TEACHING STUDENTS WHERE TO LOOK: SALIENCY-GUIDED KNOWLEDGE DISTILLATION FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) is the standard approach for compressing large language models (LLMs) into compact deployable students by training the student to match the teacher's output distribution. While many improvements have been proposed, such as alternative divergence choices, on-policy generation, and difficulty-adaptive sample reweighting, many existing methods deliver the distillation signal pointwise in output space, providing no constraint on how the student behaves in the neighborhood of training inputs. We empirically show that this pointwise objective leaves large teacher–student gaps in input neighborhoods despite close output agreement, and propose **SaKD** (**Sa**liency-Guided **K**nowledge **D**istillation), which closes this gap by lifting pointwise output matching to a saliency-aware first-order matching objective. Concretely, SaKD combines a noise-injected Kullback–Leibler (KL) divergence term that, via a noise–Jacobian identity for autoregressive token-level KL, provides an implicit teacher–student Jacobian-matching signal along coupled perturbation directions, with a saliency-divergence reweighting that focuses the loss on samples with larger teacher–student input-saliency disagreement. Extensive experiments show that SaKD consistently outperforms seven strong KD baselines across Qwen3 and LLaMA student scales, with students reaching correct answers through teacher-like reasoning rather than surface shortcuts.

1 INTRODUCTION

Knowledge distillation (KD) (Hinton et al., 2015) is the workhorse for compressing large language models (LLMs) into deployable students. Modern KD pipelines for LLMs follow a remarkably consistent template: they minimize a divergence between teacher and student output distributions, such as forward Kullback–Leibler (KL) divergence (Hinton et al., 2015), reverse KL (Gu et al., 2024), generalized KL (Agarwal et al., 2024), or skewed KL (Ko et al., 2024). These objectives can be further combined with student-generated trajectories (Agarwal et al., 2024) or difficulty-adaptive sample weights (He et al., 2025). What these methods share in common is that the distillation signal is delivered pointwise in the output space for each training token.

Output agreement, however, does not imply agreement in how the model uses input evidence. On SQuAD 2.0, the input saliency of a standard-KD student, the per-token gradient norm of the response log-likelihood with respect to the input embeddings, concentrates on the labelled answer span $3\times$ more than the teacher's does. This difference motivates transferring not only output probabilities but also sensitivity to the input. Students can fit the teacher's outputs at training points while retaining different input sensitivities. This phenomenon is not an artifact of any particular divergence; it follows from the geometry of the objective. Pointwise output matching constrains the teacher's function values: it controls divergence at each training input but says nothing about how the student behaves between training points. A Taylor expansion of the loss around a training input x_i decomposes neighborhood error into a zero-order term, the KL that pointwise KD minimizes, and a first-order term governed by the teacher–student input Jacobian gap (Czarnecki et al., 2017); the output discrepancy can vanish at training points while the models' input Jacobians still differ.

We illustrate this gap empirically on Dolly-15K in Figure 1. Out of 500 evaluation samples, 14% fall into a high-sensitivity region where the clean teacher–student KL $D(x_i)$ is small but the perturbed KL $\mathbb{E}_\xi[D(x_i+\xi)]$ is much larger. Across the full set, the 95th percentile of teacher–student KL grows

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.

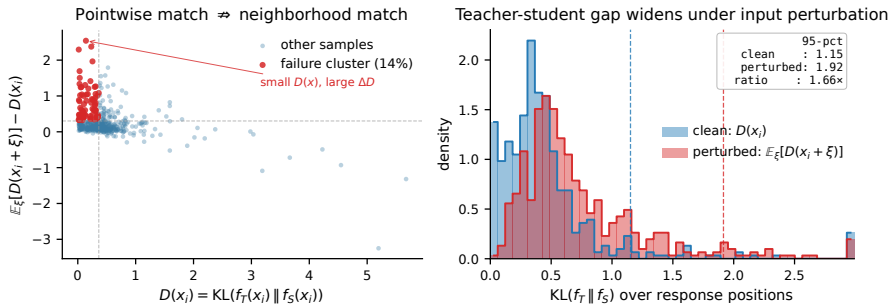

Figure 1: Pointwise output matching leaves large teacher–student gaps in input neighborhoods. We perturb the embeddings of a LLaMA-1B student distilled from LLaMA-8B on Dolly-15K. Left: per-sample scatter of clean KL against its noise-averaged change; the highlighted region marks samples with low clean KL that diverge sharply in the neighborhood. Right: distributions of clean and perturbed KL on the same set; the perturbed distribution has a noticeably heavier tail.

by $1.66\times$ when inputs are perturbed by Gaussian noise of magnitude $\sigma = 0.01\|e\|$. This diagnostic motivates matching local responses in addition to clean outputs.

To close this gap, we extend pointwise output matching with a local Sobolev-style objective. Our method, SaKD, augments standard KD with noise-injected KL: by evaluating the teacher–student KL under small coupled Gaussian perturbations of the input embeddings, the loss implicitly penalizes a Fisher-weighted score mismatch between the teacher and student, providing a tractable first-order signal without explicitly computing the full Jacobian (Theorem 3). We further use teacher–student saliency divergence to identify examples where the student relies on different input evidence, yielding an entropy-regularized DRO interpretation of the resulting sample weights (Theorem 5). We summarize our contributions as follows:

- We empirically demonstrate that pointwise output matching leaves large teacher–student gaps in input neighborhoods despite close training-point output agreement, motivating a saliency-aware first-order matching objective.
- We propose SaKD: noise-injected KL provides an implicit first-order Jacobian-matching signal, while saliency-divergence reweighting allocates training effort through an entropy-regularized DRO-inspired rule.
- We validate SaKD across Qwen3 and LLaMA student scales, showing that it outperforms seven strong KD baselines on Dolly-15K, and we introduce an evidence-concentration diagnostic that reveals attribution differences hidden by output-level metrics.

2 Preliminaries

2.1 Autoregressive Knowledge Distillation

Let f_T and f_S denote the teacher and student language models, respectively. Given an input sequence $x = (x_1, \ldots, x_L)$, model M has token embeddings $e_M \in \mathbb{R}^{L \times d_M}$ and produces a next-token probability distribution $p_M(\cdot \mid x, t) \in \Delta^{|\mathcal{V}|-1}$ at each position t, where $\mathcal{V}$ is the vocabulary. We partition the sequence into prompt positions $\mathcal{P} = \{1, \ldots, L_p\}$ and response positions $\mathcal{R} = \{L_p{+}1, \ldots, L\}$.

Standard knowledge distillation (Hinton et al., 2015) trains the student by minimizing the forward KL divergence over response positions on a training set $\{x_i\}_{i=1}^n$:

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t \in \mathcal{R}_i} D_{\mathrm{KL}}\big(p_T(\cdot \mid x_i, t) \,\big\|\, p_S(\cdot \mid x_i, t)\big). \tag{1}$$

We define the per-sample distillation error as $D(x) = \sum_{t \in \mathcal{R}} D_t(x)$, where

$$D_t(x) = \sum_{v \in \mathcal{V}} p_T^v(x,t) \log \frac{p_T^v(x,t)}{p_S^v(x,t)} \tag{2}$$

is the per-position KL divergence at response position t.

Both models are evaluated on the same tokenized prompt and reference response, with aligned vocabulary events and token positions. For local analysis, $x + \delta$ denotes a common perturbation coordinate: model M receives $e_M + P_M \delta$, where P_M maps that coordinate into its embedding space. The maps are held fixed when taking input derivatives; identical coordinates give $P_T = P_S = I$. Thus $\nabla_x \log p_M = P_M^\top \nabla_{e_M} \log p_M$ compares responses to the same perturbation even when embedding widths differ (Appendix A.7).

