# WILL LLMS EVENTUALLY BECOME INDISTINGUISHABLE? RETHINKING AUTHORSHIP ATTRIBUTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

LLM authorship attribution aims to identify the source model of a generated output by leveraging specific characteristics in its outputs. Existing approaches implicitly assume that different LLMs present sufficiently distinguishable generation patterns. However, this assumption may be challenged by knowledge distillation, which explicitly transfers knowledge between models while potentially altering the source-specific signals used for attribution. In this work, we systematically study how knowledge distillation affects black-box LLM authorship attribution. Across two benchmarks, nine source classes, and seven representative attribution methods, we find that knowledge distillation consistently impairs authorship attribution, causing distilled student outputs to become increasingly confused with their corresponding teachers. Moreover, we find that retraining attribution models on post-distillation outputs recovers only part of the lost attribution accuracy, indicating that the degradation cannot be fully explained by distribution shift. Instead, knowledge distillation appears to reduce the intrinsic distinguishability between student and teacher outputs. To theoretically explain this phenomenon, we develop an information-theoretic framework showing that knowledge distillation reduces the mutual information between generated text and its source, and further characterizing how much of this source information remains accessible to an attribution model. Together, our findings reveal an overlooked challenge for LLM provenance in an increasingly interconnected LLM ecosystem.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly deployed across online platforms and practical applications, leading to a growing volume of AI-generated content (Liang et al., 2024; Zhao et al., 2026). This trend has raised concerns about misinformation, copyright and fraud (Chen & Shu, 2024; Meeus et al., 2024; He et al., 2026), motivating research on AI-generated content detection (Bao et al., 2024; Hans et al., 2024). However, the binary distinction between AI and human authorship provides limited provenance information. In many practical settings, identifying the specific source model is essential for content tracing, accountability, and model usage auditing (Foley et al., 2023; Gao et al., 2025). Moreover, collapsing all LLM outputs into a single AI class overlooks model-specific patterns that can provide useful cues (Cava et al., 2024; Guo et al., 2024; 2025). These considerations motivate *LLM authorship attribution*, which aims to identify the underlying source model associated with a generated text (La Cava et al., 2026).

Early benchmarks such as TuringBench (Uchendu et al., 2021) formulate LLM authorship attribution as multi-class source identification, while subsequent studies show that different LLMs exhibit distinguishable output patterns, even among closely related models from the same family (Sun et al., 2025). Later work extends attribution to diverse model families, domains, languages, unseen generators, and distribution shifts, and develops increasingly sophisticated linguistic, stylistic, representation-level, and inference-based signals for more accurate attribution (He et al., 2024; La Cava & Tagarelli, 2025; Guo et al., 2024; Cava et al., 2024; Guo et al., 2025; La Cava et al., 2026). Most existing approaches therefore rely on different models exhibiting sufficiently distinguishable output patterns. To examine whether this assumption holds practically, we conduct a closer error analysis of Gemma-9B on ASQA, focusing on how its misclassified outputs are distributed across candidate sources. As shown in Fig. 1(a), 52.6% of all misclassified outputs are attributed to Gemma-27B alone, revealing a directional confusion. Motivated by this pattern, we then



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.

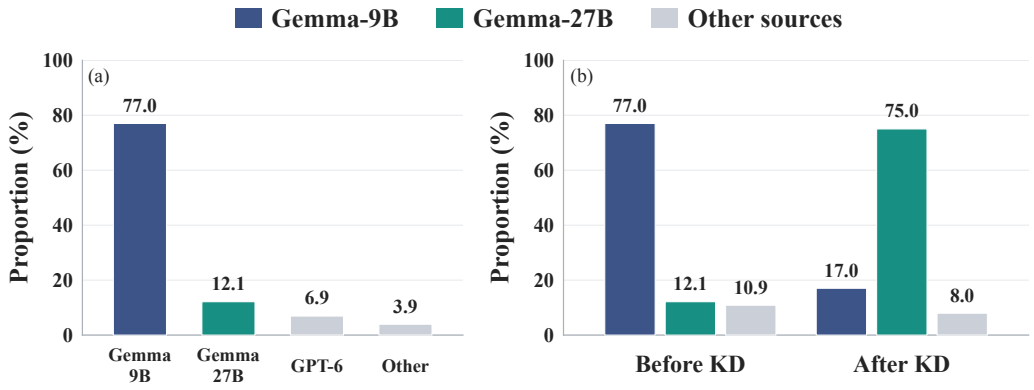



Figure 1: (a) Prediction distribution of Gemma 9B on ASQA, with six sources grouped as Other. (b) Attribution to the student, teacher, and other sources before and after KD from Gemma 27B

distill Gemma-9B using Gemma-27B as the teacher and find that self-attribution drops from 77.0% to 17.0%, while attribution to Gemma-27B rises from 12.1% to 75.0% (Fig. 1(b)). This significant observation suggests that source-specific characteristics may no longer reliably indicate the actual generator, raising a broader concern for knowledge distillation, which explicitly transfers knowledge from a model to another. This concern is practically relevant, as model-to-model knowledge transfer is already present in modern LLM development (see Appendix A for representative examples). However, its impact on black-box LLM authorship attribution remains largely underexplored. We therefore systematically study this problem through the following three research questions:

**RQ1: Does the degradation of authorship attribution induced by knowledge distillation generalize across different settings?** Motivated by the illustrative case in Fig. 1, we systematically evaluate whether the observed decrease in student source attribution and increase in teacher attribution generalize across different datasets, diverse teacher and student pairs and attribution methods.

**RQ2: What factors determine the extent of attribution degradation under knowledge distillation?** Building on RQ1, we further investigate what governs the magnitude of this effect. Specifically, we examine how attribution degradation varies with the amount of distillation data, the family relationship between teacher and student, and the form of distillation supervision.

**RQ3: Can retraining recover attribution performance after knowledge distillation?** Building on the findings of RQ1 and RQ2, we further ask whether it mainly results from a distribution shift between pre- and post-distillation student outputs. We test this by retraining the attribution models on post-distillation data and comparing the attribution accuracy before and after retraining.

Beyond these empirical findings, we provide an information-theoretic explanation showing that KD makes student generated content intrinsically harder to distinguish from the teacher, and characterizing how much of the remaining source information can be captured by a specific attribution model (see Sec. 4). Our main contributions are summarized as follows:

**Contributions.** **(1)** We identify knowledge distillation as an overlooked challenge to black-box LLM authorship attribution, raising broader concerns about the reliability of model provenance in an ecosystem where LLMs increasingly learn from one another. **(2)** We conduct extensive experiments that consistently demonstrate the degradation of LLM authorship attribution under knowledge distillation. Furthermore, retraining on post-distillation outputs only partially recovers the lost performance, indicating that the degradation cannot be explained solely by distribution shift. **(3)** We develop an information-theoretic framework based on mutual information, providing a principled explanation for why knowledge distillation impairs LLM authorship attribution.

## 2 RELATED WORK

### 2.1 LLM AUTHORSHIP ATTRIBUTION

The growing prevalence of LLM-generated content has motivated extensive research on machine-generated text detection, which primarily distinguishes human-written from AI-generated text us-

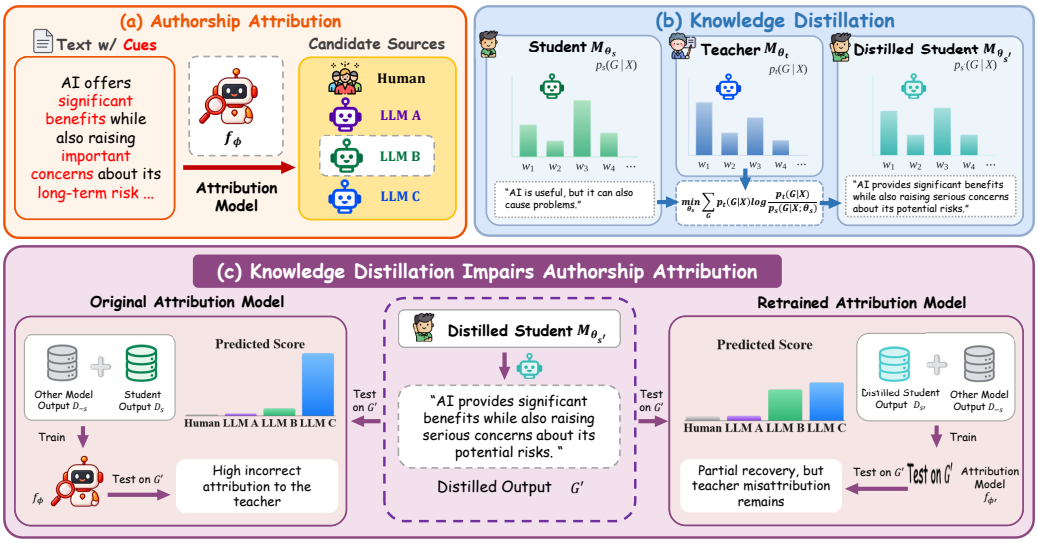



Figure 2: Overview of how knowledge distillation affects LLM authorship attribution. (a) An attribution model identifies the source of a given text among candidate human and LLM sources. (b) Knowledge distillation shifts the student's generation distribution toward that of the teacher. (c) Left: KD substantially increases teacher misattribution under the original attribution model. Right: retraining partially restores student attribution, but does not fully eliminate teacher misattribution.

ing representation-based, linguistic, and probability-based signals (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024). Beyond binary detection, authorship attribution (AA) seeks to identify the specific source of a text among a set of candidate authors, traditionally through stylometric and linguistic characteristics (Koppel et al., 2009; Stamatatos, 2009). This formulation has subsequently been extended from human authors to neural text generators, with TuringBench among the earliest benchmarks to cast source identification as multi-class attribution across different generators (Uchendu et al., 2021). In this work, we use *LLM authorship attribution* to denote the detailed task of identifying the LLM responsible for generating a given text.

Building on this formulation, later works have extended this task across models, domains, languages, and challenging deployment conditions, including paraphrasing, manipulated texts, distribution shifts, and unseen generators (Wang et al., 2024; He et al., 2024; La Cava & Tagarelli, 2025; Liu et al., 2025; La Cava et al., 2026). Meanwhile, recent methods exploit increasingly detailed source signals, including contrastive representations (Guo et al., 2024; Cava et al., 2024), topological writing patterns (Uchendu et al., 2024), and context aware inference patterns (Guo et al., 2025). Despite differences in evaluation settings and modeling strategies, these approaches implicitly rely on source models exhibiting sufficiently distinguishable characteristics for reliable attribution. Yet when and why this distinguishability may break down has received limited attention. One closely related study (Sun et al., 2025) demonstrates that different LLMs exhibit distinguishable model idiosyncrasies and that these patterns can propagate through fine-tuning. Another study investigates whether the teacher can be identified from the student's outputs, showing that teacher specific signals can enable tracing after distillation (Wadhwa et al., 2025). However, neither work examines whether knowledge distillation impairs attribution of an output to the model that generated it.

## 2.2 KNOWLEDGE DISTILLATION

Knowledge distillation transfers knowledge from a teacher model to a student model, originally for model compression and efficient inference (Hinton et al., 2015). For language models, existing work has explored knowledge transfer at multiple levels, including teacher generated sequence responses and rationales (Hsieh et al., 2023), token-level output distributions (Gu et al., 2024; Agarwal et al., 2024; Shing et al., 2025), and intermediate representations or attention patterns (Devlin et al., 2019; Wang et al., 2020; Zhang et al., 2024). Beyond task capability, recent studies further show that properties such as watermarks and membership or memorized information can also propagate through distillation (Pan et al., 2025; Zhang et al., 2025). These studies primarily focus on the propagation

of known and directly testable properties, leaving unclear how distillation affects the implicit source specific characteristics underlying black-box LLM authorship attribution.

