WHERE THE PREDICTION GOES MATTERS IN TOKEN-LEVEL KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Token-level knowledge distillation transfers a teacher's next-token distribution to a smaller student. Common distillation objectives measure how closely the student matches the teacher's next-token distribution at each prediction position. However, a smaller student cannot in general reproduce the teacher distribution exactly at every position, raising a question: when the student redistributes probability mass relative to the teacher, should the destination of this mass matter? In particular, should equal amounts of this mass be treated the same when they go to different tokens? We argue that their significance should depend on how the teacher relates those tokens across a population of contexts. To model these relations, we represent each token by a teacher-derived distribution over contexts. At each prediction position, we use the teacher and student next-token distributions to define two mixtures of the teacher-derived context distributions and compare them with the Cauchy–Schwarz divergence. We combine this divergence with a standard probability-matching objective to introduce residual composition distillation (RCD). Across two teacher–student model families and multiple evaluation sets, RCD consistently improves prediction quality and generation diversity, with a trade-off in generation quality.

1 INTRODUCTION

Logit-based knowledge distillation transfers a teacher's predictive distribution to a smaller student through its output probabilities (Hinton et al., 2015). In causal language models, token-level distillation compares next-token distributions over vocabularies of tens of thousands to over a hundred thousand outputs (Radford et al., 2019; Yang et al., 2025). Beyond the classical forward Kullback–Leibler (KL) divergence, recent methods measure the discrepancy between these distributions with reverse KL, skewed KL, or related divergences (Gu et al., 2024; Ko et al., 2024; 2025; Wang et al., 2025; Wu et al., 2025); other work applies these losses on sequences generated by the student, as in on-policy distillation (Gu et al., 2024; Agarwal et al., 2024; Xu et al., 2025).

A smaller student cannot in general reproduce the teacher distribution exactly at every prediction position, so it redistributes probability mass across tokens relative to the teacher. This raises a question: should the destination of this mass matter? Consider a teacher that assigns $(0.6, 0.2, 0.2)$ to *cat*, *kitten*, and *car*. Student A moves 0.1 from *cat* to *kitten*, giving $(0.5, 0.3, 0.2)$, and student B moves the same amount to *car*, giving $(0.5, 0.2, 0.3)$. The two students have the same KL divergence from the teacher (Figure 1(a)). However, after a prefix such as "*She adopted a*", where the teacher expects *cat*, we would easily accept *kitten* in its place, but hardly *car*.

We describe this difference using the teacher's predictions. We call the prefix at a response-token prediction position a context. At each context, the teacher assigns a probability to every token in the vocabulary. Conversely, for each vocabulary token, the teacher's predictions across contexts show where the teacher tends to assign probability to that token; we call this the token's teacher-conditioned context distribution. In Figure 1(b), the distributions for *cat* and *kitten* place mass on largely the same contexts, whereas the distribution for *car* places its mass on different ones. At each prediction position, we then define teacher and student context mixtures of these distributions, weighted by the teacher and student next-token distributions.

We compare the teacher and student context mixtures with the Cauchy–Schwarz divergence (Jenssen et al., 2006; Yu et al., 2024). At each prediction position, this divergence is computed from bilinear forms of the teacher and student next-token distributions in a teacher-derived Gram matrix, constructed

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.