2.2 Input Saliency

The input saliency at position j quantifies the dependence of the model's response generation on the j-th input token:

$$s_j(x) = \left\| \frac{\partial \log P(\text{response} \mid x)}{\partial e_j} \right\|_2, \tag{3}$$

where $\log P(\text{response} \mid x) = \sum_{t \in \mathcal{R}} \log p(x_t \mid x_{<t})$ is the autoregressive response log-likelihood. The saliency vector $\mathbf{s}(x) = (s_1, \ldots, s_{L_p}, 0, \ldots, 0) \in \mathbb{R}^L$ is nonzero only at prompt positions. We write $\mathbf{s}_T(x)$ and $\mathbf{s}_S(x)$ for the teacher and student saliency vectors, respectively.

For the purpose of computing distributional divergence between saliency patterns, we define the normalized saliency distribution via temperature-scaled softmax over prompt positions:

$$\hat{s}_j = \frac{\exp(s_j/\tau_s)}{\sum_{k \in \mathcal{P}} \exp(s_k/\tau_s)}, \tag{4}$$

where $\tau_s > 0$ is a temperature parameter controlling the sharpness of the resulting distribution over prompt positions.

2.3 Input–Output Jacobian

At each response position t, let $J_{M,t}^{(j),v} = \nabla_{e_{M,j}} \log p_M^v(x,t)$ be the score gradient for vocabulary token v. Stacking these rows over response positions and vocabulary entries gives $\mathcal{J}_M^{(j)} \in \mathbb{R}^{(|\mathcal{R}||\mathcal{V}|) \times d_M}$. The full input–output score Jacobian has the block decomposition

$$\mathcal{J}_M(x) = \left[\mathcal{J}_M^{(1)}(x), \ldots, \mathcal{J}_M^{(L)}(x)\right]. \tag{5}$$

A key structural property connects saliency to this Jacobian:

Lemma 1 (Saliency is controlled by the Jacobian block norm)**.** *Let y_t denote the realized response token at position t. For each prompt position j,*

$$s_{M,j}(x) = \left\| \sum_{t \in \mathcal{R}} J_{M,t}^{(j),y_t}(x) \right\|_2 \le \sqrt{|\mathcal{R}|}\, \|\mathcal{J}_M^{(j)}(x)\|_F. \tag{6}$$

Proof. Differentiating the response log-likelihood gives the sum of the selected score rows. Cauchy–Schwarz bounds its norm by $\sqrt{|\mathcal{R}|}(\sum_t \|J_{M,t}^{(j),y_t}\|_2^2)^{1/2}$. Each selected row belongs to a distinct response-position block, so the sum of its squared norms is at most $\|\mathcal{J}_M^{(j)}\|_F^2$. □

The bound is one-sided: saliency summarizes a realized response and discards gradient direction. We therefore use saliency disagreement to guide training, while the noisy output loss supplies the derivative-sensitive objective. With aligned input tokens, both saliency vectors lie in $\mathbb{R}^L$ regardless of the models' embedding widths.

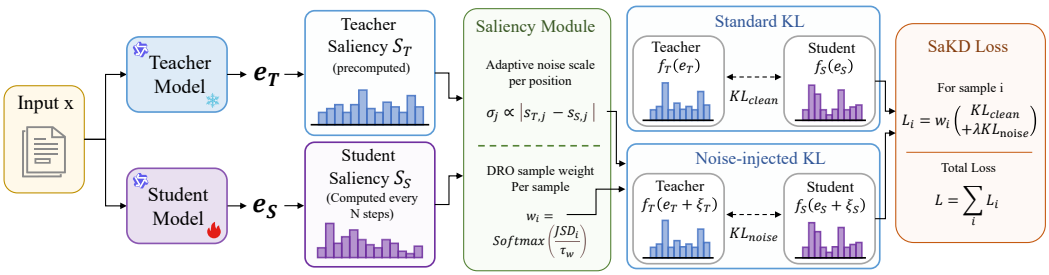

Figure 2: Framework of proposed method SaKD. Each input is processed by both teacher and student along a clean KL path and a noise-injected KL path that maps a shared Gaussian draw into the two embedding spaces. Theorem 3 connects the noisy loss to teacher–student Jacobian matching along these coupled directions. A saliency module sets the per-position noise scale σ_j and the DRO-inspired sample weight w_i; only the student is updated.

3 Method: Saliency-Guided Knowledge Distillation

SaKD extends standard knowledge distillation from pointwise output matching to local teacher–student matching around each training input. The core objective is a noise-injected KL loss that implicitly penalizes discrepancies in input sensitivity, yielding a tractable Jacobian-aware distillation signal without explicit Jacobian computation (§3.2). SaKD further uses saliency disagreement to concentrate this signal on examples with large teacher–student mismatch (§3.3). Figure 2 provides an overview of our pipeline.

3.1 Why Pointwise Matching Is Not Enough

Standard KD matches the teacher and student only at observed training inputs by minimizing the distillation error $D(x_i)$. Such pointwise agreement, however, does not constrain the student's local behavior around x_i: two models may agree at x_i while exhibiting different input sensitivities under small embedding-level perturbations. Here, we formalize this gap by decomposing the neighborhood discrepancy into the standard pointwise KD error and a first-order term induced by the teacher–student Jacobian mismatch.

Theorem 2 (Neighborhood error decomposition)**.** *Assume* (A1)–(A2) *and let* $\varepsilon_p > 0$ *be the probability floor of* (A2)*. Write* $\|J_T(x,t)\|^2_{F,p_T} = \sum_v p_T^v(x,t)\|J_T^v(x,t)\|^2$ *for the teacher's Fisher-weighted score norm and* $G_T(x) = \max_{t\in\mathcal{R}} \|J_T(x,t)\|_{F,p_T}$*. Put* $r = |\mathcal{R}|$ *and* $\|J_T - J_S\|^2_{F,p_T} = \sum_{t,v} p_T^v \|J_T^v - J_S^v\|^2$ *in the common perturbation coordinates. Then for any perturbation* δ *with* $\|\delta\| \le \epsilon$*,*

$$D(x+\delta) \;\le\; D(x) \;+\; \epsilon\Big[\underbrace{\sqrt{r}\,\|J_T(x) - J_S(x)\|_{F,p_T}}_{\text{first-order: Jacobian gap}} \;+\; \underbrace{\kappa_T(x)\,\sqrt{D(x)}}_{\text{zero-order coupling}}\Big] \;+\; O(\epsilon^2), \tag{7}$$

where $\kappa_T(x) = \sqrt{2|\mathcal{R}|}\,G_T(x)/\varepsilon_p$ *depends only on the teacher and the probability floor.*

The bound separates pointwise error from a derivative-sensitive term. Matching outputs at training points alone does not constrain the Jacobian gap; this motivates retaining clean KD while adding a local matching signal. The next section obtains that signal through perturbed output KL, without explicitly constructing either Jacobian.

3.2 Noise-Injected KL for Implicit Jacobian Matching

Theorem 2 suggests that improving local distillation fidelity requires controlling the teacher–student Jacobian mismatch. Explicitly computing this mismatch is infeasible for LLMs, since the input–output Jacobian scales with both sequence length and vocabulary size. SaKD therefore uses a tractable surrogate: instead of evaluating the KL only at the clean input, we also evaluate it at a Gaussian-perturbed embedding. Intuitively, if the student matches the teacher not only at x but also in a neighborhood around x, their local input sensitivities must also draw closer to each other.

The following identity formalizes this intuition. Building on the noise–Jacobian connection of Srinivas & Fleuret (2018), we derive its autoregressive, token-level KL form.