## 3 PRELIMINARIES

**LLM Authorship Attribution** We consider a set of $K$ candidate LLM sources $\mathcal{M} = \{M_{\theta_1}, \ldots, M_{\theta_K}\}$, together with a human source. Let $X$ denote a prompt, $G$ the corresponding response text, and $Y \in \mathcal{Y} = \{0, 1, \ldots, K\}$ its source label. We use $Y = 0$ for human-written text, while $Y = k$, for $k \in \{1, \ldots, K\}$, indicates that the response is generated by $M_{\theta_k}$. Each LLM $M_{\theta_k}$ induces a conditional generation distribution $p_{\theta_k}(g \mid x)$. Following (La Cava et al., 2026), we adopt a black-box attribution setting in which the source model itself is unknown and must be inferred solely from the generated text. Under this setting, the objective of authorship attribution is to learn a function $f_\phi : \mathcal{G} \to \mathcal{Y}$ that minimizes the source classification error:

$$\hat{\phi} = \arg\min_{\phi \in \Phi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[f_\phi(g_i) \neq y_i], \tag{1}$$

where $\mathcal{D} = \{(g_i, y_i)\}_{i=1}^{N}$ consists of generated text and source pairs, with $g_i \in \mathcal{G}$ and $y_i \in \mathcal{Y}$.

**Knowledge Distillation.** Knowledge distillation is widely adopted to improve the performance and computational efficiency of a student model by transferring knowledge from a teacher model through teacher derived supervision (Hinton et al., 2015; Kim & Rush, 2016). Let $M_{\theta_t}$ and $M_{\theta_s}$ denote the teacher and student models, respectively, with $t \neq s$. Before distillation, the teacher and student induce conditional output distributions $p_{\theta_t}(G \mid x)$ and $p_{\theta_s}(G \mid x)$ for a prompt $x$. Specifically, knowledge distillation transfers the teacher's knowledge to the student by minimizing the discrepancy between their output distributions:

$$\mathcal{L}_{\mathrm{KD}}(\theta; \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{E}_{G \sim p_{\theta_t}(\cdot \mid x)} \left[ \log \frac{p_{\theta_t}(G \mid x)}{p_\theta(G \mid x)} \right], \qquad \theta'_s = \arg\min_\theta \mathcal{L}_{\mathrm{KD}}(\theta; \mathcal{D}). \tag{2}$$

Following prior model-provenance work with their model lineage (Foley et al., 2023), we regard the distilled student $M'_s$ as retaining the same source identity $s$ as the original student $M_s$, while inducing a new output distribution $p_{\theta'_s}(G \mid x)$.

**Authorship Attribution under Knowledge Distillation.** Following Eq. 2, knowledge distillation updates the student from $M_{\theta_s}$ to $M_{\theta'_s}$, inducing a shift in its generation distribution from $p_{\theta_s}(G \mid X)$ to $p_{\theta'_s}(G \mid X)$ while preserving the same source identity $s$. Let $\mathcal{D}_s$ and $\mathcal{D}'_s$ denote attribution data generated by the student before and after distillation, respectively, and let $\mathcal{D}_{-s}$ denote data from all the remaining source models. Building on Eq. 1, we define the original attribution model $f_{\phi_s}$ and the retrained attribution model $f_{\phi'_s}$ through

$$\phi_s = \arg\min_\phi \mathcal{L}_{\mathrm{AA}}(\phi; \mathcal{D}_{-s} \cup \mathcal{D}_s), \qquad \phi'_s = \arg\min_\phi \mathcal{L}_{\mathrm{AA}}(\phi; \mathcal{D}_{-s} \cup \mathcal{D}'_s). \tag{3}$$

Here, $f_{\phi_s}$ is trained on pre-distillation student outputs, denoted by $G \sim p_{\theta_s}(\cdot \mid X)$, whereas $f_{\phi'_s}$ is retrained on post-distillation student outputs, denoted by $G' \sim p_{\theta'_s}(\cdot \mid X)$. For a student source $s$, we define the student source attribution accuracy as

$$\mathrm{Acc}_s(\phi, \theta) = \mathbb{E}_X \mathbb{E}_{G \sim p_\theta(\cdot \mid X)} [\mathbf{1}\{f_\phi(G) = s\}]. \tag{4}$$

Since knowledge distillation changes the student's generation distribution without changing its source identity, both $G$ and $G'$ retain the source label $s$. We therefore focus on two cases and quantify the attribution degradation induced by knowledge distillation under the original attribution model and the recovery obtained by retraining on post-distillation student outputs by $\Delta_b$ and $\Delta_a$:

$$\Delta_b := \mathrm{Acc}_s(\phi_s, \theta_s) - \mathrm{Acc}_s(\phi_s, \theta'_s), \qquad \Delta_a := \mathrm{Acc}_s(\phi'_s, \theta'_s) - \mathrm{Acc}_s(\phi_s, \theta'_s). \tag{5}$$

## 4 THEORETICAL ANALYSIS

Our analysis provides a local characterization of the underlying mechanism rather than a complete account of heterogeneous teacher and student architectures. We consider a shared local parameterization around the teacher, under which we show that distribution matching KD reduces intrinsic source information and characterize how much of it remains accessible to an attribution model.

### 4.1 KD Reduces Source Information in Generated Text

As discussed in Section 3, knowledge distillation changes the generation distribution associated with the student source while preserving its source identity. To characterize the intrinsic effect of this distribution shift, we focus on the distinguishability between the teacher and the student, which is the source pair directly connected by the distillation process. For this analysis, let $Y \in \{t, s\}$ denote the source identity of a response generated by the teacher or the student, with $P(Y = t) = P(Y = s) = 1/2$ and $Y \perp X$. The conditional generation distributions are given by $p(G \mid X, Y = t) = p_{\theta_t}(G \mid X)$ for the teacher and $p(G \mid X, Y = s) = p_\theta(G \mid X)$ for the student. We therefore define the intrinsic teacher and student source information as

$$\mathcal{I}(\theta) := I_\theta(Y; G \mid X) = H_\theta(Y \mid X) - H_\theta(Y \mid X, G). \tag{6}$$

For $\theta \in \Theta$ and $r > 0$, define the local neighborhood

$$B(\theta, r) := \{\vartheta \in \Theta : \|\vartheta - \theta\|_2 < r\}. \tag{7}$$

For brevity, we write $\mathcal{L}_{\mathrm{KD}}(\theta) := \mathcal{L}_{\mathrm{KD}}(\theta; \mathcal{D})$ in the following analysis.

**Assumption 1.** There exist $r_0 > 0$, $L_K > 0$, and $L_I > 0$ such that, for any $\theta_1, \theta_2 \in B(\theta_t, r_0)$, $\left\|\nabla^2 \mathcal{L}_{\mathrm{KD}}(\theta_1) - \nabla^2 \mathcal{L}_{\mathrm{KD}}(\theta_2)\right\|_2 \le L_K \|\theta_1 - \theta_2\|_2$ and $\left\|\nabla^2 \mathcal{I}(\theta_1) - \nabla^2 \mathcal{I}(\theta_2)\right\|_2 \le L_I \|\theta_1 - \theta_2\|_2$.

**Assumption 2.** At $\theta_t$, let $F := \mathbb{E}_{X,G \sim p_{\theta_t}(\cdot|X)}\left[\nabla_\theta \log p_\theta(G \mid X)\big|_{\theta=\theta_t} \nabla_\theta \log p_\theta(G \mid X)^\top\big|_{\theta=\theta_t}\right]$ denote the Fisher information matrix. There exists a constant $\mu > 0$ such that $v^\top F v \ge \mu \|v\|_2^2$ for every $v \in \mathbb{R}^d$.

Assumptions 1 and 2 are standard local regularity conditions. Locally Lipschitz Hessians have been adopted in theoretical analyses of knowledge distillation (Xie et al., 2026), while non-degenerate Fisher information similarly underlies recent Fisher-based analyses of LLM distillation (Hamman et al., 2025).

**Theorem 1.** *Let $\tau \ge 0$ denote the continuous distillation time and let $M_{\theta(\tau)}$ denote the student model at time $\tau$, with $\theta(0) = \theta_s$. Under Assumptions 1 and 2, suppose $0 < \|\theta_s - \theta_t\|_2 < \min\left\{r_0, \frac{2\mu}{L_K + 4L_I}\right\}$. Then, for any $0 \le \tau_1 < \tau_2$,*

$$\mathcal{I}(\theta(\tau_2)) < \mathcal{I}(\theta(\tau_1)). \tag{8}$$

*In particular, for $\theta'_s = \theta(T)$ with $T > 0$, $\mathcal{I}(\theta'_s) < \mathcal{I}(\theta_s)$.*

Theorem 1 shows that the generated text itself carries progressively less information for distinguishing the student from its teacher. Since $\mathcal{I}(\theta)$ is independent of the attribution model, this reduction reflects an intrinsic loss of source distinguishability rather than merely a mismatch of the attribution boundary. The complete proof is provided in Appendix B.1.

### 4.2 Source Information Retained by Attribution Models under KD

Theorem 1 establishes that KD reduces the source information intrinsically carried by generated text. We next characterize how much of this information is accessible to a particular attribution model. For an attribution model $f_\phi$, let $Z_\phi = f_\phi(G)$ denote its output, and define

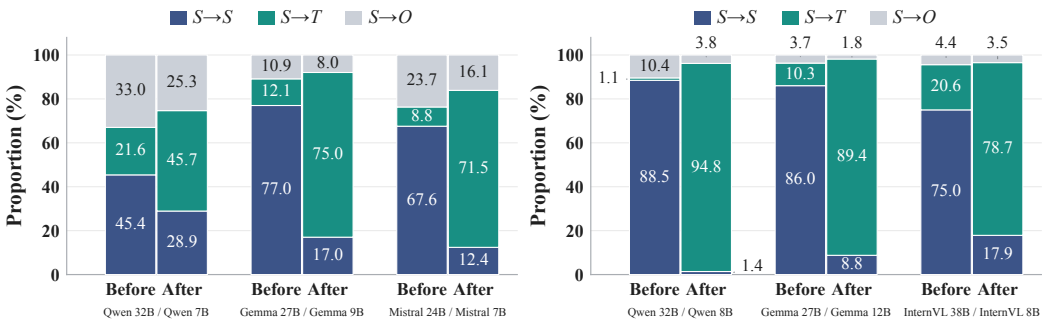
$$\mathcal{I}_\phi(\theta) := I_\theta(Y; Z_\phi \mid X). \tag{9}$$

Since $Z_\phi$ is determined by $G$, the data processing inequality immediately yields $\mathcal{I}_\phi(\theta) \le \mathcal{I}(\theta)$.

To characterize how source information is retained by different attribution models, consider a local perturbation $\theta_t + \epsilon v$ around the teacher parameters $\theta_t$, where $v$ denotes a fixed parameter direction. For a generated response $(X, G)$, define its local source evidence as

$$h(X, G) := \frac{d}{d\epsilon} \log \frac{p_{\theta_t + \epsilon v}(G \mid X)}{p_{\theta_t}(G \mid X)}\bigg|_{\epsilon=0}. \tag{10}$$

Here, $h(X, G)$ denotes the local source discriminative signal along $v$. Its variance $\mathrm{Var}(h)$ quantifies the total dispersion of this signal, while $\mathbb{E}[\mathrm{Var}(h \mid X, Z_\phi)]$ quantifies the residual dispersion after conditioning on the attribution output.




(a) ASQA. (b) ScienceQA.

Figure 3: Attribution results before and after knowledge distillation. Bars show the proportions of student responses attributed to the student ($S \to S$), teacher ($S \to T$), or other sources ($S \to O$), averaged across the seven attribution methods in Table 1.

**Theorem 2.** *If the regularity conditions of Theorem 1 hold and* $0 < \mathrm{Var}(h) < \infty$*, then the accessible source information satisfies*

$$\lim_{\epsilon \to 0} \frac{\mathcal{I}_\phi(\theta_t + \epsilon v)}{\mathcal{I}(\theta_t + \epsilon v)} = 1 - \frac{\mathbb{E}[\mathrm{Var}(h \mid X, Z_\phi)]}{\mathrm{Var}(h)}. \tag{11}$$

Theorem 2 shows that the fraction of intrinsic source information retained by an attribution model is determined by the unresolved variation $\mathbb{E}[\mathrm{Var}(h \mid X, Z_\phi)]$ relative to the total variation $\mathrm{Var}(h)$. A smaller conditional variation implies greater retention of $\mathcal{I}(\theta)$ in $\mathcal{I}_\phi(\theta)$. Consequently, different attribution models may preserve different fractions of the same intrinsic source information. The complete proof is provided in Appendix B.2.

## 5 Experiments

We conduct experiments to systematically investigate how knowledge distillation affects LLM authorship attribution. We first examine whether knowledge distillation degrades the performance of existing attribution methods **(RQ1)**, then investigate what factors determine the extent of this degradation **(RQ2)**, and finally assess whether, and to what extent, retraining on post-distillation outputs can recover attribution performance **(RQ3)**.

