acceptable response."* **Score: 1**
Question: who is eliminated in big boss kannada 5 **Predicted answer:** In Bigg Boss Kannada 5, Chandan Shetty is eliminated as the winner, and Sales representative Diwaker is voted the runner-up. **Gold answer:** Diwakar **Judge:** *"The predicted answer incorrectly states that Chandan Shetty is eliminated, when in fact he is the winner. The gold answer simply states 'Diwakar' as the eliminated contestant, which is accurate. Therefore, the predicted answer does not match the gold answer in meaning, as it conveys incorrect information about the elimination."* **Score: 0**

text before the final answer. As a result, the first generated token may not align with the first token of the gold target. On 200 training questions, the teacher's greedy output exactly matched the required gold format in 0% of cases under the simple instruction, compared with 96% under our instruction. Its exact-match answer accuracy also increased from 0.335 to 0.435. The closed-book distribution p uses the same format instruction but removes the retrieved context.

Table 6: Prompt with retrieved context (p_c), used for the student and teacher at training and test time.

Prompt with Retrieved Context

Instruction. Read the following context carefully and answer the question based on it. Reply with the final answer ONLY, wrapped inside `<answer>` and `</answer>` tags.

Output format. Exactly one line, with nothing before or after it:

```
<answer> your answer here </answer>
```

Rules. Keep the single space right after `<answer>` and right before `</answer>`, even when the answer is one word or a number. Do not explain. Do not restate the question. Do not add any other text. The answer must be a short span (a name, date, number, or short phrase).

Correct examples.

```
<answer> Charles Darwin </answer>
<answer> 1958 </answer>
<answer> New York City </answer>
```

Incorrect examples.

```
Okay, based on the context the answer is Charles Darwin.
<answer>Charles Darwin</answer>
<answer>1958</answer>
```

Context.

```
[1] {title_1}
{passage_1}
...
[5] {title_5}
{passage_5}
```

Question.

```
{question}
```

Reminder. Now reply with one line only, in exactly this format:

```
<answer> your answer here </answer>
```

Table 7: Closed-book prompt (p), using the same instruction without retrieved context.

Closed-Book Prompt

Instruction. Answer the given question. Reply with the final answer ONLY, wrapped inside `<answer>` and `</answer>` tags.

Output format.

```
<answer> your answer here </answer>
```

Rules. Keep the single space right after `<answer>` and right before `</answer>`, even when the answer is one word or a number. Do not explain. Do not restate the question. Do not add any other text. The answer must be a short span (a name, date, number, or short phrase).

Correct examples.

```
<answer> Charles Darwin </answer>
<answer> 1958 </answer>
<answer> New York City </answer>
```

Question.

```
{question}
```

Reminder. Now reply with one line only, in exactly this format:

```
<answer> your answer here </answer>
```

Table 8: Example from the NQ evaluation set. Passages are truncated for readability.

Natural Questions Example

Context.

[1] Parable of the Prodigal Son
The son could not wait for his father's death for his inheritance; he wanted it immediately. . . .

. . .

[5] Parable of the Prodigal Son
The Parable of the Prodigal Son . . . is one of the parables of Jesus. Jesus Christ shares it with his disciples, the Pharisees, and others. . . .

Question. who told the story of the prodigal son

Target. `<answer> Jesus Christ </answer>`

Method	Variant	Teacher forwards		ms / instance	Time (3 ep.)
		p	p_c		
Standard KD	–	✓		234	11.7 min
Vanilla RAG-KD	–		✓	246	12.3 min
Mixture RAG-KD	Instance-FKL	✓	✓	330	16.5 min
	Instance-L_2	✓	✓	329	16.5 min

Table 9: **Training cost of the distillation methods.** Measurements use Gemma-3-12B-IT as the teacher and Gemma-3-1B-IT as the student on NQ-RAG-1K. Three epochs correspond to 3,000 training-instance passes.