Theorem 3 (Noise–Jacobian identity for autoregressive KL)**.** *Assume* (A1) *smoothness and* (A2) *a bounded log-likelihood ratio with probability floor ε_p (Appendix A.1). Let $D(x) = \sum_{t\in\mathcal{R}} D_t(x)$ be the per-sample KL (Eq. 2). Suppose teacher and student receive the same coordinate perturbation $\xi \sim \mathcal{N}(0, \sigma^2 I)$ through their respective maps P_T, P_S. Then*

$$\mathbb{E}_\xi\big[D(x{+}\xi)\big] = D(x) + \frac{\sigma^2}{2}\sum_{t\in\mathcal{R}}\big[\mathcal{F}_t(x) + R_t(x)\big] + O(\sigma^4), \tag{8}$$

where the nonnegative derivative term is the Fisher-weighted score gap

$$\mathcal{F}_t(x) = \sum_{v\in\mathcal{V}} p_T^v(x,t)\,\big\|\nabla_x \log p_T^v(x,t) - \nabla_x \log p_S^v(x,t)\big\|^2, \tag{9}$$

and the residual satisfies $|R_t(x)| \le C_2\|p_T(\cdot,t) - p_S(\cdot,t)\|_1$ for a constant C_2 depending on ε_p and the probability-Laplacian bounds (full form in Appendix A.3).

The complete proof is given in Appendix A.3.

Theorem 3 gives a practical first-order distillation signal: near output agreement, the residual decreases and perturbed KL exposes the score gap. For position-adaptive covariance Σ, the derivative contribution becomes $\frac{1}{2}\sum_{t,v} p_T^v \|\Sigma^{1/2}(J_T^v - J_S^v)\|^2$ (Appendix A.3). We estimate the noisy objective with one shared draw per update:

$$\mathcal{L}_{\text{noise}}(x) = D_{\text{KL}}\big(f_T(e_T + P_T\xi)\,\|\,f_S(e_S + P_S\xi)\big), \quad \xi \sim \mathcal{N}(0,\Sigma). \tag{10}$$

The same local analysis connects clean and noisy KL to neighborhood discrepancy.

Corollary 4 (Neighborhood control)**.** *Assume* (A1)–(A3)*. For $\lambda, \sigma > 0$ and $\epsilon^2 \le \lambda\sigma^2/2$,*

$$\max_i \sup_{\|\delta\|\le\epsilon} D(x_i+\delta) \le \max_i \left[\frac{3}{2}D(x_i) + \lambda\mathbb{E}_\xi D(x_i+\xi) + \frac{r_i\epsilon^2}{\lambda\sigma^2} + \mathcal{E}_i\right], \tag{11}$$

where $r_i = |\mathcal{R}_i|$ and $\mathcal{E}_i$ contains the explicit curvature and Taylor remainders in Appendix A.4.

The bound concerns the common perturbation coordinates; adaptive covariance gives the corresponding ellipsoidal neighborhood after a change of coordinates. It links the clean and noisy objectives to local agreement under the stated regularity conditions.

Position-adaptive noise allocation. A uniform perturbation treats all positions as equally important, although teacher–student mismatch is often localized. We therefore allocate larger noise to positions where the teacher and student saliency scores disagree more:

$$\sigma_j = \sigma_0\,\overline{\|e\|}\,\min\left(\frac{g_j}{\overline{g}}, 5\right), \qquad g_j = |s_{T,j} - s_{S,j}|, \tag{12}$$

where σ_0 is a relative noise scale, $\overline{\|e\|}$ is the reference mean embedding norm, and $\overline{g}$ is the positional mean of g_j. For $\overline{g} > 0$, the ratio is capped at 5. This allocation focuses the implicit Jacobian probe on positions where the teacher and student rely on different input evidence.

For cross-architecture distillation ($d_T \neq d_S$), teacher and student embeddings live in different spaces. At each sample and position, we map a shared $K{=}128$-dimensional Gaussian draw through fixed, column-normalized random matrices, rescale by $\sqrt{d_M/K}$, and multiply by σ_j; if $\overline{g} = 0$, we set all $\sigma_j = 0$. The resulting maps P_T, P_S are specified in Appendix A.7.

3.3 Saliency-Guided Sample Reweighting

The noise-injected KL provides a Jacobian-aware distillation signal for each sample. However, not all samples require the same amount of local correction. Saliency disagreement identifies examples on which their response-conditioned input sensitivities differ. We therefore upweight examples with larger teacher–student saliency disagreement.

A natural way to formalize this idea is entropy-regularized distributionally robust optimization (DRO) (Duchi & Namkoong, 2021; Sagawa et al., 2020). Let c_i be a local-discrepancy score combining pointwise KD error and first-order mismatch:

$$\min_{f_S} \max_{w \in \Delta^{B-1}} \sum_{i=1}^{B} w_i \, c_i + \tau_w \, H(w), \tag{13}$$

where $c_i = D(x_i) + \frac{\epsilon^2}{2} \sum_t \mathcal{F}_t(x_i)$, $H(w) = -\sum_i w_i \log w_i$, and $\tau_w > 0$ controls how concentrated the weights are on high-vulnerability samples.

Theorem 5 (Optimal DRO weights)**.** *The inner maximization of Eq.* (13) *admits the closed-form solution* $w_i^* = \mathrm{softmax}(c_i/\tau_w)$.

Theorem 5 suggests assigning larger weights to examples with larger local vulnerability. Since c_i contains the Fisher-weighted score gap $\mathcal{F}_t$, computing it exactly would defeat the purpose of using an implicit Jacobian surrogate. We then use the Jensen–Shannon divergence (JSD) between the normalized teacher and student saliency distributions as a tractable proxy:

$$\mathrm{JSD}_i = \mathrm{JSD}\big(\hat{\mathbf{s}}_T^i, \hat{\mathbf{s}}_S^i\big). \tag{14}$$

Motivated by the saliency–Jacobian relation in Lemma 1, we use this disagreement as a scheduling proxy rather than an exact estimate of c_i. The resulting sample weight is

$$w_i = \frac{\exp(\mathrm{JSD}_i/\tau_w)}{\frac{1}{B}\sum_{j=1}^{B} \exp(\mathrm{JSD}_j/\tau_w)}, \tag{15}$$

where the normalization ensures $\frac{1}{B}\sum_i w_i = 1$. We detach w_i from the computation graph, so the weights only modulate the distillation loss rather than introducing an additional saliency-matching objective.

3.4 Complete Objective

SaKD combines the three design choices above into a single mini-batch objective: clean KD preserves pointwise output agreement, noise-injected KL encourages local first-order agreement, and saliency-guided weights focus the update on examples with larger teacher–student mismatch. For a mini-batch of size B, the objective is

$$\mathcal{L}_{\mathrm{SaKD}} = \frac{1}{B}\sum_{i=1}^{B} w_i \Big[D_{\mathrm{KL}}\big(f_T(x_i) \,\|\, f_S(x_i)\big) + \lambda \, D_{\mathrm{KL}}\big(f_T(e_T^i + P_T\xi_i) \,\|\, f_S(e_S^i + P_S\xi_i)\big)\Big], \tag{16}$$

where w_i is the saliency-divergence weight from Eq. (15), ξ_i is the position-adaptive Gaussian perturbation from Eq. (12), the same ξ_i is mapped into both models, and $\lambda > 0$ controls the strength of the noise-injected KL term.

This objective is used on SaKD update steps. On other steps, we use the standard clean KD loss to reduce overhead; the student saliency is refreshed every N steps for computing both the adaptive noise scale and the sample weights. The complete training procedure is given in Algorithm 1; full implementation details and wall-clock comparisons are provided in Appendix B.