### 5.1 Experimental Setup

**Benchmarks and Models.** We evaluate on ASQA (Stelmakh et al., 2022) and ScienceQA (Lu et al., 2022), covering text-only and multimodal question answering, respectively. Attribution is performed in a black-box setting using only generated response text. Each benchmark contains nine source classes: six open-weight models from three families (Qwen2.5, Gemma-2, and Mistral on ASQA; Qwen3-VL, Gemma-3, and InternVL3.5 on ScienceQA), two closed-source models (GPT-6 Astra and MiniMax-M3), and one human source. Importantly, this model pool also allows us to distill the same student from different teachers while keeping the student and training protocol fixed. This design enables us to examine whether attribution changes systematically follow the source of teacher supervision, rather than merely reflecting an undirected shift caused by student adaptation. Detailed settings are provided in Appendix C.1.

**Attribution Methods and Metrics.** To provide broad coverage of existing LLM authorship attribution approaches, we evaluate seven representative baselines spanning complementary modeling paradigms: BERT-MLP (Devlin et al., 2019), WhoSAI (Cava et al., 2024), DeTeCtive (Guo et al., 2024), MDOK (Macko, 2025), PROFILER (Guo et al., 2025), OTBDetector (La Cava & Tagarelli, 2025), and HLD-Detector (Guo et al., 2026). Collectively, these methods cover supervised representation-based classification, contrastive learning, fine-tuned language-model classifiers, inference patterns, and hierarchical linguistic-distribution signals. This selection also follows recent fine-grained attribution studies that consider diverse source signals and modeling

Table 1: Student source attribution accuracy (%) before and after KD across three teacher and student pairs per benchmark. Positive $\Delta_b$ (Eq. 5) indicates KD induced attribution degradation.

| Dataset | Teacher / Student | Stage | BERT-MLP. | WhosAI | DeTeCtive | mdok | PROF | OTB | HLD | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| ASQA | Qwen2.5-32B Qwen2.5-7B | Before | 20.57 | 64.66 | 68.78 | 70.25 | 21.73 | 31.75 | 40.08 | **45.40** |
| | | After | 12.87 | 47.68 | 33.44 | 35.44 | 18.99 | 22.89 | 31.22 | **28.93** |
| | | $\Delta_b$ | +7.70 | +16.98 | +35.34 | +34.81 | +2.74 | +8.86 | +8.86 | **+16.47** |
| | Gemma-2-27B Gemma-2-9B | Before | 80.06 | 89.56 | 87.66 | 94.73 | 45.36 | 89.03 | 52.64 | **77.01** |
| | | After | 25.84 | 16.98 | 10.23 | 17.19 | 11.71 | 15.93 | 21.31 | **17.03** |
| | | $\Delta_b$ | +54.22 | +72.58 | +77.43 | +77.54 | +33.65 | +73.10 | +31.33 | **+59.98** |
| | Mistral-24B Mistral-7B | Before | 56.33 | 70.36 | 80.80 | 80.91 | 52.74 | 79.85 | 52.00 | **67.57** |
| | | After | 16.03 | 3.80 | 8.33 | 4.96 | 34.60 | 5.38 | 13.71 | **12.40** |
| | | $\Delta_b$ | +40.30 | +66.56 | +72.47 | +75.95 | +18.14 | +74.47 | +38.29 | **+55.17** |
| ScienceQA | Qwen3-VL-32B Qwen3-VL-8B | Before | 86.63 | 96.17 | 98.30 | 96.51 | 57.58 | 97.44 | 86.71 | **88.48** |
| | | After | 2.04 | 0.26 | 0.09 | 0.26 | 2.21 | 0.26 | 4.60 | **1.39** |
| | | $\Delta_b$ | +84.59 | +95.91 | +98.21 | +96.25 | +55.37 | +97.18 | +82.11 | **+87.09** |
| | Gemma-3-27B Gemma-3-12B | Before | 86.12 | 97.87 | 95.40 | 96.17 | 48.04 | 93.36 | 85.26 | **86.03** |
| | | After | 9.37 | 8.94 | 4.26 | 9.63 | 12.35 | 4.60 | 12.27 | **8.77** |
| | | $\Delta_b$ | +76.75 | +88.93 | +91.14 | +86.54 | +35.69 | +88.76 | +72.99 | **+77.26** |
| | InternVL3.5-38B InternVL3.5-8B | Before | 72.91 | 82.37 | 86.88 | 86.88 | 46.34 | 76.06 | 73.59 | **75.00** |
| | | After | 14.48 | 10.99 | 14.22 | 22.57 | 24.70 | 17.89 | 20.19 | **17.86** |
| | | $\Delta_b$ | +58.43 | +71.38 | +72.66 | +64.31 | +21.64 | +58.17 | +53.40 | **+57.14** |

paradigms (La Cava & Tagarelli, 2025; La Cava et al., 2026). Detailed implementation settings are provided in Appendix C.3.

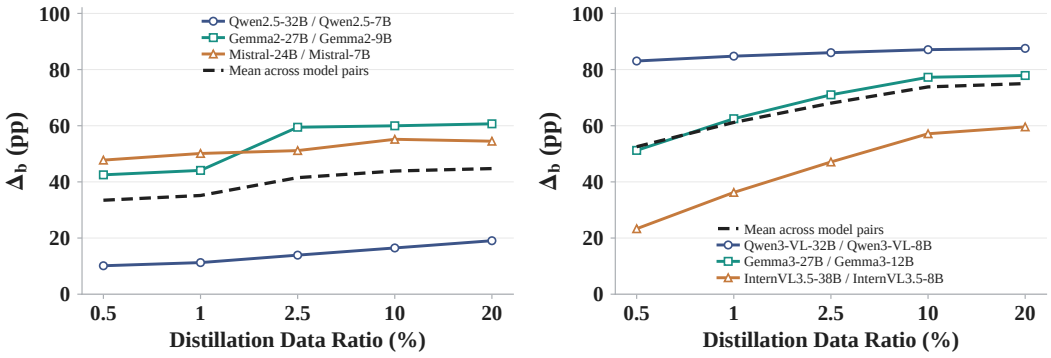
### 5.2 RQ1: How Does Knowledge Distillation Affect LLM Authorship Attribution?

Table 1 summarizes student source attribution accuracy before and after KD across all evaluated settings and shows a consistent degradation in attribution performance after distillation. Specifically, we observe average reductions in student source accuracy of 16.47, 59.98, and 55.17 percentage points for the three model pairs on ASQA, and 87.09, 77.26, and 57.14 percentage points for those on ScienceQA, respectively. The degradation is particularly pronounced for several methods. For example, DeTeCtive on the Qwen3-VL pair in ScienceQA drops from 98.30% to 0.09%, corresponding to a 98.21-point decrease. Similar reductions are observed for WhosAI, mdok, and OTBDetector. More generally, attribution accuracy falls below 20% in most evaluated settings. This indicates that, in many cases, existing attribution methods fail to correctly attribute distilled outputs to the student model. Figure 3 further shows where the lost student attribution is redirected after KD. Across all model pairs, the decrease in $S \to S$ is largely mirrored by an increase in $S \to T$, indicating that most attribution errors shift toward the corresponding teacher. For example, on Gemma-2 in ASQA, $S \to S$ drops from 77.0% to 17.0%, while $S \to T$ rises from 12.1% to 75.0%.

To rule out ineffective distillation as the cause of attribution degradation, we also report downstream task utility before and after KD. As shown in Table D.1, task performance improves across all evaluated settings, showing that the degradation persists despite successful knowledge transfer.

### 5.3 RQ2: What Factors Influence Attribution Degradation under Knowledge Distillation?

Given the consistent degradation observed in RQ1, we further investigate the factors that govern its magnitude. We focus on the amount of distillation data, the teacher and student family relation, and the form of distillation supervision. First, Figure 4 examines the effect of the amount of distillation data on attribution degradation. Across both benchmarks, the mean $\Delta_b$ increases monotonically with the distillation ratio, indicating that greater exposure to teacher supervision generally leads to stronger attribution degradation. Notably, the degradation also exhibits a saturation trend, with the saturation point varying across different teacher and student pairs. For example, on ASQA, the Gemma-2 pair changes little beyond a 2.5% ratio, whereas the Mistral pair continues to increase until approximately 10%. These results show that, for some teacher and student pairs, even a small fraction of distillation data is sufficient to substantially impair LLM authorship attribution.




(a) ASQA. (b) ScienceQA.

Figure 4: Effect of the distillation data ratio on attribution degradation. $\Delta_b$ denotes the decrease in attribution accuracy. Each solid line represents one teacher and student pair, averaged across the seven attribution methods. The dashed line denotes the mean across the three model pairs.

Second, we examine the effect of the teacher and student family relation in Table 2 and find that cross-family distillation also substantially reduces attribution accuracy. More strikingly, all evaluated cross-family teachers induce larger degradation than the corresponding same-family teacher. On ASQA, the same-family Gemma-2 teacher yields a 59.98-point decrease, whereas Qwen2.5 and Mistral produce decreases of 76.44 and 71.15 points, respectively. Similarly, on ScienceQA, the same-family Gemma-3 teacher causes a 77.26-point decrease, compared with 84.39 and 83.83 points for Qwen3-VL and InternVL3.5. These results show that attribution degradation is not restricted to within-family distillation and can be even more pronounced under cross-family transfer in the evaluated settings.

Table 2: Effect of model family relation on student source attribution accuracy. Values are seven attribution method means (%).

| Dataset / Student | Teacher | Before | After | $\Delta_b$ |
|---|---|---|---|---|
| ASQA Gemma-2-9B | Qwen2.5-32B | 77.01 | 0.57 | +76.44 |
| | Gemma-2-27B | 77.01 | 17.03 | +59.98 |
| | Mistral-24B | 77.01 | 5.86 | +71.15 |
| | GPT-6 Astra | 77.01 | 0.17 | +76.84 |
| | MiniMax-M3 | 77.01 | 1.59 | +75.42 |
| ScienceQA Gemma-3-12B | Qwen3-VL-32B | 86.03 | 1.64 | +84.39 |
| | Gemma-3-27B | 86.03 | 8.77 | +77.26 |
| | InternVL3.5-38B | 86.03 | 2.20 | +83.83 |
| | GPT-6 Astra | 86.03 | 0.38 | +85.65 |
| | MiniMax-M3 | 86.03 | 2.60 | +83.43 |

Finally, we examine whether attribution degradation depends on the form of teacher supervision. Table 3 compares three pure distillation objectives: sequence-level, distribution-level, and hidden-state distillation. All three objectives cause substantial degradation in student source attribution accuracy, with $\Delta_b$ ranging from 56.42 to 59.98 percentage points on ASQA and from 73.44 to 77.26 points on ScienceQA. This consistency indicates that the degradation is not specific to a particular distillation objective. Sequence-level distillation yields the largest degradation on both benchmarks, with $\Delta_b$ of 59.98 points on ASQA and 77.26 points on ScienceQA, although the differences across objectives are relatively modest. One possible explanation is that sequence-level distillation directly encourages the student to imitate teacher-generated responses, and thus may more directly transfer source-specific textual patterns observable to black-box attribution methods. In contrast, distribution- and hidden-state-level supervision operate on internal model signals, whose effects must first be expressed through generated text before becoming observable to the attribution model.

### 5.4 RQ3: To What Extent Can Retraining Recover Attribution Performance after Knowledge Distillation?

RQ1 and RQ2 show that knowledge distillation consistently degrades attribution accuracy across model pairs and distillation settings. A natural question is whether this simply reflects an outdated attribution model trained on pre-distillation outputs. To test this, we retrain each attribution method using post-distillation student responses, while keeping all other source classes and evaluation questions unchanged.