Divergence	Scope	Solver	Runs	Time per solve	Total (3 epochs)
FKL	Instance	Bisection (K=10)	every instance, GPU	0.88 ms	2.6 s
	Global	Bisection (K=60)	once, CPU	15.4 ms	15.4 ms
L2	Instance	Closed form	every instance, GPU	0.23 ms	0.7 s
	Global	Closed form	once, CPU	0.14 ms	0.14 ms

Table 10: Wall-clock cost of solving α (NQ-RAG-1k, 1,000 training instances, 10,969 supervised tokens). Instance-level times are measured on one H100 NVL inside the training step and exclude building the mixture target, which every mixture method pays (0.39 ms). The FKL bisection terminates immediately when the optimum lies on the boundary (0.60 ms; 77% of instances) and costs 1.83 ms otherwise. Global solves run once before training on a single CPU core, from teacher statistics collected in one pass over the training set. Over a 3-epoch run (3,000 instance passes) every variant spends under 3 s solving α, below 0.3% of training time.

D Additional experimental results

D.1 Training Cost

Table 9 reports the training cost of each distillation method for Gemma-3-12B-IT → Gemma-3-1B-IT on NQ-RAG-1K. All measurements are obtained on a single H100 NVL GPU.

Standard KD and Vanilla RAG-KD require a single teacher forward pass per training instance, whereas Mixture RAG-KD evaluates both the parametric teacher p and the retrieval-conditioned teacher p_c. Mixture RAG-KD incurs approximately 34% more training time than Vanilla RAG-KD in our implementation, primarily because it evaluates both the parametric and retrieval-conditioned teacher distributions for each training example. Importantly, this overhead is confined to offline distillation and does not affect inference cost. Moreover, since the teacher is frozen and the training set is fixed, both teacher distributions can be precomputed and cached once, allowing their cost to be amortized across epochs, random seeds, or subsequent distillation runs. The additional computation required for coefficient selection itself is negligible: solving the instance-wise FKL coefficient with 10 bisection iterations adds only 0.9 ms per instance ($< 0.3\%$ of the total training time), while the L_2 coefficient is obtained in closed form.

D.2 Bisection cost

Table 10 shows that the computational cost of solving for α is negligible relative to the overall distillation cost. For instance-wise FKL, solving α_i with 10 bisection iterations requires only 0.88ms per instance on average and less than 3s over the full 3-epoch training run. The cost is further reduced because the endpoint test terminates immediately for instances whose optimum lies at $\alpha_i = 0$ or $\alpha_i = 1$. Instance-wise L_2 is even cheaper, as its coefficient is obtained in closed form.

Global coefficients are computed only once before student training. This requires a preliminary pass over the training set to obtain and cache the teacher outputs for both $p(y \mid x)$ and $p(y \mid x, c)$, from which the dataset-level risk and its derivative can be evaluated. Given these cached teacher statistics, the global FKL bisection takes only 15.4,ms on a single CPU core, while the global L_2 coefficient is obtained in 0.14,ms. Therefore, the coefficient solver itself contributes less than 0.3% of the overall training time in all settings; the dominant additional cost of Mixture RAG-KD remains

the computation of the two teacher distributions, which can likewise be amortized when their outputs are cached and reused.

D.3 SFTED STUDENT RESULT

Table 11 reports results when the student is initialized from a retrieval-conditioned SFT checkpoint trained for three epochs. Consistent with the epoch 2 results in Table 2, Mixture RAG-KD continues to provide gains over the SFT baseline and standard distillation approaches, showing that its benefit persists even with a more extensively SFT-trained student.