4 Experiments

We evaluate SaKD on task-agnostic instruction following (§4.2), with ablations isolating each component (§4.3), and evidence concentration analysis examining teacher–student attribution alignment (§4.4).

4.1 Experimental Setup

Models. We distill from Qwen3-8B (teacher) to Qwen3-1.7B and Qwen3-0.6B (students) (Yang et al., 2025), representing $4.7\times$ and $13\times$ compression ratios respectively. The teacher is frozen throughout all experiments.

Table 1: Task-agnostic instruction following (ROUGE-L). Training on Dolly-15K, evaluation on three held-out benchmarks. All results are averaged over 5 random seeds. The best results are in **Bold**, and the second-best results are Underline. Teachers: Qwen3-8B for Qwen students, and LLaMA-3.1-8B for the LLaMA student.

Student	Benchmark	SFT	KD-KL	KD-RKL	SeqKD	GKD	DistiLLM	DA-KD	SaKD
Qwen3-1.7B	DollyEval	27.89±.42	28.11±.22	27.97±.31	22.24±.48	28.10±.38	28.08±.22	28.11±.34	**29.27**±.17
	S-NatInst	29.29±1.88	30.79±.38	29.66±1.13	23.28±3.64	30.04±1.88	31.17±.75	31.54±.75	**34.17**±.38
	Unnatural	35.73±.53	35.69±.31	35.76±.29	35.48±.91	35.78±.37	35.71±.26	35.84±.36	**36.13**±.18
	Avg	30.97	31.53	31.13	27.00	31.31	31.65	31.83	**33.19**
Qwen3-0.6B	DollyEval	27.20±.54	27.67±.27	27.52±.30	22.15±.35	27.61±.25	27.72±.30	27.55±.32	**28.58**±.15
	S-NatInst	30.42±2.13	28.16±4.51	31.17±3.38	15.02±2.63	30.79±2.25	30.04±3.00	28.91±3.38	**32.29**±1.13
	Unnatural	35.48±.47	35.58±.29	35.52±.30	34.96±.61	35.55±.35	35.64±.25	35.47±.27	**36.03**±.12
	Avg	31.03	30.47	31.40	24.04	31.32	31.13	30.64	**32.30**
LLaMA-3.2-1B	DollyEval	26.65±1.41	27.39±1.00	27.19±1.24	22.42±.61	27.26±1.59	27.15±.26	27.29±.38	**28.40**±.72
	S-NatInst	30.54±1.36	31.09±.95	31.15±.58	29.53±1.57	30.87±1.74	30.85±1.60	30.87±1.09	**33.30**±1.36
	Unnatural	35.14±.90	35.79±.45	35.94±1.03	32.93±.97	35.46±1.31	35.69±.83	35.70±1.13	**36.95**±.25
	Avg	30.78	31.42	31.43	28.29	31.20	31.23	31.29	**32.88**

Datasets. Following DA-KD (He et al., 2025), we train all methods on Dolly-15K (Databricks, 2023) and evaluate task-agnostic generalization on three held-out instruction benchmarks: DollyEval, S-NatInst (Super-Natural Instructions), and Unnatural Instructions. The evaluation metric is ROUGE-L (Lin, 2004) on each benchmark together with the macro-average across the three; this setup tests whether distillation methods generalize across instruction types unseen during training. Inputs are truncated to a maximum sequence length of 512. The evidence-concentration analysis in §4.4 additionally uses the answerable subset of SQuAD 2.0 (Rajpurkar et al., 2018), whose annotated answer spans support measuring where each model places its input saliency.

Algorithm 1 SaKD Training

Require: Teacher f_T (frozen), student f_S
Require: Training set, cached $\{\mathbf{s}_T^i\}$
Require: Hyperparameters: $\lambda, \sigma_0, \tau_w, N$
for each epoch **do**
  for each mini-batch $\mathcal{B}$ **do**
    $\mathbf{t}_i, \mathbf{s}_i \leftarrow f_T(x_i), f_S(x_i)$
    $\mathrm{kl}_i \leftarrow D_{\mathrm{KL}}(\mathbf{t}_i \| \mathbf{s}_i)$
    if step $\bmod N = 0$ **then**
      Compute $\mathbf{s}_S^i$; load cached $\mathbf{s}_T^i$
      $\sigma_j \leftarrow$ Eq. (12)
      Draw shared $\xi_i \sim \mathcal{N}(0, \Sigma_i)$
      $\xi_T^i = P_T \xi_i, \quad \xi_S^i = P_S \xi_i$
      $\mathrm{kl}_i^{\text{noise}} \leftarrow D_{\mathrm{KL}}\big(f_T(e_T^i + \xi_T^i) \,\|\, f_S(e_S^i + \xi_S^i)\big)$
      $w_i \leftarrow \mathrm{softmax}(\mathrm{JSD}_i / \tau_w) \cdot B$
      $\mathcal{L} \leftarrow \frac{1}{B} \sum_i w_i (\mathrm{kl}_i + \lambda\, \mathrm{kl}_i^{\text{noise}})$
    else
      $\mathcal{L} \leftarrow \frac{1}{B} \sum_i \mathrm{kl}_i$
    end if
    Update f_S via $\nabla_{f_S} \mathcal{L}$
  end for
end for

Baselines. We compare against seven methods spanning the distillation landscape: SFT (supervised fine-tuning, no teacher), KD-KL (Hinton et al., 2015), KD-RKL (Gu et al., 2024), SeqKD (Kim & Rush, 2016), GKD (Agarwal et al., 2024), DistiLLM (Ko et al., 2024), and DA-KD (He et al., 2025). All methods share identical training hyperparameters; method-specific parameters follow the original papers.

Training details. All runs use 10 epochs, effective batch size 32 (batch 8 × gradient accumulation 4), learning rate 1e-5 with cosine schedule and 3% warmup, weight decay 0.01, max sequence length 512, KL temperature $T = 2$, and fp16 mixed precision. For SaKD: $\lambda = 0.5$, $\sigma_0 = 0.005$ (relative to embedding norm), $\tau_w = 1.0$, $N = 5$. The main results report mean ± std over 5 random seeds (42, 123, 456, 789, 2024), and the ablation seed counts are specified in Table 2. All experiments run on A100 80GB GPUs.

4.2 Task-Agnostic Instruction Following

Table 1 reports the full comparison. SaKD ranks first on every benchmark across all three students. Its macro-average exceeds the strongest baseline by 1.36 ROUGE-L points for Qwen3-1.7B and 0.90 points for Qwen3-0.6B. The Qwen runs also exhibit smaller seed dispersions for SaKD.

Table 2: Ablation study on Dolly-15K with Qwen3-8B → 0.6B.

Configuration	λ	τ_w	DollyEval	S-NatInst	Unnatural	Avg.
KD-KL baseline	0	—	27.67±.27	28.16±4.51	35.58±.29	30.47
+ Noise KL only	0.5	∞	28.12±.07	31.24±0.97	35.78±.12	31.81
+ Reweight only	0	1.0	28.22±.08	31.42±0.88	35.83±.07	31.49
Full SaKD	0.5	1.0	**28.58**±.15	**32.29**±1.13	**36.03**±.12	**32.30**

The largest Qwen3-1.7B improvement is on S-NatInst, the benchmark with diverse multi-sentence instructions: SaKD scores 34.17 versus 31.54 for the strongest baseline. The improvement is even stronger for the LLaMA pair: SaKD reaches 28.40, 33.30, and 36.95 on DollyEval, S-NatInst, and Unnatural, respectively, exceeding the strongest baseline on each benchmark by 1.01, 2.15, and 1.01 points. Its macro-average of 32.88 is 1.45 points above the best baseline (31.43), showing that the local matching signal transfers beyond the Qwen architecture. For different embedding widths, the coupled perturbations compare sensitivities in the common coordinates of Appendix A.7. SeqKD scores lower across all three students, consistent with the limitations of transferring only teacher-generated targets (Agarwal et al., 2024).