Table 4: Effect of retraining the attribution model after knowledge distillation. Values are seven attribution method mean changes in percentage points. $\Delta_b$, $\Delta_a$, and $\Delta_{\text{res}} = \Delta_b - \Delta_a$ denote KD-induced degradation, retraining recovery, and the residual gap in student source accuracy, respectively. Here, TAR is the proportion of student responses attributed to the corresponding teacher.

| Dataset | Type | Teacher → Student | Student Source Accuracy | | | Teacher Attribution Rate | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\Delta_b$ | $\Delta_a$ | $\Delta_{\text{res}}$ | $\Delta\text{TAR}_b$ | $\Delta\text{TAR}_a$ | $\Delta\text{TAR}_{\text{res}}$ |
| ASQA | Open | Qwen2.5-32B → Qwen2.5-7B | 16.47 | 6.85 | 9.62 | 24.11 | 12.56 | 11.55 |
| | Open | Gemma-2-27B → Gemma-2-9B | 59.98 | 48.63 | 11.35 | 62.88 | 49.08 | 13.80 |
| | Open | Mistral-24B → Mistral-7B | 55.17 | 43.92 | 11.25 | 62.75 | 50.93 | 11.82 |
| | API | GPT-6 Astra → Qwen2.5-7B | 45.12 | 25.43 | 19.69 | 84.93 | 59.51 | 25.42 |
| | API | MiniMax-M3 → Qwen2.5-7B | 42.71 | 23.97 | 18.74 | 78.00 | 58.15 | 19.85 |
| ScienceQA | Open | Qwen3-VL-32B → Qwen3-VL-8B | 87.09 | 69.21 | 17.88 | 93.70 | 68.61 | 25.09 |
| | Open | Gemma-3-27B → Gemma-3-12B | 77.26 | 61.49 | 15.77 | 79.12 | 61.18 | 17.94 |
| | Open | InternVL3.5-38B → InternVL3.5-8B | 57.14 | 44.78 | 12.36 | 58.08 | 44.61 | 13.47 |
| | API | GPT-6 Astra → Qwen3-VL-8B | 88.42 | 72.10 | 16.32 | 98.56 | 71.72 | 26.84 |
| | API | MiniMax-M3 → Qwen3-VL-8B | 86.53 | 71.99 | 14.54 | 89.69 | 70.56 | 19.13 |

As shown in Table 4, retraining recovers a substantial portion of the lost student source accuracy, but a non-negligible residual gap remains in every setting. On ASQA, $\Delta_{\text{res}}$ ranges from 9.62 to 19.69 percentage points, while on ScienceQA it ranges from 12.36 to 17.88 points. The largest residual gap is observed for GPT-6 Astra → Qwen2.5-7B on ASQA, reaching 19.69 percentage points. Thus, although retraining adapts the attribution model to post-distillation student outputs, it does not restore the pre-distillation level of student-source attribution accuracy. The same pattern is even more apparent in teacher attribution. After retraining, $\Delta\text{TAR}_{\text{res}}$ remains positive for all ten settings, ranging from 11.55 to 26.84 percentage points and reaching 26.84 points for GPT-6 Astra → Qwen3-VL-8B on ScienceQA. In other words, even when the attribution model is explicitly retrained on outputs from the distilled student, these outputs remain more likely to be attributed to the teacher than before distillation.

These results show that distribution shift in the attribution model explains only part of the degradation. Retraining can adapt the decision boundary to the post-distillation distribution, but it does not fully restore pre-distillation attribution performance. This persistent residual gap, together with the remaining teacher-directed confusion, suggests that distribution shift alone is insufficient to explain the observed degradation and is consistent with reduced source distinguishability between the student and teacher.

Table 3: Effect of pure knowledge distillation form on student source attribution accuracy. Values are seven attribution method means (%).

| Dataset / T–S Pair | KD Form | Before | After | $\Delta_b$ |
|---|---|---|---|---|
| ASQA | Sequence-level | 77.01 | 17.03 | +59.98 |
| Gemma-2-27B | Distribution-level | 77.01 | 20.59 | +56.42 |
| Gemma-2-9B | Hidden-state | 77.01 | 20.05 | +56.96 |
| ScienceQA | Sequence-level | 86.03 | 8.77 | +77.26 |
| Gemma-3-27B | Distribution-level | 86.03 | 11.95 | +74.08 |
| Gemma-3-12B | Hidden-state | 86.03 | 12.59 | +73.44 |

## 6 Conclusion

In this work, we systematically study how knowledge distillation affects black-box LLM authorship attribution. Across two benchmarks and seven attribution methods, we consistently find that KD reduces student-source attribution accuracy while redirecting predictions toward the corresponding teacher. This effect persists across the amount of distillation data, model family relations, and distillation forms. Although retraining on post-distillation outputs recovers some performance, it does not fully restore the pre-distillation level, leaving a considerable residual gap. Beyond these empirical findings, our information-theoretic analysis shows that KD reduces the intrinsic source information distinguishing the student from its teacher. Together, these findings highlight knowledge distillation as a key challenge to reliable LLM provenance.

Our study has several limitations. First, we consider one distilled student at a time, while real-world model ecosystems may involve multiple knowledge distillation processes. Second, we examine three major factors affecting attribution degradation, but do not claim that these are the only relevant factors. Finally, our work focuses on identifying and characterizing this challenge, while developing attribution methods that remain reliable is left for future work.

### AI USE STATEMENT

Generative AI tools were used only for language polishing and improving readability. All scientific ideas, theoretical analyses, experiments, and conclusions were developed by the authors, who take full responsibility for the final content.

### ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. We use publicly available datasets and model generated text solely for research purposes. We do not identify or profile real individuals, and no additional ethical concerns are identified.

### REPRODUCIBILITY STATEMENT

Details of the datasets, model configurations, experimental settings, evaluation metrics, and implementation are provided in the main paper and appendix. Our code and experimental configurations can be found in the anonymous repository at https://anonymous.4open.science/r/kd-authorship-112D.

## REFERENCES

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

Anthropic. Detecting and preventing distillation attacks, February 2026a. URL https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks. Anthropic.

Anthropic. Countering misuse of ai: September 2026, September 2026b. URL https://www.anthropic.com/threat-intelligence-report-september-2026. Anthropic Threat Intelligence Report.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, 2024.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. Is contrasting all you need? contrastive learning for the detection and attribution of ai-generated text. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 3179–3186, 2024.

Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? In *International Conference on Learning Representations*, pp. 34687–34726, 2024.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Myles Foley, Ambrish Rawat, Taesung Lee, Yufang Hou, Gabriele Picco, and Giulio Zizzo. Matching pairs: Attributing fine-tuned models to their pre-trained large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7423–7442, 2023.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this api serving? In *International Conference on Learning Representations*, pp. 86369–86382, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, pp. 32694–32717, 2024.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Guangyu Shen, Kaiyuan Zhang, Shengwei An, Guanhong Tao, and Xiangyu Zhang. Profiler: Black-box AI-generated text origin detection via context-aware inference pattern analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24892–24912, 2025.

Rui Guo, Weibin Zeng, Fuzhang Wu, Yan Kong, Sicheng Shen, Yanjun Wu, and Weiming Dong. Hld: Approximate hierarchical linguistic distribution modeling for llm-generated text detection. In *International Conference on Learning Representations*, 2026.

Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning. In *Advances in Neural Information Processing Systems*, pp. 88320–88347, 2024.

Faisal Hamman, Pasan Dissanayake, Yanjun Fu, and Sanghamitra Dutta. Few-shot knowledge distillation of llms with counterfactual explanations. In *Advances in Neural Information Processing Systems*, volume 38, pp. 165977–166008, 2025.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 17519–17537, 2024.

Haorui He, Yupeng Li, Dacheng Wen, Yang Chen, Reynold Cheng, Donglong Chen, and Francis C. M. Lau. Debating truth: Debate-driven claim verification with multiple large language model agents. In *Proceedings of the ACM Web Conference 2026*, pp. 8851–8861, 2026.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 2251–2265, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.

Lucio La Cava and Andrea Tagarelli. OpenTuringBench: An open-model-based benchmark and framework for machine-generated text detection and attribution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26655–26671, 2025.

Lucio La Cava, Dominik Macko, Robert Moro, Ivan Srba, and Andrea Tagarelli. Authorship attribution in multilingual machine-generated texts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45136–45152, 2026.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, Jinkai Hu, Jiayao Li, Rui Gao, Zekun Li, Songquan Zhu, Jingkai Zhou, and Pengyu Zhao. Minimax sparse attention, 2026.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel Mcfarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 29575–29620, 2024.

Yule Liu, Zhiyuan Zhong, Yifan Liao, Zhen Sun, Jingyi Zheng, Jiaheng Wei, Qingyuan Gong, Fenghua Tong, Yang Chen, Yang Zhang, and Xinlei He. On the generalization and adaptation ability of machine-generated text detectors in academic writing. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5674–5685, 2025.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Dominik Macko. mdok of kinit: Robustly fine-tuned LLM for binary and multiclass ai-generated text detection. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, pp. 3819–3826, 2025.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre De Montjoye. Copyright traps for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 35296–35309, 2024.

Mistral AI Team. Mistral Small 3. Mistral AI, 2025.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

OpenAI. Model distillation in the api, October 2024. URL https://openai.com/index/api-model-distillation/. OpenAI.

OpenAI. GPT-6 Astra System Card. OpenAI, 2026a.

OpenAI. Updated stakes for american-led, democratic ai, February 2026b. Memo to the U.S. House Select Committee on Strategic Competition between the United States and the Chinese Communist Party.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. Can LLM watermarks robustly prevent unauthorized knowledge distillation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13228–13251, 2025.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. Taid: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *International Conference on Learning Representations*, pp. 92233–92256, 2025.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57854–57885. PMLR, 13–19 Jul 2025.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2001–2016, 2021.

Adaku Uchendu, Thai Le, and Dongwon Lee. Topformer: Topology-aware authorship attribution of deepfake texts with diverse writing styles. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 1446–1454, 2024.

Somin Wadhwa, Chantal Shaib, Silvio Amir, and Byron C Wallace. Who taught you that? tracing teachers in model distillation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3307–3315, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3964–3992, 2024.

Audrey Xie, Ludwig Schmidt, and John Duchi. Two mathematical models of knowledge distillation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, pp. 4726–4734, 2026.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, 2024.

Ziqi Zhang, Ali Shahin Shamsabadi, Hanxiao Lu, Yifeng Cai, and Hamed Haddadi. Membership and memorization in LLM knowledge distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20074–20084, 2025.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

## Appendix Contents


**A Real-World Examples of Knowledge Distillation** **14**

**B Theoretical Analysis** **15**

B.1 Proof of Theorem 1 . . . . . 15

B.2 Proof of Theorem 2 . . . . . 17

**C Experiment Setup** **18**

C.1 Benchmarks and Models. . . . . 18

C.2 Illustrative Examples . . . . . 19

C.2.1 ASQA Example . . . . . 19

C.2.2 ScienceQA Example . . . . . 21

C.3 Attribution Baselines . . . . . 24

C.4 Experiment Settings. . . . . 25

**D Supplementary Results** **27**

D.1 Task Utility . . . . . 27


## A Real-World Examples of Knowledge Distillation

Model-to-model knowledge transfer is already present in contemporary LLM development. Beyond controlled academic studies, several model developers have explicitly released distilled models or production workflows that train smaller models using outputs generated by more capable LLMs. In addition, frontier model providers have reported large-scale cross-provider distillation activity. We summarize several representative examples below.

**DeepSeek-R1-Distill.** DeepSeek provides a direct example of model-to-model distillation in deployed LLM development. DeepSeek-R1 generated reasoning data that were used to fine-tune six smaller models based on Qwen2.5 and Llama3, including 1.5B, 7B, 8B, 14B, 32B, and 70B variants (DeepSeek-AI, 2025). These models are publicly released as DeepSeek-R1-Distill models, demonstrating that outputs from one LLM can serve directly as supervision for models originating from other model families.

**OpenAI Model Distillation.** OpenAI has also released an integrated Model Distillation workflow that allows developers to collect outputs from more capable models, such as GPT-4o and o1-preview, and use them to fine-tune smaller, more cost-efficient models such as GPT-4o mini (OpenAI, 2024). This provides a production-level example in which outputs from one LLM are explicitly reused as training supervision for another model.

**Reported cross-provider distillation.** Recent reports further suggest that model-to-model knowledge transfer also occurs across model providers. In February 2026, Anthropic reported detecting large-scale campaigns attributed to DeepSeek, Moonshot, and MiniMax that collected Claude outputs for training their own models (Anthropic, 2026a). A subsequent Anthropic report in September 2026 described additional campaigns attributed to several model developers, including Alibaba, DeepSeek, Zhipu, Xiaomi, and MiniMax, involving responses or reasoning traces (Anthropic, 2026b). OpenAI has similarly reported adversarial distillation attempts targeting its frontier models, including activity associated with DeepSeek (OpenAI, 2026b). These latter cases are provider reports rather than independently verified training disclosures, but they further illustrate the practical relevance of knowledge transfer between independently developed LLMs.