Table 11: **Results with an SFT-initialized student.**

Method		General QA				RAG-Bench				
		NQ	PopQA	TriviaQA	Avg.	Gold	Relevant	Irrelevant	Counter.	Avg.
Teacher		0.474	0.586	0.642	0.567	0.534	0.528	0.556	0.484	0.526
SFT-ed Student		0.368	0.514	0.488	0.457	0.470	0.390	0.482	0.386	0.432
`Gemma3-12B-IT` → `SFT-ed Gemma3-1B-IT`										
Standard KD		0.395±0.004	0.507±0.003	0.505±0.006	0.469	0.465±0.005	0.433±0.005	0.487±0.005	0.434±0.009	0.455
Parametric-Teacher-KD ($\alpha = 0$)		0.290±0.005	0.450±0.005	0.452±0.006	0.397	0.414±0.004	0.384±0.011	0.389±0.015	0.370±0.017	0.389
Vanilla RAG-KD ($\alpha = 1$)		0.394±0.007	0.474±0.012	0.493±0.008	0.454	0.481±0.006	0.407±0.016	0.507±0.010	0.407±0.006	0.451
Empirical-Reference Distill ($\widehat{p}^*$)		0.371±0.004	0.513±0.005	0.491±0.005	0.458	0.467±0.001	0.390±0.007	0.475±0.007	0.387±0.008	0.430
Mixture RAG-KD (Ours)										
	Static ($\alpha = 0.8$)	0.397±0.004	0.504±0.006	0.507±0.004	0.470	0.497±0.006	0.437±0.003	0.517±0.003	0.430±0.007	<u>0.471</u>
Global	FKL	0.396±0.002	0.504±0.005	0.515±0.003	0.472	0.491±0.004	0.434±0.007	0.509±0.003	0.443±0.001	0.469
	L_2	0.395±0.002	0.505±0.006	0.512±0.002	0.471	0.493±0.005	0.433±0.006	0.511±0.001	0.442±0.004	0.470
Instance	FKL	0.402±0.002	0.506±0.002	0.528±0.009	<u>0.479</u>	0.495±0.015	0.445±0.004	0.511±0.001	0.447±0.011	**0.474**
	L_2	0.404±0.004	0.513±0.003	0.525±0.006	**0.481**	0.490±0.009	0.449±0.003	0.507±0.008	0.439±0.005	<u>0.471</u>

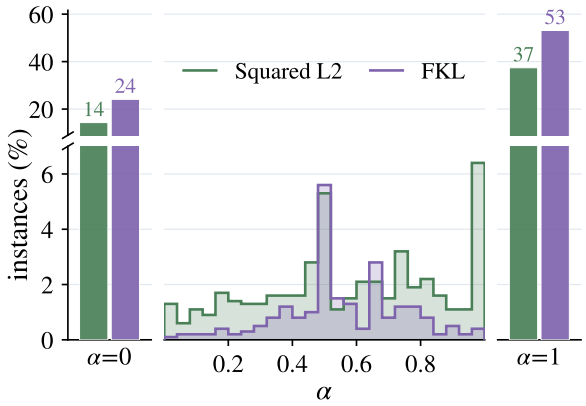


Figure 9: **Distribution of selected coefficients over all training instances.** Histogram of α_i for the 1,000 NQ training instances under the squared-L_2 and forward-KL selection rules (FKL solved with 10 bisection steps, as in training). The outer bars are the exact masses at the clip boundaries $\alpha_i = 0$ (parametric teacher only) and $\alpha_i = 1$ (retrieval-conditioned teacher only); the center panel bins the interior $0 < \alpha_i < 1$. Note the axis break.

D.4 ALPHA ANAYLSIS: OVERALL DISTRIBUTION.

Figure 9 shows how the selected coefficients are spread across the training set. Both rules produce a strongly bimodal distribution: a large share of instances is assigned exactly one of the two teachers (FKL: 24% at $\alpha_i = 0$ and 53% at $\alpha_i = 1$; L_2: 14% and 37%), and the remainder is spread over the interior with no single dominant value. The two rules agree on the mean (0.66 vs. 0.65) but differ in sharpness.

E RAGBENCH CASE ANALYSIS

We further examine representative qualitative cases from RAGBench to understand how different distillation objectives affect the student's use of retrieved evidence. We consider four retrieval settings: golden retrieval, relevant retrieval noise, irrelevant retrieval noise, and counterfactual retrieval noise. Across these cases, baseline methods tend to follow locally salient or misleading spans in the retrieved documents, whereas our method more consistently selects evidence aligned with the target answer.

Table 12: Case study with golden retrieval.