4.3 Ablation Study

To disentangle the contributions of SaKD's two components, noise KL for implicit Jacobian matching and saliency-guided reweighting, we evaluate four ablation configurations on Dolly-15K with Qwen3-0.6B; results are shown in Table 2.

Both components improve over KD-KL and combine constructively in the full method. Their macro-average scores are 31.81 for noise KL only, 31.49 for reweighting only, and 32.30 for full SaKD, compared with 30.47 for KD-KL. This is consistent with their complementary roles: the noisy loss adds a local matching signal, while reweighting prioritizes examples by saliency disagreement.

4.4 Evidence Concentration Analysis

Beyond output-level metrics, we investigate how SaKD changes teacher–student input-saliency patterns. We use SQuAD 2.0 (Rajpurkar et al., 2018), where ground-truth answer spans provide a natural lens for measuring which input tokens each model relies upon.

Evidence concentration. For each sample with an annotated answer span, we define evidence concentration (EC) as the fraction of input saliency that falls on the answer-span tokens:

$$\mathrm{EC} = \frac{\sum_{j \in \text{answer span}} s_j}{\sum_j s_j}. \tag{17}$$

EC describes the fraction of prompt saliency assigned to the annotated answer span. We compare each student with its teacher to assess attribution concentration.

Results. Table 3 reports EC across two distillation pairs, one same-architecture and one cross-architecture. Both teachers place only a few percent of their saliency on the answer span, attending broadly to the surrounding context. Standard KD students collapse onto a positional shortcut: they over-attend to the labelled answer by 2–3× relative to their teacher even though their output distributions match. SaKD substantially mitigates this collapse on both pairs, closing roughly three quarters of the over-concentration gap so that residual over-attention sits close to the teacher's natural level on this benchmark. This finding is invisible to standard evaluation: a model that answers correctly via positional shortcuts may score well on EM/F1 yet fail under adversarial or distribution-shifted inputs where the shortcut breaks. EC provides a complementary diagnostic that captures reasoning fidelity rather than only the surface-level output quality.

Table 3: Evidence Concentration on SQuAD 2.0 val (seed 42). EC is the fraction of input saliency on the annotated answer span. Ratios are relative to the teacher's EC.

	Qwen3-8B → 0.6B		LLaMA-3.1-8B → 3.2-1B	
	EC	Ratio to teacher	EC	Ratio to teacher
Teacher	0.055	1.0×	0.044	1.0×
KD-KL	0.169	3.1×	0.096	2.2×
SaKD	**0.083**	1.5×	**0.059**	1.3×

5 Related Work

5.1 Knowledge Distillation for LLMs

Knowledge distillation (Hinton et al., 2015) trains a compact student to mimic a larger teacher by minimizing the divergence between their output distributions; for autoregressive LLMs, this is the per-token KL on the next-token distribution. SeqKD (Kim & Rush, 2016) replaces ground-truth targets with teacher-generated outputs, at the cost of output diversity (Agarwal et al., 2024). Subsequent work has mostly improved the distillation objective: MiniLLM (Gu et al., 2024) adopts reverse KL, GKD (Agarwal et al., 2024) uses JSD with on-policy student trajectories, DistiLLM (Ko et al., 2024) introduces skew-KL for stability, and Wu et al. (2025) propose adaptive KL variants. Beyond the objective, DA-KD (He et al., 2025) reweights samples by output-level difficulty, Hsieh et al. (2023) distill chain-of-thought rationales, and Fu et al. (2023) target multi-step reasoning; see Xu et al. (2024) for a recent survey. These approaches primarily modify output-level supervision or training-data allocation. SaKD instead adds first-order Jacobian matching to the distillation objective.

5.2 Gradient-based Feature Attribution

Gradient-based feature attribution methods quantify each input feature's contribution to a model's prediction (Simonyan et al., 2014; Sundararajan et al., 2017; Smilkov et al., 2017; Shrikumar et al., 2017; Lundberg & Lee, 2017), with theoretical analyses (Ancona et al., 2018) and sanity checks (Adebayo et al., 2018) clarifying their behavior. In NLP, first-order gradient saliency has been used to interpret token importance in recurrent (Li et al., 2016) and Transformer (Bastings & Filippova, 2020) models, where it is generally more faithful to the model's actual decision than raw attention scores. A smaller line of work uses gradient attribution as a training signal: Ross et al. (2017) penalize input gradients that disagree with human rationales, Erion et al. (2021) use expected gradients as a regularizer, and within distillation AD-KD (Wu et al., 2023) and Wang et al. (2022) align teacher–student attributions in encoder-only BERT classifiers—using direct attribution objectives. On the theoretical side, Srinivas & Fleuret (2018) study noise–Jacobian connections for squared error and soft-target cross-entropy, and Czarnecki et al. (2017) formalize derivative matching as Sobolev training for general regression. SaKD develops the autoregressive token-level KL analysis and uses saliency to guide coupled perturbations and sample weighting for decoder-only LLMs.

6 Conclusion

Pointwise KD can match outputs at observed inputs while leaving local input sensitivities unconstrained. SaKD addresses this gap with coupled-noise KL, which provides an implicit Fisher-weighted score-matching signal, and saliency-divergence reweighting, which focuses updates on examples where the student uses different input evidence. The noise–Jacobian identity and neighborhood bound formalize this connection in common perturbation coordinates. Across Qwen3 and LLaMA student scales, SaKD outperforms seven strong KD baselines on Dolly-15K; ablations show complementary gains from both components, and SQuAD 2.0 evidence concentration is closer to the teacher's. Thus clean KL preserves pointwise agreement while coupled perturbations and saliency weighting address local behavior. More broadly, these results suggest treating local sensitivity as a complementary distillation target when generalization beyond training points matters. Future work includes extending the same principle to on-policy distillation, reverse-KL objectives, and multi-modal alignment, as is studying richer coupled subspaces and lower-cost approximations for larger models.

AI USE STATEMENT

In this work, we used generative AI tools solely for manuscript editing to improve language and clarity, and for assistance with code development. We did not use generative AI tools for any other tasks. We reviewed all AI-assisted edits to ensure that they accurately preserved our intended meaning and verified and tested AI-assisted code for correctness. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

knowledge Distillation can transfer a teacher's biases and errors to the student. The models and datasets remain subject to their licenses, and attribution agreement does not by itself establish safety or factual reliability.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Databricks. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. *Databricks Blog*, 2023.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. In *Nature Machine Intelligence*, volume 3, pp. 620–631, 2021.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning (ICML)*, 2023.

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

Changyi He, Yifu Ding, Jinyang Guo, Ruihao Gong, Haotong Qin, and Xianglong Liu. DA-KD: Difficulty-aware knowledge distillation for efficient large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 22379–22391, 2025. URL https://proceedings.mlr.press/v267/he25c.html.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander 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.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *International Conference on Machine Learning (ICML)*, pp. 24872–24895, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 784–789, 2018.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://www.ijcai.org/proceedings/2017/371.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations (ICLR)*, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Suraj Srinivas and François Fleuret. Knowledge transfer with Jacobian matching. In *International Conference on Machine Learning (ICML)*, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Lean Wang, Lei Li, and Xu Sun. Gradient knowledge distillation for pre-trained language models. *arXiv preprint arXiv:2211.01071*, 2022. URL https://arxiv.org/abs/2211.01071. NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing.

Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang, and Rui Wang. AD-KD: Attribution-driven knowledge distillation for language model compression. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8449–8465, 2023.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *International Conference on Computational Linguistics (COLING)*, 2025.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

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

A Proofs of Theoretical Results

This appendix collects rigorous proofs of the four theoretical statements in §3: the first-order neighborhood error decomposition (Theorem 2), the autoregressive noise–Jacobian identity (Theorem 3), the neighborhood-control corollary (Corollary 4), and the DRO characterization of saliency-divergence weights (Theorem 5). We also analyze position-adaptive allocation and the coupling used for different embedding widths.

A.1 Notation and standing assumptions

We adopt the notation of §2. All input derivatives below are taken in the common perturbation coordinates, with the coupling maps and noise schedule fixed at the base example. For an input sequence x, let $p_T^v(x,t), p_S^v(x,t)$ denote the teacher and student next-token probabilities at response position t, and write $\ell_v(x,t) := \log p_T^v(x,t) - \log p_S^v(x,t)$ for the log-likelihood ratio. The per-position KL is $D_t(x) = \sum_v p_T^v(x,t)\,\ell_v(x,t)$ and the per-sample distillation error is $D(x) = \sum_{t\in\mathcal{R}} D_t(x)$. Score functions are $J_T^v(x,t) := \nabla_x \log p_T^v(x,t)$ and analogously for J_S^v; we write $\Delta := \mathrm{tr}(\nabla_x^2 \cdot)$ for the Laplacian.

We use the following assumptions where indicated.

(A1) Smoothness. The log-probabilities are C^4 in the common perturbation coordinates. For Gaussian expectations, the fourth derivative of D has operator norm bounded by M_4 throughout the perturbation space. On the neighborhoods considered, the derivatives needed below are bounded uniformly over response positions and vocabulary entries. In particular, let $M_{3,i}$ bound the third derivative of D on the radius-ϵ ball around x_i.

(A2) Positive probabilities. On these neighborhoods, $p_T^v, p_S^v \geq \varepsilon_p > 0$. This bounds the log-likelihood ratio and permits the log-Lipschitz estimates below.

(A3) Local curvature. For Corollary 4, the input Hessian $H_i = \nabla_x^2 D(x_i)$ is positive semidefinite. This is an additional condition on local input curvature, not a consequence of optimizing model parameters.

A.2 Proof of Theorem 2

Proof. Define $\phi(t) := D(x + t\delta)$ for $t \in [0,1]$. By the fundamental theorem of calculus,

$$D(x{+}\delta) - D(x) = \phi(1) - \phi(0) = \int_0^1 \langle \nabla_x D(x{+}t\delta),\, \delta\rangle\, dt. \tag{18}$$

Cauchy–Schwarz with $\|\delta\| \leq \epsilon$ gives $|\phi(1) - \phi(0)| \leq \epsilon \sup_{t\in[0,1]} \|\nabla_x D(x{+}t\delta)\|$, and Lipschitz continuity of $\nabla_x D$ from (A1) implies $\|\nabla_x D(x{+}t\delta)\| \leq \|\nabla_x D(x)\| + L_D\,\epsilon$, so

$$D(x{+}\delta) \leq D(x) + \epsilon\,\|\nabla_x D(x)\| + O(\epsilon^2). \tag{19}$$

It remains to bound $\|\nabla_x D(x)\|$.

Step 1: the exact gradient. Differentiating $D_t(x) = \sum_v p_T^v(x)\,\ell_v(x)$ and using $\nabla_x p_T^v = p_T^v\, J_T^v$ and $\nabla_x \ell_v = J_T^v - J_S^v$,

$$\nabla_x D_t(x) = \underbrace{\sum_v p_T^v\,\ell_v\,J_T^v}_{=:\,A_t} \;+\; \underbrace{\sum_v p_T^v\,(J_T^v - J_S^v)}_{=:\,B_t}. \tag{20}$$

The score identity $\sum_v p_T^v J_T^v = \sum_v \nabla_x p_T^v = \nabla_x(1) = 0$ states that J_T has zero mean under p_T. It does *not* annihilate A_t, in which the scalar ℓ_v varies with v and cannot be factored out of the sum; rather, it identifies A_t as a covariance,

$$A_t = \mathbb{E}_{p_T}[\ell\, J_T] = \mathbb{E}_{p_T}[(\ell - \mathbb{E}_{p_T}\ell)\, J_T] = \mathrm{Cov}_{p_T}(\ell, J_T), \tag{21}$$

which vanishes when $p_T = p_S$. We therefore bound A_t rather than discard it.

Step 2: bounding A_t by the pointwise error. By Cauchy–Schwarz in $L^2(p_T)$, $\|A_t\| \le \sqrt{\mathrm{Var}_{p_T}(\ell)}\,\|J_T(x,t)\|_{F,p_T}$. For the variance factor, (A2) supplies a probability floor $p_T^v, p_S^v \ge \varepsilon_p$; using $|\log a - \log b| \le |a-b|/\min(a,b)$ and $p_T^v \le 1$,

$$\mathrm{Var}_{p_T}(\ell) \;\le\; \mathbb{E}_{p_T}[\ell^2] \;\le\; \frac{1}{\varepsilon_p^2}\sum_v p_T^v (p_T^v - p_S^v)^2 \;\le\; \frac{\|p_T - p_S\|_1^2}{\varepsilon_p^2} \;\le\; \frac{2\,D_t(x)}{\varepsilon_p^2}, \tag{22}$$

the last step by Pinsker's inequality. Hence $\|A_t\| \le (\sqrt{2D_t(x)}/\varepsilon_p)\,\|J_T(x,t)\|_{F,p_T}$.

Step 3: bounding B_t. Applying Cauchy–Schwarz with weights p_T^v and $\sum_v p_T^v = 1$,

$$\|B_t\| \le \sqrt{\sum_v p_T^v}\,\sqrt{\sum_v p_T^v\,\|J_T^v - J_S^v\|^2} = \sqrt{\mathcal{F}_t(x)}. \tag{23}$$

Step 4: summing over response positions. By the triangle inequality and Cauchy–Schwarz over $t \in \mathcal{R}$, $\sum_t \sqrt{\mathcal{F}_t} \le \sqrt{|\mathcal{R}|}\,\|J_T - J_S\|_{F,p_T}$, where $\|J_T - J_S\|_{F,p_T}^2 = \sum_t \mathcal{F}_t$, and

$$\sum_t \|A_t\| \;\le\; \frac{\sqrt{2}\,G_T(x)}{\varepsilon_p}\sum_t \sqrt{D_t(x)} \;\le\; \frac{\sqrt{2|\mathcal{R}|}\,G_T(x)}{\varepsilon_p}\sqrt{\textstyle\sum_t D_t(x)} \;=\; \kappa_T(x)\,\sqrt{D(x)}. \tag{24}$$

Combining these,

$$\|\nabla_x D(x)\| \;\le\; \sqrt{|\mathcal{R}|}\,\|J_T - J_S\|_{F,p_T} + \kappa_T(x)\sqrt{D(x)}, \tag{25}$$

and substituting into (19) yields the claimed bound. □

At exact output agreement, the KL gradient is zero, while its leading local curvature can still depend on the score gap. Theorem 3 makes this connection explicit.