Table 5: Representative real-world examples of model-to-model knowledge distillation. "Reported" cases are claims disclosed by the corresponding model provider and should not be interpreted as independently verified training disclosures.

| Example | Knowledge source | Target / use |
|---|---|---|
| DeepSeek-R1-Distill | DeepSeek-R1 | Qwen2.5 / Llama3 models |
| OpenAI Model Distillation | GPT-4o / o1-preview | Smaller OpenAI models |
| Anthropic report (Feb. 2026) | Claude | DeepSeek / Moonshot / MiniMax |
| Anthropic report (Sep. 2026) | Claude | Multiple external LLM developers |
| OpenAI report (Feb. 2026) | OpenAI frontier models | Reported external distillation activity |

# B Theoretical Analysis

## B.1 Proof of Theorem 1

We first establish several auxiliary results used in the proof. Throughout this section, let $p_t(g \mid X) := p_{\theta_t}(g \mid X)$ and $p_\theta(g \mid X) := p_\theta(g \mid X)$.

**Lemma 1.** *At the teacher parameters $\theta_t$,*

$$\begin{aligned} \nabla \mathcal{L}_{\mathrm{KD}}(\theta_t) = 0, \quad \nabla^2 \mathcal{L}_{\mathrm{KD}}(\theta_t) = F, \\ \nabla \mathcal{I}(\theta_t) = 0, \qquad \nabla^2 \mathcal{I}(\theta_t) = \frac{1}{4} F. \end{aligned} \tag{12}$$

*Proof.* From Eq. 2, the KD objective can be written as

$$\mathcal{L}_{\mathrm{KD}}(\theta) = \mathbb{E}_X \sum_g p_t(g \mid X) \log \frac{p_t(g \mid X)}{p_\theta(g \mid X)}.$$

The first term is independent of $\theta$, and hence $\nabla \mathcal{L}_{\mathrm{KD}}(\theta) = -\mathbb{E}_X \sum_g p_t \nabla p_\theta / p_\theta$. Evaluating at $\theta_t$ and using $\sum_g p_t(g \mid X) = 1$ gives $\nabla \mathcal{L}_{\mathrm{KD}}(\theta_t) = 0$. Differentiating once more yields

$$\nabla^2 \mathcal{L}_{\mathrm{KD}}(\theta_t) = -\mathbb{E}_X \sum_g \nabla^2 p_t + \mathbb{E}_X \sum_g \frac{\nabla p_t \nabla p_t^\top}{p_t} = F,$$

where the first term again vanishes by normalization.

Under the equal source prior $P(Y = t) = P(Y = s) = 1/2$, the intrinsic source information is the conditional Jensen–Shannon divergence

$$\mathcal{I}(\theta) = \mathbb{E}_X D_{\mathrm{JS}} \left( p_{\theta_t}(\cdot \mid X), p_\theta(\cdot \mid X) \right).$$

Equivalently,

$$\mathcal{I}(\theta) = \frac{1}{2} \mathbb{E}_X \sum_g p_t \log \frac{2p_t}{p_t + p_\theta} + \frac{1}{2} \mathbb{E}_X \sum_g p_\theta \log \frac{2p_\theta}{p_t + p_\theta}.$$

Differentiating gives $\nabla \mathcal{I}(\theta) = \frac{1}{2} \mathbb{E}_X \sum_g \nabla p_\theta \log \frac{2p_\theta}{p_t + p_\theta}$, and therefore $\nabla \mathcal{I}(\theta_t) = 0$. A second differentiation at $\theta_t$ gives

$$\nabla^2 \mathcal{I}(\theta_t) = \frac{1}{4} \mathbb{E}_X \sum_g \frac{\nabla p_t \nabla p_t^\top}{p_t} = \frac{1}{4} F.$$

This proves Eq. 12. □

**Lemma 2.** *Let $\Delta := \theta - \theta_t$ and $d := \|\Delta\|_2$. For any $\theta \in B(\theta_t, r_0)$,*

$$\begin{aligned} \nabla \mathcal{L}_{\mathrm{KD}}(\theta) = F\Delta + r_K(\Delta), \quad \|r_K(\Delta)\|_2 \le \frac{L_K}{2} d^2, \\ \nabla \mathcal{I}(\theta) = \frac{1}{4} F\Delta + r_I(\Delta), \quad \|r_I(\Delta)\|_2 \le \frac{L_I}{2} d^2. \end{aligned} \tag{13}$$

*Proof.* By Lemma 1 and the fundamental theorem of calculus,

$$\nabla \mathcal{L}_{\mathrm{KD}}(\theta) = F\Delta + \int_0^1 \left[\nabla^2 \mathcal{L}_{\mathrm{KD}}(\theta_t + u\Delta) - F\right] \Delta \, du.$$

The integral defines $r_K(\Delta)$. Assumption 1 gives $\|\nabla^2 \mathcal{L}_{\mathrm{KD}}(\theta_t + u\Delta) - F\|_2 \le L_K u d$, and hence $\|r_K(\Delta)\|_2 \le L_K d^2/2$. Applying the same argument to $\mathcal{I}$, using $\nabla^2 \mathcal{I}(\theta_t) = F/4$, gives the second relation in Eq. 13. □

**Lemma 3.** *Let $u \neq 0$ and let $a, b$ be vectors satisfying $\|a\|_2 + \|b\|_2 < \|u\|_2$. Then*

$$(u+a)^\top (u+b) \ge \|u\|_2 \left(\|u\|_2 - \|a\|_2 - \|b\|_2\right) > 0. \tag{14}$$

*Proof.* Let $s := \|a\|_2 + \|b\|_2$. Using the polarization identity,

$$4(u+a)^\top (u+b) = \|2u + a + b\|_2^2 - \|a - b\|_2^2.$$

Since $s < \|u\|_2$, the triangle inequality gives $\|2u + a + b\|_2 \ge 2\|u\|_2 - s$ and $\|a - b\|_2 \le s$. Therefore,

$$4(u+a)^\top (u+b) \ge (2\|u\|_2 - s)^2 - s^2 = 4\|u\|_2(\|u\|_2 - s),$$

which proves Eq. 14. □

*Proof of Theorem 1.* Let

$$r_c := \min\left\{r_0, \, \frac{2\mu}{L_K + 4L_I}\right\}.$$

Consider the continuous KD gradient flow

$$\frac{d\theta(\tau)}{d\tau} = -\nabla \mathcal{L}_{\mathrm{KD}}(\theta(\tau)), \qquad \theta(0) = \theta_s. \tag{15}$$

We first establish that $\mathcal{I}$ decreases whenever $0 < \|\theta - \theta_t\|_2 < r_c$. Let $\Delta = \theta - \theta_t$, $d = \|\Delta\|_2$, and set

$$u := F\Delta, \qquad a := 4r_I(\Delta), \qquad b := r_K(\Delta).$$

By Lemma 2,

$$4\left\langle \nabla \mathcal{I}(\theta), \nabla \mathcal{L}_{\mathrm{KD}}(\theta)\right\rangle = (u+a)^\top (u+b). \tag{16}$$

Assumption 2 and Cauchy–Schwarz imply $\|F\Delta\|_2 \ge \mu d$ for $d > 0$. Moreover, Lemma 2 gives

$$\|a\|_2 + \|b\|_2 \le 2L_I d^2 + \frac{L_K}{2} d^2 = \frac{L_K + 4L_I}{2} d^2.$$

Hence, whenever $d < 2\mu/(L_K + 4L_I)$,

$$\|a\|_2 + \|b\|_2 < \mu d \le \|u\|_2.$$

Lemma 3 applied to Eq. 16 therefore yields

$$\left\langle \nabla \mathcal{I}(\theta), \nabla \mathcal{L}_{\mathrm{KD}}(\theta)\right\rangle > 0. \tag{17}$$

Combining Eq. 17 with Eq. 15,

$$\frac{d}{d\tau} \mathcal{I}(\theta(\tau)) = -\left\langle \nabla \mathcal{I}(\theta(\tau)), \nabla \mathcal{L}_{\mathrm{KD}}(\theta(\tau))\right\rangle < 0 \tag{18}$$

whenever $0 < \|\theta(\tau) - \theta_t\|_2 < r_c$.

It remains to verify that the KD trajectory stays in this region. Consider $V(\theta) := \frac{1}{2}\|\theta - \theta_t\|_2^2$. Using Lemma 2,

$$\begin{aligned}
\frac{dV}{d\tau} &= -\Delta^\top \nabla \mathcal{L}_{\mathrm{KD}}(\theta) \\
&= -\Delta^\top F\Delta - \Delta^\top r_K(\Delta) \\
&\le -\mu d^2 + \frac{L_K}{2} d^3 = -d^2\left(\mu - \frac{L_K}{2} d\right).
\end{aligned}$$

Because $r_c \le 2\mu/(L_K + 4L_I) < 2\mu/L_K$, we have $dV/d\tau < 0$ throughout the contraction region whenever $d > 0$.

Choose $r$ such that $\|\theta_s - \theta_t\|_2 < r < r_c$. If the trajectory first reached $\|\theta(\tau) - \theta_t\|_2 = r$ at some finite $\tau^\star > 0$, then $V$ would be strictly decreasing on $[0, \tau^\star]$, which would imply $\|\theta(\tau^\star) - \theta_t\|_2 < \|\theta_s - \theta_t\|_2 < r$, a contradiction. Thus the trajectory remains inside $B(\theta_t, r_c)$.

Finally, Assumption 1 implies that $\nabla\mathcal{L}_{\mathrm{KD}}$ is locally Lipschitz. Since $\nabla\mathcal{L}_{\mathrm{KD}}(\theta_t) = 0$, uniqueness of the gradient-flow solution implies that a trajectory initialized at $\theta_s \neq \theta_t$ cannot reach $\theta_t$ in finite time. Hence Eq. 18 holds for every finite $\tau \ge 0$.

For any $0 \le \tau_1 < \tau_2$,

$$\mathcal{I}(\theta(\tau_2)) - \mathcal{I}(\theta(\tau_1)) = \int_{\tau_1}^{\tau_2} \frac{d}{d\tau}\mathcal{I}(\theta(\tau))\, d\tau < 0.$$

Therefore,

$$\mathcal{I}(\theta(\tau_2)) < \mathcal{I}(\theta(\tau_1)). \tag{19}$$

Taking $\theta'_s = \theta(T)$ for any $T > 0$ gives $\mathcal{I}(\theta'_s) < \mathcal{I}(\theta_s)$, completing the proof. □

### B.2 Proof of Theorem 2

Throughout this section, all expectations and variances are evaluated over the prompt distribution and $G \sim p_{\theta_t}(\cdot \mid X)$. For compactness, write

$$p_t(g \mid X) := p_{\theta_t}(g \mid X), \qquad p_\epsilon(g \mid X) := p_{\theta_t + \epsilon v}(g \mid X).$$

Recall from Eq. 10 that

$$h(X, G) = \frac{d}{d\epsilon} \log p_\epsilon(G \mid X)\Big|_{\epsilon=0}.$$

**Lemma 4.** *As $\epsilon \to 0$,*

$$\mathcal{I}(\theta_t + \epsilon v) = \frac{\epsilon^2}{8}\,\mathrm{Var}(h) + o(\epsilon^2). \tag{20}$$

*Proof.* Since $\sum_g p_\epsilon(g \mid X) = 1$ for every $\epsilon$,

$$\mathbb{E}[h \mid X] = \frac{d}{d\epsilon} \sum_g p_\epsilon(g \mid X)\Big|_{\epsilon=0} = 0.$$

Hence $\mathbb{E}[h] = 0$ and $\mathrm{Var}(h) = \mathbb{E}[h^2]$.