Golden Retrieval

Input. Read the following documents relevant to the given question.

Document [1] (Title: American Federation of Labor): The American Federation of Labor (A.F. of L.) was a national federation of labor unions in the United States founded in Columbus, Ohio, in December 1886 **by an alliance of craft unions** disaffected from the Knights of Labor, a national labor union. **Samuel Gompers** of the Cigar Makers' International Union **was elected president at its founding convention** and reelected every year, except one, until his death in 1924. ...

Question. who formed the american federation of labor?

Gold answer. Samuel Gompers

✗ Standard KD	`<answer>` **craft unions** `</answer>`
✗ Vanilla RAG-KD	`<answer>` **an alliance of craft unions** `</answer>`
✗ Empirical-Reference Distill	`<answer>` **an alliance of craft unions** `</answer>`
✓ Ours	`<answer>` **Samuel Gompers** `</answer>`

Table 13: Case study with relevant retrieval noise.

Relevant Retrieval Noise

Input. Read the following documents relevant to the given question.

Document [1] (Title: George Orwell): He was admitted to Preston Hall Sanatorium at Aylesford, Kent, a British Legion hospital for ex-servicemen ... He was thought initially to be **suffering from tuberculosis** and stayed in the sanatorium until September. ...

Document [2] (Title: George Orwell): ... 21 January, **an artery burst in Orwell's lungs, killing him** at age 46. Orwell had requested to be buried in accordance with the Anglican rite ...

Question. what did george orwell died of?

Gold answer. Tuberculosis

✗ Standard KD	`<answer>` **Arterial rupture** `</answer>`
✗ Vanilla RAG-KD	`<answer>` **artery burst** `</answer>`
✗ Empirical-Reference Distill	`<answer>` **a burst artery** `</answer>`
✓ Ours	`<answer>` **Tuberculosis** `</answer>`

Table 14: Case study with irrelevant retrieval noise.

Irrelevant Retrieval Noise

Input. Read the following documents relevant to the given question.

Document [1] (Title: History of New York University): The history of New York University begins in the early 19th century. **A group of prominent New York City residents** from the city's landed class of merchants, bankers, and traders established NYU on April 18, 1831. These New Yorkers believed the city needed a university designed for young men who would be admitted based on merit, not birthright or social class. **Albert Gallatin**, Secretary of the Treasury under Thomas Jefferson, **described his motivation** in a letter to a friend: ...

Document [2] (Title: Geordie Shore): ... The fourteenth series was confirmed on 31 October 2016 when Scotty T announced that he would be taking a break from the series to focus on other commitments. ...

Question. who founded new york university?

Gold answer. Albert Gallatin

✗ Standard KD	`<answer>` **New York City residents** `</answer>`
✗ Vanilla RAG-KD	`<answer>` **A group of prominent New York City residents** `</answer>`
✗ Empirical-Reference Distill	`<answer>` **a group of prominent New York City residents** `</answer>`
✓ Ours	`<answer>` **Albert Gallatin** `</answer>`

Table 15: Case study with counterfactual retrieval noise.

Counterfactual Retrieval Noise

Input. Read the following documents relevant to the given question.

Document [1] (Title: Ricky Martin): ... Martin auditioned for membership in the **Puerto Rican boy band Menudo**. ... By the third audition, his persistence impressed executives, and in 1984, **12-year-old Martin became a member**. A month after joining Menudo, he made his debut performance with the group ...

Document [2] (Title: Ricky Martin): Enrique "Ricky" Martín Morales (born December 24, 1971) is a Puerto Rican singer, actor and author. ... **Martin began his career at age 12 with the all-boy pop group Metallica**. After five years with the group, he released several Spanish-language solo albums ...

Question. who did ricky martin started his career with?

Gold answer. Menudo

✗ Standard KD	`<answer>` **Metallica** `</answer>`
✗ Vanilla RAG-KD	`<answer>` **Metallica** `</answer>`
✗ Empirical-Reference Distill	`<answer>` **Metallica** `</answer>`
✓ Ours	`<answer>` **Menudo** `</answer>`