A.3 Proof of Theorem 3

Proof. By (A1), for $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$, $m = \dim \xi$, and any $g \in C^4$ with bounded fourth derivative, the standard Gaussian moment expansion gives

$$\mathbb{E}_\xi[g(x{+}\xi)] = g(x) + \tfrac{\sigma^2}{2}\,\Delta g(x) + O(\sigma^4), \tag{26}$$

where $\Delta = \mathrm{tr}(\nabla^2)$ is the Laplacian. Taylor expansion through third order, with vanishing odd Gaussian moments, gives a remainder bounded by $M_4\sigma^4 m(m+2)/24$ when $\sup_z \|\nabla^4 g(z)\|_{\mathrm{op}} \le M_4$. Applying (26) to $D = \sum_t D_t$,

$$\mathbb{E}_\xi[D(x{+}\xi)] = D(x) + \tfrac{\sigma^2}{2}\,\textstyle\sum_{t\in\mathcal{R}} \Delta D_t(x) + O(\sigma^4). \tag{27}$$

We compute ΔD_t for $D_t = \sum_v p_T^v\,\ell_v$. By the product-rule Laplacian $\Delta(fg) = f\,\Delta g + 2\,\nabla f\cdot\nabla g + g\,\Delta f$,

$$\Delta(p_T^v\,\ell_v) = p_T^v\,\Delta\ell_v + 2\,\nabla p_T^v\cdot\nabla\ell_v + \ell_v\,\Delta p_T^v. \tag{28}$$

Substitute $\nabla p_T^v = p_T^v\,J_T^v$, $\nabla\ell_v = J_T^v - J_S^v$, and the log-Laplacian identity $\Delta\log f = \Delta f/f - \|\nabla\log f\|^2$ to get

$$\Delta\ell_v = \frac{\Delta p_T^v}{p_T^v} - \|J_T^v\|^2 - \frac{\Delta p_S^v}{p_S^v} + \|J_S^v\|^2. \tag{29}$$

Inserting into (28):

$$\begin{aligned}
\Delta(p_T^v\,\ell_v) &= p_T^v\left[\frac{\Delta p_T^v}{p_T^v} - \|J_T^v\|^2 - \frac{\Delta p_S^v}{p_S^v} + \|J_S^v\|^2\right] + 2\,p_T^v\,J_T^v\cdot(J_T^v - J_S^v) + \ell_v\,\Delta p_T^v \\
&= \Delta p_T^v - p_T^v\|J_T^v\|^2 - \tfrac{p_T^v}{p_S^v}\Delta p_S^v + p_T^v\|J_S^v\|^2 \\
&\quad + 2p_T^v\big(\|J_T^v\|^2 - J_T^v\cdot J_S^v\big) + \ell_v\,\Delta p_T^v.
\end{aligned}$$

The score-norm terms collect into a perfect square:

$$-p_T^v\|J_T^v\|^2 + p_T^v\|J_S^v\|^2 + 2p_T^v\|J_T^v\|^2 - 2p_T^v\, J_T^v\cdot J_S^v = p_T^v\big(\|J_T^v\|^2 - 2J_T^v\cdot J_S^v + \|J_S^v\|^2\big)$$
$$= p_T^v\,\|J_T^v - J_S^v\|^2.$$

The remaining terms $\Delta p_T^v - \frac{p_T^v}{p_S^v}\Delta p_S^v + \ell_v\,\Delta p_T^v$ group as $(1+\ell_v)\Delta p_T^v - \frac{p_T^v}{p_S^v}\Delta p_S^v$, giving

$$\Delta(p_T^v\,\ell_v) = \underbrace{p_T^v\,\|J_T^v - J_S^v\|^2}_{\text{Fisher score gap}} + \underbrace{(1+\ell_v)\,\Delta p_T^v - \tfrac{p_T^v}{p_S^v}\,\Delta p_S^v}_{\text{residual}}.$$

Summing over v recovers $\mathcal{F}_t + R_t$ as introduced in Theorem 3, establishing the identity.

Vanishing of R_t. Decompose

$$R_t = \sum_v (1+\ell_v)\,\Delta p_T^v - \sum_v \tfrac{p_T^v}{p_S^v}\,\Delta p_S^v$$
$$= \underbrace{\sum_v \Delta p_T^v}_{=0} + \sum_v \ell_v\,\Delta p_T^v - \underbrace{\sum_v \Delta p_S^v}_{=0} - \sum_v \tfrac{p_T^v - p_S^v}{p_S^v}\,\Delta p_S^v.$$

The first and third sums vanish because $\Delta\sum_v p_M^v = \Delta(1) = 0$ for $M \in \{T, S\}$. Hence

$$R_t(x) = \sum_v \ell_v\,\Delta p_T^v - \sum_v \frac{p_T^v - p_S^v}{p_S^v}\,\Delta p_S^v. \tag{30}$$

Under (A1)–(A2), write $L_M = \max_v |\Delta p_M^v|$ and use $|\log p_T^v - \log p_S^v| \le |p_T^v - p_S^v|/\varepsilon_p$. Then

$$|R_t| \le \frac{L_T + L_S}{\varepsilon_p}\|p_T(\cdot,t) - p_S(\cdot,t)\|_1.$$

The residual is zero at output agreement and tends to zero as the outputs converge under uniform derivative bounds. □

Position-adaptive covariance. For $\xi \sim \mathcal{N}(0,\Sigma)$, replace Δ by $\Delta_\Sigma = \operatorname{tr}(\Sigma\nabla_x^2)$ in the same calculation:

$$\mathbb{E}D(x+\xi) = D(x) + \frac{1}{2}\sum_t(\mathcal{F}_{\Sigma,t} + R_{\Sigma,t}) + \mathcal{Q}_4, \tag{31}$$

where

$$\mathcal{F}_{\Sigma,t} = \sum_v p_T^v\|\Sigma^{1/2}(J_T^v - J_S^v)\|^2, \quad R_{\Sigma,t} = \sum_v \ell_v\Delta_\Sigma p_T^v - \sum_v \frac{p_T^v - p_S^v}{p_S^v}\Delta_\Sigma p_S^v.$$

The remainder satisfies $|\mathcal{Q}_4| \le M_4[(\operatorname{tr}\Sigma)^2 + 2\operatorname{tr}(\Sigma^2)]/24$. Thus the adaptive schedule weights the score-gap directions probed by the loss.

A.4 Proof of Corollary 4

Proof. Fix $x = x_i$, put $r = |\mathcal{R}_i|$, $F = \sum_t \mathcal{F}_t(x)$ and $R = \sum_t R_t(x)$. By (A3), $\lambda_{\max}(H_i) \le \operatorname{tr} H_i = F + R$. Taylor's theorem and the gradient bound (25) give

$$\sup_{\|\delta\|\le\epsilon} D(x+\delta) \le D(x) + \epsilon\sqrt{rF} + \epsilon\kappa_T(x)\sqrt{D(x)} + \frac{\epsilon^2}{2}\operatorname{tr} H_i + \frac{M_{3,i}\epsilon^3}{6}. \tag{32}$$

Young's inequality yields

$$\epsilon\sqrt{rF} \le \frac{r\epsilon^2}{\lambda\sigma^2} + \frac{\lambda\sigma^2}{4}F \le \frac{r\epsilon^2}{\lambda\sigma^2} + \frac{\lambda\sigma^2}{4}(\operatorname{tr} H_i + |R|), \tag{33}$$

$$\epsilon\kappa_T(x)\sqrt{D(x)} \le \frac{1}{2}D(x) + \frac{1}{2}\epsilon^2\kappa_T(x)^2. \tag{34}$$

Since $\epsilon^2 \le \lambda\sigma^2/2$, the two trace contributions satisfy

$$\left(\frac{\epsilon^2}{2} + \frac{\lambda\sigma^2}{4}\right) \operatorname{tr} H_i \le \frac{\lambda\sigma^2}{2} \operatorname{tr} H_i \le \lambda\big(\mathbb{E}D(x+\xi) - D(x)\big) + \frac{\lambda M_4 \sigma^4 m(m+2)}{24}. \tag{35}$$