From the local geometry established in Lemma 1,

$$\nabla\mathcal{I}(\theta_t) = 0, \qquad \nabla^2\mathcal{I}(\theta_t) = \frac{1}{4}F.$$

A second-order expansion around $\theta_t$ therefore gives

$$\mathcal{I}(\theta_t + \epsilon v) = \frac{\epsilon^2}{8} v^\top F v + o(\epsilon^2).$$

By the definition of the Fisher information matrix,

$$v^\top F v = \mathbb{E}\left[\left(v^\top \nabla_\theta \log p_\theta(G \mid X)\big|_{\theta=\theta_t}\right)^2\right] = \mathbb{E}[h^2] = \mathrm{Var}(h).$$

Substituting this identity yields Eq. 20. □

**Lemma 5.** *Let*

$$q_{\phi,\epsilon}(z \mid X) := \Pr_{\theta_t + \epsilon v}(Z_\phi = z \mid X)$$

*denote the distribution induced by the fixed attribution model $f_\phi$. Then, as $\epsilon \to 0$,*

$$\mathcal{I}_\phi(\theta_t + \epsilon v) = \frac{\epsilon^2}{8}\,\mathrm{Var}\left(\mathbb{E}[h \mid X, Z_\phi]\right) + o(\epsilon^2). \tag{21}$$

*Proof.* Because $Z_\phi = f_\phi(G)$ is deterministic,

$$q_{\phi,\epsilon}(z \mid X) = \sum_{g: f_\phi(g)=z} p_\epsilon(g \mid X).$$

Differentiating at $\epsilon = 0$ gives

$$\frac{d}{d\epsilon} q_{\phi,\epsilon}(z \mid X)\Big|_{\epsilon=0} = \sum_{g: f_\phi(g)=z} p_t(g \mid X) h(X, g).$$

Therefore, for every $z$ with $q_{\phi,0}(z \mid X) > 0$,

$$\frac{d}{d\epsilon} \log q_{\phi,\epsilon}(z \mid X)\Big|_{\epsilon=0} = \mathbb{E}[h \mid X, Z_\phi = z]. \tag{22}$$

Thus, the local source evidence preserved in the attribution output is the conditional expectation $\mathbb{E}[h \mid X, Z_\phi]$.

Under equal source priors,

$$\mathcal{I}_\phi(\theta_t + \epsilon v) = \mathbb{E}_X D_{\mathrm{JS}}\left(q_{\phi,0}(\cdot \mid X), q_{\phi,\epsilon}(\cdot \mid X)\right).$$

Applying the same local Jensen–Shannon expansion as in Lemma 4, with the induced score in Eq. 22, yields

$$\mathcal{I}_\phi(\theta_t + \epsilon v) = \frac{\epsilon^2}{8} \mathbb{E}\left[\left(\mathbb{E}[h \mid X, Z_\phi]\right)^2\right] + o(\epsilon^2).$$

Since $\mathbb{E}[h] = 0$, the tower property gives $\mathbb{E}[\mathbb{E}[h \mid X, Z_\phi]] = 0$. Hence

$$\mathbb{E}\left[\left(\mathbb{E}[h \mid X, Z_\phi]\right)^2\right] = \mathrm{Var}\left(\mathbb{E}[h \mid X, Z_\phi]\right),$$

which proves Eq. 21. □

*Proof of Theorem 2.* By the law of total variance,

$$\mathrm{Var}(h) = \mathrm{Var}\left(\mathbb{E}[h \mid X, Z_\phi]\right) + \mathbb{E}\left[\mathrm{Var}(h \mid X, Z_\phi)\right]. \tag{23}$$

Combining Lemmas 4 and 5, and using $0 < \mathrm{Var}(h) < \infty$, gives

$$\lim_{\epsilon \to 0} \frac{\mathcal{I}_\phi(\theta_t + \epsilon v)}{\mathcal{I}(\theta_t + \epsilon v)} = \frac{\mathrm{Var}\left(\mathbb{E}[h \mid X, Z_\phi]\right)}{\mathrm{Var}(h)}.$$

Applying Eq. 23 yields

$$\lim_{\epsilon \to 0} \frac{\mathcal{I}_\phi(\theta_t + \epsilon v)}{\mathcal{I}(\theta_t + \epsilon v)} = 1 - \frac{\mathbb{E}\left[\mathrm{Var}(h \mid X, Z_\phi)\right]}{\mathrm{Var}(h)}. \tag{24}$$

This is precisely Eq. 11, completing the proof. □

# C Experiment Setup

## C.1 Benchmarks and Models.

We conduct experiments on two benchmarks, ASQA (Stelmakh et al., 2022) and ScienceQA (Lu et al., 2022), covering text-only question answering and image-conditioned science reasoning, respectively. ASQA contains ambiguous factoid questions whose answers must account for multiple plausible interpretations. We use its 4,353 official training questions as the training partition and its 948 development questions as the held-out evaluation partition. For the human-source class, we use the first provided human long-answer annotation for each question.

For ScienceQA, we initially retain questions containing both an image and non-empty textual context, yielding 6,532 questions. For response generation and knowledge distillation, we combine the retained official training and test sets into a training pool of 5,201 questions and reserve the 1,331

retained validation questions for held-out evaluation. For nine-class attribution, we additionally require a non-empty official solution so that human references include an explanation rather than only an answer option. We exclude questions without an official solution from all nine source classes, removing 672 training questions and 157 evaluation questions. This yields an attribution training partition of 4,529 questions and an evaluation partition of 1,174 questions. Each human-source response consists of the correct option letter and option text, followed by the unmodified official solution. This additional filtering applies to attribution training and evaluation; KD retains the original 5,201-question training pool and uses teacher-generated responses as supervision. Within each benchmark, all nine source classes share identical question coverage in each attribution partition. Pre-distillation, post-distillation, and post-retraining attribution evaluations use the same held-out questions, which are excluded from both KD training and attribution-model training.

For each benchmark, the model pool contains six open-weight models from three model families and two API-based models. On ASQA, we use Qwen2.5-32B-Instruct and Qwen2.5-7B-Instruct (Qwen et al., 2025), Gemma-2-27B-IT and Gemma-2-9B-IT (Gemma Team et al., 2024), and Mistral-Small-24B-Instruct-2501 and Mistral-7B-Instruct-v0.3 (Mistral AI Team, 2025; Jiang et al., 2023). On SCIENCEQA, we use Qwen3-VL-32B-Instruct and Qwen3-VL-8B-Instruct (Bai et al., 2025), Gemma-3-27B-IT and Gemma-3-12B-IT (Gemma Team et al., 2025), and InternVL3.5-38B-Instruct and InternVL3.5-8B-Instruct (Wang et al., 2025). Both benchmarks additionally include responses from GPT-6 Astra (OpenAI, 2026a) and MiniMax-M3 (Lai et al., 2026), accessed through hosted APIs. For the human source, we use the first provided long-answer annotation for each ASQA question. For SCIENCEQA, we concatenate the gold option letter and answer text with the unmodified official solution.

Within each benchmark, the six open-weight models and two closed-sourced models use the same user prompt, with model-specific chat templates or API wrappers applied externally. For ASQA, the input is simply the original ambiguous question,

$$\{\texttt{ambiguous_question}\}.$$

For SCIENCEQA, the model receives the question image together with the following textual prompt:

```
Use the image and the text context to answer the
multiple-choice science question.
Text context: {hint}
Question: {question}
Choices:
(A) {choice_1}
(B) {choice_2}
(C) {choice_3}
...
Give the best answer and a brief explanation.
```

The available choices are expanded line by line according to each question. Open-weight models use 4-bit NF4 quantization with temperature 0.7, top-$p = 1.0$, and at most 256 newly generated tokens; the same decoding settings are retained for distilled students. Both API-based models, GPT-6 Astra and MiniMax-M3, are queried with a requested completion budget of 256 tokens. GPT-6 Astra uses low reasoning effort, without explicitly specified temperature or top-$p$ parameters. MiniMax-M3 is queried with thinking disabled, temperature 0.7, and top-$p = 1.0$. Attribution uses only the returned final response text, excluding any separately returned reasoning content.

### C.2 ILLUSTRATIVE EXAMPLES

#### C.2.1 ASQA EXAMPLE

Table 6: Illustrative ASQA example.

| Source | Content |
|---|---|
| **Input** | When was the first helicopter made and flown? |

*Continued on next page*

*Table 6 continued*

| Source | Content |
|---|---|
| **Human** | The first helicopter was created in 1861 by the French inventor, Gustave de Ponton d'Amécourt. However, his model never took flight. The Italian engineer, Enrico Forlanini, build an unmanned version of the helicopter in 1878 that rose to a height of 12 meters (39 feet), where it hovered for about 20 seconds after a vertical take-off. The first helicopter model to carry a pilot was in 1907. Two French brothers, Jacques and Louis Breguet, built the "Gyroplane No.1" sometime between 14 August and 29 September 1907. It lifted its pilot 2 ft into the air for about a minute. The Gyroplane No.1 proved to be extremely unsteady and required a man at each corner of the airframe to hold it steady. On 13 November 1907, another French inventor, Paul Cornu, designed and built the Cornu helicopter. It lifted its inventor off the ground and remained aloft for 20 seconds. Even though this flight did not surpass in height the flight of the Gyroplane No. 1, it was reported to be the first truly free flight with a pilot as it did not need to be held steady. Almost three decades later, German aviation pioneer and co-founder of the Focke-Wulf company, invented the Fw 61. This was the world's first practical transverse twin-rotor helicopter. It first flew on 26 June 1936 to more than 8,000 feet (2,400 m) at speeds of 120 miles per hour (190 km/h). A few years later, on 11 December 1951, turbine engines started being used in helicopters. This meant the helicopter could fly with a large amount of power and a low weight penalty. |
| **Qwen2.5-32B** | The first practical helicopter is generally attributed to Igor Sikorsky, who flew his VS-300 on May 13, 1940, in the United States. However, if we consider earlier experimental designs, the French inventor Étienne Oehmichen conducted successful helicopter flights earlier. He set the first world record for helicopters on April 14, 1924, flying a distance of one kilometer.<br>But it's worth noting that earlier attempts at vertical flight were made throughout the 19th and early 20th centuries. For instance, in 1907, Paul Cornu, also from France, managed brief hovering flights with his helicopter design. The development of the helicopter involved many pioneers and iterations before it reached practicality and widespread use. |
| **Qwen2.5-7B** | The first successful powered flight of a helicopter occurred on July 26, 1909 in France. This historic event was achieved by French aviation pioneer Louis Breguet with his Breguet No. 15 helicopter. However, it's important to note that several other prototypes existed before this, and there were attempts to create helicopters in different countries around that time.<br>The story of early helicopter development begins much earlier, with various designs and experiments from the 17th century onwards. Key figures in the history of helicopter design include Leonardo da Vinci, who sketched designs for flying machines in the late 15th century, although these were not built during his lifetime.<br>The Breguet No. 15 was one of the earliest successful designs. It had three rotors and was powered by a two-cylinder engine producing about 30 horsepower. The helicopter managed to lift off the ground, hover briefly, and land safely, marking a significant milestone in the history of aviation. |