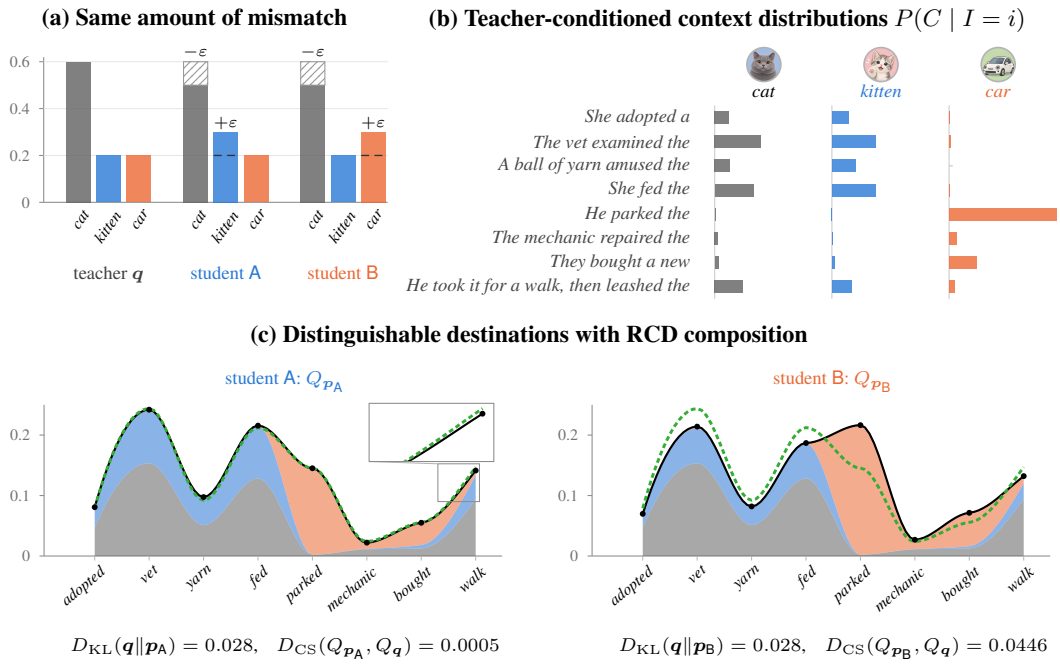


Figure 1: Why the destination of redistributed probability mass can matter in token-level knowledge distillation. (a) Teacher distribution $\boldsymbol{q}$ and two students that each move ε of probability away from *cat*: student A to *kitten* and student B to *car*. (b) Teacher-conditioned context distributions $P(C \mid I = i)$ for the three tokens, estimated from GPT-2 predictions on eight illustrative contexts. The *car* distribution clearly differs from the other two. (c) Context mixtures over the contexts of (b), each labeled by a keyword, obtained by weighting the distributions in (b) with the student probabilities (solid) and the teacher probabilities (green dashed), smoothed for visualization; shaded areas show the weighted contribution of *cat* (gray), *kitten* (blue), and *car* (orange); the inset zooms in near *walk*. The two students have the same KL divergence, but the composition term clearly distinguishes them: the mixture of student A nearly coincides with that of the teacher, whereas that of student B does not.

once before training from teacher predictions on training contexts. We call it the composition divergence and combine it with a probability-matching objective to obtain residual composition distillation (RCD), in which the composition divergence uses teacher-derived relations between tokens to take into consideration where the residual that the student cannot match exactly goes.

Our contributions are:

- We argue that when a student redistributes probability mass across tokens relative to the teacher, the destination of this mass should matter, depending on how the teacher relates those tokens across contexts, as illustrated in Figure 1.
- We construct a teacher-conditioned context distribution for each token. At each prediction position, we define two mixtures of these distributions weighted by the teacher and student next-token distributions, compare them with the Cauchy–Schwarz divergence, and combine this divergence with a probability-matching objective to introduce RCD.
- We implement RCD and evaluate it with two teacher–student model families on multiple evaluation sets, comparing it with token-level distillation baselines. Relative to the probability-matching objective it builds on, RCD consistently improves prediction quality and generation diversity, with a trade-off in generation quality.

2 Related Work and Preliminaries

2.1 Related work

Token-level distillation. Knowledge distillation was introduced for model compression (Bucilă et al., 2006; Hinton et al., 2015). In language models, token-level distillation commonly matches teacher and student next-token distributions using forward KL or related objectives, including reverse KL, skewed KL, and adaptive KL variants (Gu et al., 2024; Ko et al., 2024; 2025; Wang et al., 2025; Wu et al., 2025). Other work also trains on student-generated sequences, with the teacher providing token-level targets, through on-policy optimization (Gu et al., 2024; Agarwal et al., 2024), an adaptive off-policy approach (Ko et al., 2024), or interleaved sampling in which the student proposes tokens and the teacher replaces poorly ranked ones (Xu et al., 2025).