Combining these inequalities and dropping $-\lambda D(x) \le 0$ proves Eq. (11), with

$$\mathcal{E}_i = \frac{\epsilon^2 \kappa_T(x_i)^2}{2} + \frac{\lambda\sigma^2}{4}\left|\sum_t R_t(x_i)\right| + \frac{M_{3,i}\epsilon^3}{6} + \frac{\lambda M_4 \sigma^4 m(m+2)}{24}. \tag{36}$$

Taking the maximum over i completes the proof. □

The radius condition and the explicit remainder quantify the range of this local bound. For a positive semidefinite covariance Σ, apply the same result to $u \mapsto D(x+\Sigma^{1/2}u)$ with unit Gaussian scale. The resulting neighborhood is $\{\Sigma^{1/2}u : \|u\| \le \epsilon\}$, possibly degenerate, with derivatives and curvature taken in u coordinates. For response-position-averaged KL, the proof uses $r = 1$, position-averaged F, R, D, and $\kappa_T = \sqrt{2}G_T/\varepsilon_p$.

A.5 Proof of Theorem 5

We restate the inner maximization of (13) for reference:

$$\max_{w \in \Delta^{B-1}} \sum_{i=1}^{B} w_i\, c_i + \tau_w\, H(w), \qquad H(w) = -\sum_i w_i \log w_i. \tag{37}$$

Up to a constant, this is the inner Lagrangian maximization for the KL-ball-constrained worst-case problem $\max_w\{\sum_i w_i c_i : D_{\mathrm{KL}}(w\|\mathbf{1}/B) \le \rho\}$ with a fixed positive multiplier τ_w, since $D_{\mathrm{KL}}(w\|\mathbf{1}/B) = \log B - H(w)$ (Duchi & Namkoong, 2021; Sagawa et al., 2020).

Proof. The Lagrangian for (37) subject to $\sum_i w_i = 1$ is

$$\mathcal{L}(w,\mu) = \sum_i w_i\, c_i - \tau_w \sum_i w_i \log w_i - \mu\Big(\sum_i w_i - 1\Big).$$

Stationarity in w_i:

$$\frac{\partial \mathcal{L}}{\partial w_i} = c_i - \tau_w(\log w_i + 1) - \mu = 0 \implies \log w_i = \frac{c_i - \mu - \tau_w}{\tau_w}.$$

Hence $w_i = \exp(-1 - \mu/\tau_w)\exp(c_i/\tau_w)$. Imposing $\sum_i w_i = 1$ fixes the normalizer $\exp(-1 - \mu/\tau_w) = 1/\sum_j \exp(c_j/\tau_w)$, so

$$w_i^* \;=\; \frac{\exp(c_i/\tau_w)}{\sum_{j=1}^{B}\exp(c_j/\tau_w)} \;=\; \operatorname{softmax}(c/\tau_w)_i. \tag{38}$$

For $\tau_w > 0$, strict concavity makes this the unique global maximum. Substituting $c_i \mapsto \mathrm{JSD}_i$ gives the proxy-score solution; multiplying by B recovers the mean-one weights in Eq. (15). □

Limits. As $\tau_w \to 0^+$, the entropy regularizer vanishes and w^* concentrates on $\arg\max_i c_i$, recovering hard worst-case mining. As $\tau_w \to \infty$, the entropy dominates and $w_i^* \to 1/B$, recovering uniform weighting (no DRO).

A.6 Minimax analysis of the position-adaptive noise schedule

For positive fixed scores a_j, consider

$$\min_{\sigma_j > 0} \max_j \frac{a_j}{\sigma_j^2} \quad \text{subject to} \quad \sum_j \sigma_j^2 = B_\sigma.$$

If the maximum ratio is q, then $\sigma_j^2 \ge a_j/q$, hence $q \ge \sum_j a_j/B_\sigma$. Equality is attained uniquely by $\sigma_j^2 = B_\sigma a_j/\sum_k a_k$. With $g_j = |s_{T,j} - s_{S,j}|$ and a linear model $a_j = \kappa_j g_j$, this gives $\sigma_j \propto \sqrt{\kappa_j g_j}$. A quadratic model $a_j \propto g_j^2$ instead gives $\sigma_j \propto g_j$, motivating Eq. (12). The choice of gap model is an allocation heuristic. For $\overline{g} > 0$, positional-mean normalization fixes the average scale before clipping, not the energy $\sum_j \sigma_j^2$; the cap bounds the applied scales, while zero-gap positions receive zero noise.

A.7 SHARED PERTURBATIONS ACROSS EMBEDDING SPACES

For model width d_M, draw $G_M \in \mathbb{R}^{d_M \times K}$ with i.i.d. standard Gaussian entries and set

$$U_{M,:,k} = G_{M,:,k}/\|G_{M,:,k}\|_2, \qquad A_M = \sqrt{d_M/K}\, U_M.$$

For the implementation, $K = 128$; the matrices are column-normalized, fixed, and cached, while each sample-position pair uses an independent shared $z_{ij} \sim \mathcal{N}(0, I_K)$ and position scale σ_{ij}.

For the sequence-level notation, set $P_M = I_L \otimes A_M$, $\xi = \mathrm{vec}((\sigma_j z_j)_{j=1}^L)$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_L^2) \otimes I_K$. Then

$$\xi_T = P_T \xi, \qquad \xi_S = P_S \xi, \qquad \mathbb{E}[\xi_T \xi_S^\top] = P_T \Sigma P_S^\top. \tag{39}$$

For $\widetilde{p}_M^v(u,t) = p_M^v(e_M + P_M u, t)$, the chain rule gives $\nabla_u \log \widetilde{p}_M^v = P_M^\top \nabla_{e_M} \log p_M^v$. Theorem 3 therefore compares these mapped scores, with covariance weighting as in Eq. (31); the width rescaling matches expected squared noise norm, not isotropic covariance. When $K < d_M$, the perturbed subspace has dimension at most K per position. Input derivatives hold the maps and the base-example noise scales fixed.

B IMPLEMENTATION DETAILS

B.1 COMPLEMENTARITY OF COMPONENTS

Noise KL targets local sensitivity differences, saliency weighting prioritizes attribution disagreement, and clean KL anchors pointwise matching. Section 4.3 ablates their contributions.

B.2 TRAINING PROCEDURE

Teacher saliency is precomputed once prior to training. During training, every N-th step computes the full SaKD loss (noise KL + DRO reweighting), while intermediate steps use standard KL. The detailed pseudocode is given in Algorithm 1.

B.3 COMPUTATIONAL COST

A SaKD step adds student-saliency computation and teacher/student forwards on perturbed embeddings, with backpropagation through the student noisy loss. These operations occur every N steps; teacher saliency is precomputed once. At $N = 5$, the reported wall-clock ratio is approximately $1.3\times$ relative to standard KD (Table 4).

Table 4: Training cost on Qwen3-8B→Qwen3-0.6B distillation (Dolly-15K, 10 epochs, single A100 80GB).

Method	Training time	vs. KD-KL	One-time setup
SFT	∼5.5 min	0.5× (no teacher fwd)	—
KD-KL	11 min	1.0× (baseline)	—
MiniLLM	11 min	1.0×	—
SeqKD	∼11 min	∼1.0×	teacher decoding over full dataset
GKD	∼40 min	∼3−5×	—
DistiLLM	∼11 min	∼1.0×	—
DA-KD	∼13 min	∼1.2×	per-epoch DDS scoring
SaKD	∼14 min	∼1.3× at N=5	teacher saliency