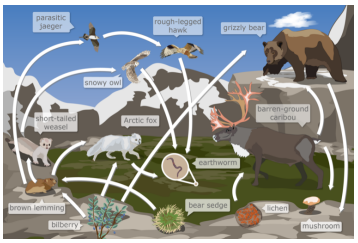
*Continued on next page*

*Table 6 continued*

| Source | Content |
|---|---|
| **Gemma-2-27B** | The first successful helicopter flight was achieved by **Paul Cornu on 19th November 1907** in France.<br>It's important to note that this wasn't the first attempt to build a helicopter. Many people had tried and failed before Cornu. **However, his invention is recognized as the "first" because his was the first documented instance of a *free flight* being achieved by a heavier-than-air craft using a rotor system.**<br>This means his machine actually lifted off the ground and flew for a short distance, powered solely by its rotors, unlike earlier attempts that only achieved short hops or gliding.<br>Cornu's helicopter was a **small, unmanned craft** with a **single rotor** and a set of **counterweights** to counteract the torque and achieve a stable hover.<br>While it only achieved a **short, uncontrolled flight** of about a minute, it was a milestone achievement that proved a helicopter with a single rotor could create lift and achieve a stable hover, albeit briefly.<br>This achievement paved the way for further development and innovation in the field of helicopter design. |
| **Gemma-2-9B** | The first successful helicopter flight was made by **Igor Sikorsky** on **September 14, 1939**. This was with his **VS-300** model.<br>While there were earlier attempts at creating helicopters, Sikorsky's VS-300 was the first to achieve controlled, sustained flight. |
| **Mistral-24B** | The first successful helicopter, often credited for its vertical takeoff and landing capabilities, was the Focke-Wulf Fw 61. It was designed by German engineer Henrich Focke and flew for the first time on June 26, 1936. This helicopter was notable for its innovative design and demonstrated the feasibility of vertical flight. However, it's important to note that there were earlier attempts and designs, such as the work by Igor Sikorsky, which led to the first practical helicopter, the VS-300, which flew in 1939. |
| **Mistral-7B** | The first helicopter that was successfully flown and controlled in a sustained hover was the Focke-Wulf Fw 61 "V1" (Flugelrad-Wirbelwind), designed by Heinrich Focke and Barnim Diekwolf in Germany. It was first flown on June 22, 1936. The helicopter, also known as the "Focke-Achtrotor," was powered by eight small engines and had a rotor with eight rotor blades. However, the first powered helicopter with a single rotor, the Cornu helicopter, was built by Emile Cornu in France in 1907, but it was not capable of sustained flight. |
| **GPT-6 Astra** | The first widely recognized **successful helicopter flight** took place on **November 13, 1907**, near Paris, France. Paul Cornu flew his twin-rotor helicopter about **30 cm (1 ft)** above the ground for roughly **20 seconds**. It is often credited as the first untethered, piloted helicopter flight.<br>However, it was not fully controllable. The first practical, controllable helicopter was the **Focke-Wulf Fw 61**, first flown in Germany on **June 26, 1936**. Igor Sikorsky's **VS-300**, first flown in the United States on **September 14, 1939**, established the single-main-rotor and tail-rotor configuration used by most modern helicopters. |
| **MiniMax-M3** | The first helicopter to achieve a true free flight was the Focke-Wulf Fw 61, which first flew on June 26, 1936, piloted by Ewald Rohlfs in Bremen, Germany. It was the first practical, fully controllable helicopter.<br>However, the first helicopter to successfully fly, though only briefly and tethered, was the Cornu helicopter, built by Paul Cornu, which lifted off the ground on November 13, 1907, in Lisieux, France. Igor Sikorsky's VS-300, which first flew on September 14, 1939, became the first mass-produced helicopter.<br>The Fw 61 is widely credited as the first fully functional helicopter, while the Cornu helicopter holds the distinction of being the first to lift off. |

### C.2.2 SCIENCEQA EXAMPLE

Table 7: Illustrative example from ScienceQA.

| Source | Content |
|---|---|
| **Image** |   |
| **Context** | Below is a food web from a tundra ecosystem in Nunavut, a territory in Northern Canada.<br>A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem. |
| **Question** | Which of these organisms contains matter that was once part of the bear sedge? |
| **Choices** | (A) grizzly bear (B) bilberry (C) short-tailed weasel |
| **Human** | (C) short-tailed weasel<br>Use the arrows to follow how matter moves through this food web. For each answer choice, try to find a path of arrows that starts from the bear sedge.<br>The grizzly bear has two arrows pointing to it. One arrow starts from the bilberry. The bilberry does not have any arrows pointing to it. The other arrow pointing to the grizzly bear starts from the barren-ground caribou. The only arrow pointing to the barren-ground caribou starts from the lichen. The lichen does not have any arrows pointing to it. So, in this food web, matter does not move from the bear sedge to the grizzly bear.<br>The bilberry does not have any arrows pointing to it. So, in this food web, matter does not move from the bear sedge to the bilberry.There is one path matter can take from the bear sedge to the short-tailed weasel: bear sedge->brown lemming->short-tailed weasel. There is one path matter can take from the bear sedge to the Arctic fox: bear sedge->brown lemming->Arctic fox. There is one path matter can take from the bear sedge to the rough-legged hawk: bear sedge->brown lemming->parasitic jaeger->rough-legged hawk. |
| **Qwen3-VL-32B** | **Answer: (A) grizzly bear**<br>**Explanation:** In the food web, there is an arrow pointing from **bear sedge** to **grizzly bear**, which means the grizzly bear consumes bear sedge. Since the arrows represent the flow of matter through the ecosystem, the matter from the bear sedge moves into the grizzly bear when it eats it. Therefore, the grizzly bear contains matter that was once part of the bear sedge.<br>The other choices:<br>• **Bilberry** is consumed by the brown lemming, but there's no direct or indirect pathway from bear sedge to bilberry.<br>• **Short-tailed weasel** eats the brown lemming, but again, no connection exists between bear sedge and the short-tailed weasel in this food web.<br>Thus, the correct answer is **(A) grizzly bear**. |
| **Qwen3-VL-8B** | The best answer is: **(C) short-tailed weasel**<br>**Explanation:**<br>In a food web, arrows indicate the direction of matter flow—from the organism being eaten to the organism that eats it. Looking at the diagram, the bear sedge is eaten by the brown lemming (arrow points from bear sedge → brown lemming). The brown lemming is then eaten by the short-tailed weasel (arrow points from brown lemming → short-tailed weasel). |

*Continued on next page*

*Table 7 continued*

| Source | Content |
|---|---|
| | Therefore, matter from the bear sedge flows through the brown lemming and then into the short-tailed weasel. The grizzly bear does not eat bear sedge directly (no arrow from bear sedge to grizzly bear), and bilberry is a plant that is not shown to be eaten by the bear sedge.<br>So, the short-tailed weasel is the organism that contains matter originally from the bear sedge, via the brown lemming. |
| **Gemma-3-27B** | Here's the best answer and explanation:<br>**Answer:** (C) short-tailed weasel<br>**Explanation:**<br>1. **Tracing the Arrows:** The arrows in a food web show the flow of matter (energy/nutrients). We need to find a path that starts with bear sedge and leads to the short-tailed weasel.<br>2. **Following the Path:**<br>• Bear sedge is eaten by the brown lemming.<br>• The brown lemming is eaten by the short-tailed weasel.<br>3. **Therefore:** Matter from the bear sedge has been transferred to the brown lemming and then to the short-tailed weasel.<br>Let's look at why the other options are incorrect:<br>• **(A) grizzly bear:** The grizzly bear eats barren-ground caribou, not bear sedge.<br>• **(B) bilberry:** Bilberry does not directly consume bear sedge. |
| **Gemma-3-12B** | **The best answer is (A) grizzly bear.**<br>Here's the explanation:<br>• **Food webs and matter flow:** Food webs show how energy and matter move from one organism to another. The arrows indicate the direction of this flow.<br>• **Bear sedge and its connections:** The image shows that the grizzly bear has an arrow pointing *from* the bear sedge *to* the grizzly bear. This means the grizzly bear consumes the bear sedge, so it contains matter that was originally part of the bear sedge.<br>Let's look at why the other options are incorrect:<br>• **(B) bilberry:** The bilberry is a producer, it doesn't get matter *from* the bear sedge.<br>• **(C) short-tailed weasel:** The short-tailed weasel consumes the brown lemming, which eats the bilberry. It doesn't directly consume the bear sedge.<br>I hope this helps! |
| **InternVL3.5-38B** | (C) short-tailed weasel<br>In the given text context, the food web is described as a model that shows how matter moves through an ecosystem. The matter from the bear sedge would be transferred to organisms that feed on it directly or indirectly. Among the provided choices, the short-tailed weasel is most likely to have consumed matter that was once part of the bear sedge, either by eating plants directly or by consuming other animals that had eaten the sedge. The grizzly bear and bilberry do not fit the pattern of consuming matter from bear sedge based on the provided image and context. |
| **InternVL3.5-8B** | (A) grizzly bear<br>The grizzly bear contains matter that was once part of the bear sedge. In the food web, the arrow from the bear sedge points directly to the grizzly bear, indicating that the bear consumes plant material, including bear sedge. The other choices (bilberry and short-tailed weasel) are not connected to the bear sedge in the food web. |
| **GPT-6 Astra** | **(C) short-tailed weasel.** The brown lemming eats bear sedge, and the short-tailed weasel eats the lemming. Matter moves along this path: bear sedge → brown lemming → short-tailed weasel. |
| **MiniMax-M3** | **Answer: (A) grizzly bear**<br>**Explanation:** Following the arrows in the food web, matter flows from bear sedge to the barren-ground caribou (which eats the bear sedge), and then from the barren-ground caribou to the grizzly bear (which eats the caribou). |

*Continued on next page*

*Table 7 continued*

| Source | Content |
|---|---|
| | • Bilberry is a plant (producer) and does not consume bear sedge. |
| | • Short-tailed weasels eat brown lemmings, not bear sedge or any organism that consumes it. |
| | Therefore, only the grizzly bear contains matter that was once part of the bear sedge. |

### C.3 Attribution Baselines

We consider seven representative baselines covering supervised classification, contrastive representation learning, linguistic distribution modeling, and context-aware inference patterns. All seven methods perform attribution over the same nine source classes using response text as input, without requiring access to the parameters or intermediate activations of the source models. For BERT-MLP, WhosAI, DeTeCtive, mdok, and OTBDetector, inputs are truncated to 512 tokens under the respective tokenizers. Method-specific configurations are detailed below.

**BERT-MLP** (Devlin et al., 2019). We use a frozen `bert-base-uncased` encoder and extract the final-layer `[CLS]` representation of each response. A multilayer perceptron with one 256-unit hidden layer and a multiclass output layer is trained over the source label space. Only the classification head is trained; the BERT encoder remains fixed. This provides a standard supervised baseline based on pretrained text representations.

**WhosAI** (Cava et al., 2024). WhosAI learns source-discriminative representations through triplet-based contrastive learning. We use `bert-base-uncased` with attention-mask-aware mean pooling and $\ell_2$ normalization. Training uses cosine-distance triplet loss with a dynamic margin initialized at $0.10$ and increased by $0.025$ every 750 optimization steps. Informative triplets are selected using multi-similarity mining with $\epsilon = 0.10$. At inference time, each response is assigned to the source whose normalized training-set centroid has the highest cosine similarity.

**DeTeCtive** (Guo et al., 2024). DeTeCtive organizes contrastive supervision according to source identity, model family, and the distinction between human and machine-generated text. We use the `unsup-simcse-roberta-base` encoder with mean pooling and $\ell_2$ normalization. The contrastive temperature is set to $\tau = 0.07$, and the model-level, family-level, machine-level, and human-level loss components receive weights of 1, 1, 1, and 3, respectively. An auxiliary source-classification cross-entropy loss is added with weight 1. Final attribution uses cosine $k$-nearest-neighbor classification with $k = 5$ over the training embeddings.

**mdok** (Macko, 2025). We implement mdok as supervised sequence classification using `Qwen3-4B-Base`. The backbone is adapted with 4-bit NF4 quantization and QLoRA, using adapter rank $r = 64$, scaling parameter $\alpha_{\text{LoRA}} = 16$, and dropout $0.10$. The sequence-classification head predicts the source identity directly. Training uses cross-entropy with class weights inversely proportional to training-set source frequencies; these weights are equal when the source classes are balanced.

**PROFILER** (Guo et al., 2025). PROFILER characterizes responses through context-aware loss patterns computed by a frozen surrogate language model. We use `Falcon-7B` as the surrogate and a context width of $W = 6$, corresponding to token offsets $\{-2, -1, 0, 1, 2, 3\}$. For each offset, we extract the mean, maximum, minimum, standard deviation, and median of both the context-loss sequence and its residual sequence. Together with pairwise symmetric KL features between context-loss sequences, this produces 75 features per response. Attribution uses one-versus-rest random forests with 200 trees and maximum depth 7, selecting the source with the highest positive-class score. Responses are clipped to 4,000 characters before feature extraction.

**OTBDetector** (La Cava & Tagarelli, 2025). OTBDetector learns source-aware embeddings through triplet-based contrastive learning. We use `longformer-base-4096` with masked mean pooling and $\ell_2$ normalization. The cosine-distance triplet margin is fixed at $0.05$, with multi-similarity mining using $\epsilon = 0.10$. Attribution is performed by selecting the nearest normalized source centroid under cosine similarity. Although the backbone supports longer sequences, our implementation uses a 512-token input limit. Thus, the principal configuration differences from WhosAI are the Longformer backbone and the fixed triplet margin.

**HLD-Detector** (Guo et al., 2026). HLD-Detector combines linguistic distributional signals at the word, part-of-speech, dependency, and semantic levels. For the discrete distributions, we use context order $k = 5$ with recursive backoff and smoothing parameter $\delta = 10^{-6}$. Semantic distributions are estimated using `all-MiniLM-L6-v2` embeddings, four-word contexts, two-word targets, a stride of two words, $M = 4$ nearest neighbors, and temperature $\tau = 0.10$. We adapt the binary formulation to nine-class attribution by training $\binom{9}{2} = 36$ pairwise XGBoost classifiers on the distributional features. Predictions are combined by majority voting, with summed pairwise probabilities used to resolve ties. XGBoost configurations are selected through 60 Optuna trials using a prompt-disjoint 80:20 split of the attribution training partition.

### C.4 Experiment Settings.

**Distillation data and sampling.** Let $T$ and $S$ denote the teacher and the original student, respectively, and let $S'$ denote the distilled student. We construct distillation examples by pairing each selected training question with its teacher-generated response. For ScienceQA, the input additionally includes the original image, textual context, and answer choices. Human reference answers are not used as distillation targets.

The distillation ratio $\alpha$ denotes the fraction of training questions selected for KD, rather than a mixing coefficient between teacher and human supervision. Unless otherwise specified, all experiments use $\alpha = 0.10$, corresponding to 435 questions on ASQA and 520 questions on ScienceQA. All reported results are averaged over three runs with random seeds $7, 22, 41$. For the distillation-ratio analysis, we additionally consider $\alpha \in 0.005, 0.01, 0.025, 0.10, 0.20$. For each seed, the training subsets are nested across ratios, and teacher comparisons under the same dataset and ratio use identical question IDs.

**Student adaptation.** Each KD run starts from the original student checkpoint. We train for five epochs using AdamW with an initial learning rate of $2\times10^{-4}$, linear learning-rate decay, no warmup, and zero weight decay. The micro-batch size is one, with eight gradient accumulation steps. We use 4-bit NF4 quantization and LoRA with rank $r = 16$, scaling parameter $\alpha_{\text{LoRA}} = 32$, and dropout $0.05$. Adapters are applied to the attention projections ($Q$, $K$, $V$, and $O$) and the feed-forward gate, up, and down projections of the language decoder. For multimodal students, the vision encoder and multimodal projector remain frozen. Each KD configuration is independently trained with random seeds $\{7, 22, 41\}$, and the reported results are averaged across the three runs.
**Distribution-level knowledge distillation.** We use Gemma-2-27B-IT as the teacher and Gemma-2-9B-IT as the student on ASQA, and Gemma-3-27B-IT as the teacher and Gemma-3-12B-IT as the student on ScienceQA. Within each pair, the teacher and student use identical token-to-ID mappings and receive the same encoded teacher-forced inputs. For ScienceQA, they additionally share the same encoded image inputs and multimodal preprocessing.

For model $M \in \{S, T\}$, define

$$p_M^{(\tau)}(\cdot \mid c_t) = \text{softmax}(z_M(c_t)/\tau)\,,$$

where $z_M(c_t)$ denotes the next-token logits produced by model $M$ under the teacher-forced context $c_t$. We minimize the forward KL divergence from teacher to student:

$$\mathcal{L}_{\text{dist}} = \frac{\tau^2}{L} \sum_{t=1}^{L} D_{\text{KL}}\left(p_T^{(\tau)}(\cdot \mid c_t) \,\middle\|\, p_S^{(\tau)}(\cdot \mid c_t)\right), \tag{25}$$

where $\tau = 2$ and

$$D_{\text{KL}}(p\|q) = \sum_{v\in\mathcal{V}} p(v) \log \frac{p(v)}{q(v)}.$$

Here, $L$ denotes the number of valid response-prediction positions; prompt and padding positions are excluded.

The KL divergence is computed over the full shared tokenizer vocabulary: $|\mathcal{V}| = 256{,}000$ for the ASQA Gemma-2 pair and $262{,}145$ for the ScienceQA Gemma-3 pair. The Gemma-2 output heads contain exactly $256{,}000$ entries. The Gemma-3 output heads contain $262{,}208$ entries, of which the additional 63 entries beyond the tokenizer vocabulary are removed before softmax normalization.

No top-$k$ approximation is used. The distribution-level loss coefficient is set to one, while the sequence-level and hidden-state loss coefficients are set to zero.

**Hidden-state knowledge distillation.** We align the outputs of the final language-decoder blocks, before the models' final RMSNorm operations, at valid response-prediction positions. Using one-based layer numbering, the aligned student and teacher layers are $(42, 46)$ for the ASQA Gemma-2 pair and $(48, 62)$ for the SCIENCEQA Gemma-3 pair. Only the final decoder-block pair is aligned for each benchmark.

Let $h_S(c_t) \in \mathbb{R}^{d_S}$ and $h_T(c_t) \in \mathbb{R}^{d_T}$ denote the corresponding student and teacher hidden states at the causal position used to predict $y_t^T$. Because the teacher and student have different hidden dimensions, we introduce a trainable bias-free linear projection $W \in \mathbb{R}^{d_T \times d_S}$. We minimize the token-wise cosine distance:

$$\mathcal{L}_{\text{hidden}} = \frac{1}{L} \sum_{t=1}^{L} \left[ 1 - \frac{\langle W h_S(c_t), h_T(c_t) \rangle}{\max(\|W h_S(c_t)\|_2, \epsilon) \max(\|h_T(c_t)\|_2, \epsilon)} \right], \tag{26}$$

where $\epsilon = 10^{-8}$. The projection dimensions are $3584 \to 4608$ for ASQA and $3840 \to 5376$ for SCIENCEQA.

The projection weights are stored in FP32 and initialized using the default linear-layer initialization,

$$W_{ij} \sim \mathcal{U}\left( -\frac{1}{\sqrt{d_S}}, \frac{1}{\sqrt{d_S}} \right).$$

They are optimized jointly with the student LoRA adapters using the same optimizer and learning rate, and are discarded at inference time. The hidden-state loss coefficient is set to one, while the sequence-level and distribution-level loss coefficients are set to zero. The teacher remains frozen throughout training, and its hidden states are detached from gradient computation. As in distribution-level KD, the loss is computed only over valid response-prediction positions.

On SCIENCEQA, the aligned decoder states remain conditioned on the image input. The vision tower and multimodal projector are kept frozen, and no explicit alignment loss is imposed on vision-tower representations or attention maps.

**Shared training settings.** The main comparison uses Gemma-2-27B→Gemma-2-9B on ASQA and Gemma-3-27B→Gemma-3-12B on ScienceQA, with the default distillation ratio $\alpha = 0.10$. Within each dataset, all three KD objectives use the same teacher-supervised questions and saved target responses, corresponding to 435 training questions for ASQA and 520 for ScienceQA. Unless otherwise specified, each experiment is repeated with three random seeds $\{7, 22, 41\}$, and the reported results are averaged across the three runs.

Training is performed for five epochs using AdamW with a learning rate of $2 \times 10^{-4}$, linear learning-rate decay, no warmup, and zero weight decay. The per-device batch size is one with eight gradient accumulation steps. Student base weights are frozen, while LoRA adapters are applied to the query, key, value, output, gate, up, and down projections throughout the language decoder, using rank 16, scaling parameter 32, and dropout 0.05. Maximum training sequence lengths are 1,024 for ASQA and 4,096 for ScienceQA; for ScienceQA, teacher responses are additionally truncated to 256 tokens before chat-template encoding.

For distribution-level and hidden-state KD, both teacher and student are loaded in NF4 4-bit precision with double quantization, using FP16 computation on ASQA and BF16 on ScienceQA. The teacher is kept frozen and in evaluation mode throughout training. Data selection and optimization settings are held fixed across the three KD objectives, with hidden-state KD additionally optimizing the projection matrix $W$. API-based teachers are evaluated only with sequence-level KD because the APIs used do not expose teacher logits or hidden states.

**Attribution and retraining protocol.** All main attribution experiments use nine source classes: six open-weight models, two API-based models, and one human source. Each attribution method is initially trained on responses from the original sources. For the frozen evaluation, this same attribution model is applied to distilled-student responses, which retain the student's source label.

For retraining, we replace the student-class responses in the attribution training partition with responses generated by $S^{\text{KD}}$ on the same training questions. The other eight source classes remain

Table 8: Task utility before and after sequence-level knowledge distillation ($\alpha = 0.10$). We report DR on ASQA and accuracy on SCIENCEQA. Before and Post denote the student performance before and after KD, respectively; $\Delta$ = Post − Before.

| ASQA | | | | ScienceQA | | | |
|---|---|---|---|---|---|---|---|
| **Teacher → Student** | **Before** | **Post** | **Δ** | **Teacher → Student** | **Before** | **Post** | **Δ** |
| Qwen2.5-32B→Qwen2.5-7B | 20.32 | 23.95 | +3.63 | Qwen3-VL-32B→Qwen3-VL-8B | 88.69 | 91.31 | +2.62 |
| Gemma-2-27B→Gemma-2-9B | 24.49 | 26.08 | +1.59 | Gemma-3-27B→Gemma-3-12B | 81.96 | 84.06 | +2.10 |
| Mistral-24B→Mistral-7B | 22.97 | 26.10 | +3.13 | InternVL3.5-38B→InternVL3.5-8B | 86.80 | 93.44 | +6.64 |
| Qwen2.5-32B→Gemma-2-9B | 24.49 | 27.63 | +3.14 | Qwen3-VL-32B→Gemma-3-12B | 81.96 | 85.28 | +3.32 |
| Mistral-24B→Gemma-2-9B | 24.49 | 27.12 | +2.63 | InternVL3.5-38B→Gemma-3-12B | 81.96 | 84.71 | +2.75 |
| GPT-6 Astra→Gemma-2-9B | 24.49 | 29.18 | +4.69 | GPT-6 Astra→Gemma-3-12B | 81.96 | 90.72 | +8.76 |
| MiniMax-M3→Gemma-2-9B | 24.49 | 28.27 | +3.78 | MiniMax-M3→Gemma-3-12B | 81.96 | 87.61 | +5.65 |

unchanged. We then fit a new attribution model using the method-specific configuration and evaluate it on held-out distilled-student responses. Thus, the original, frozen-KD, and retrained-KD stages share the same evaluation questions and source label space.

**Implementation and hardware.** Experiments are conducted on a server equipped with four NVIDIA RTX 6000 Ada Generation GPUs, each with 48 GB of memory, running Ubuntu 20.04.6 LTS. Our software environment uses Python 3.12.13, PyTorch 2.12.1, Transformers 5.12.1, PEFT 0.19.1, Accelerate 1.14.0, and bitsandbytes 0.49.2. Teacher responses are cached and reused across KD runs. API-based teacher generation is performed through remote inference services.

# D SUPPLEMENTARY RESULTS

## D.1 TASK UTILITY

We use the standard task-specific evaluation metric for each benchmark. For ASQA, we report the Disambiguation-ROUGE (DR) score introduced with the benchmark (Stelmakh et al., 2022). DR jointly evaluates whether a generated long-form answer covers the information required by the different interpretations of an ambiguous question and whether it remains consistent with the reference long-form answer. Specifically, it is defined as the geometric mean of Disambig-F1 and ROUGE-L:

$$\mathrm{DR} = \sqrt{\text{Disambig-F1} \times \text{ROUGE-L}}. \tag{27}$$

Disambig-F1 measures whether the generated answer contains correct answers to the benchmark's disambiguated questions, while ROUGE-L measures the overlap between the generated long-form response and the human reference. A higher DR score therefore indicates better overall answer quality across both ambiguity resolution and long-form response quality.

For SCIENCEQA, we report answer accuracy, following the standard multiple-choice evaluation (Lu et al., 2022). The predicted answer choice is compared with the ground-truth option, and accuracy is computed as the proportion of evaluation questions answered correctly. Because our models also generate explanations, only the predicted answer choice is used for computing task accuracy.

To verify that the observed attribution degradation is not caused by ineffective distillation, we further evaluate task utility before and after knowledge distillation. As shown in Table 8, sequence-level KD consistently improves downstream task performance across all evaluated teacher–student pairs. On ASQA, DR improves across all seven teacher–student settings, with gains ranging from 1.59 to 4.69 points. On SCIENCEQA, answer accuracy similarly increases in every setting, with improvements ranging from 2.10 to 8.76 percentage points. These results confirm that the observed attribution degradation does not arise from unsuccessful student adaptation or reduced downstream task capability. Instead, it occurs even when distillation improves the student's performance on the underlying task.