Relations between vocabulary outputs. Similar to RCD, other work relates distinct outputs so that a loss depends on where probability goes, but not through the teacher's predictions across contexts. Prior approaches have used inter-class correlations in vision (Huang et al., 2025). Optimal-transport approaches provide another related perspective. For language-model training, EMO replaces the forward cross-entropy with an earth mover distance whose transport cost between two tokens is the cosine distance between their output embeddings (Ren et al., 2024). SinKD measures the teacher–student discrepancy with the Sinkhorn distance, computed across the samples of a batch instead of within each sample (Cui et al., 2024). In vision, Wasserstein distillation performs cross-category comparison of probabilities to leverage interrelations among categories, over output spaces much smaller than a language-model vocabulary (Lv et al., 2024). Concurrently, Verbii & Lee (2026) study multi-temperature teacher views and transport-based objectives under an embedding-based ground cost, solved iteratively at each evaluated position. Closest to us, CSD relates tokens through relative logit differences (Kim et al., 2026). Our focus is the composition effect: we pair RCD with various probability-matching objectives to examine its consistency and trade-offs.

2.2 Token-level distillation

At prediction position t, the model predicts the next token W_t from the preceding prefix $W_{<t}$. Let $x = W_{<t}$ denote this prefix, and let f_ϕ and f_θ denote a fixed teacher and a trainable student sharing a vocabulary $\mathcal{V}$ of size m. At prefix x, their logits $g^T, g^S \in \mathbb{R}^m$ define, at a distillation temperature $T_{\mathrm{KD}} > 0$ that is typically selected in advance, the next-token distributions

$$\boldsymbol{q} = \mathrm{softmax}\left(\frac{g^T}{T_{\mathrm{KD}}}\right), \qquad \boldsymbol{p} = \mathrm{softmax}\left(\frac{g^S}{T_{\mathrm{KD}}}\right), \tag{1}$$

which lie in the probability simplex Δ^{m-1}.

Standard token-level distillation matches these distributions with the forward KL divergence,

$$\mathcal{L}_{\mathrm{KL}}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i}. \tag{2}$$

More generally, discrete f-divergences take the form

$$D_f(\boldsymbol{q} \| \boldsymbol{p}) = \sum_{i=1}^{m} p_i \, f\left(\frac{q_i}{p_i}\right), \tag{3}$$

where f is convex and $f(1) = 0$. Forward and reverse KL (Gu et al., 2024) are instances of this form. Other token-level objectives, including skewed KL (Ko et al., 2024) and the α–β divergence (Wang et al., 2025), likewise compare teacher and student probabilities at corresponding vocabulary indices.

Figure 1(a) illustrates a simple phenomenon. Suppose the teacher assigns equal probability to *kitten* and *car*. Student A moves mass ε from *cat* to *kitten*, whereas student B moves the same mass from *cat* to *car*. For an f-divergence,

$$D_f(\boldsymbol{q} \| \boldsymbol{p}_{\mathsf{A}}) = D_f(\boldsymbol{q} \| \boldsymbol{p}_{\mathsf{B}}).$$

In this case, f-divergences besides forward and reverse KL, such as total variation, squared Hellinger, and Jensen–Shannon divergence, also do not distinguish between the two destinations.

2.3 TEACHER-INDUCED JOINT DISTRIBUTION

Let C be a discrete context variable over prediction prefixes from the training data, with distribution $P_{\mathrm{ctx}}(c) = P(C = c)$. For example, in Figure 1(b), "*She adopted a*" is one realization of C. Given $C = c$, let $I \in \mathcal{V}$ be an auxiliary token drawn from the teacher next-token distribution $P(I \mid C = c)$, i.e., $\boldsymbol{q}$ evaluated at prefix c. Together they define the joint distribution

$$P(C = c, I = i) = P_{\mathrm{ctx}}(c)\, P(I = i \mid C = c) = P(C = c \mid I = i)\, P(I = i), \tag{4}$$

where

$$P(I = i) = \sum_{c} P_{\mathrm{ctx}}(c)\, P(I = i \mid C = c)$$

is the marginal probability of token i under the teacher across the context population. By Bayes' rule, this joint distribution associates each vocabulary item i with the teacher-conditioned context distribution $P(C \mid I = i)$, which describes the contexts in which the teacher tends to assign probability to item i.

3 RESIDUAL COMPOSITION DISTILLATION

We propose a distillation objective that distinguishes where redistributed probability mass goes, using relations derived from the teacher's predictions across contexts. Each vocabulary item is represented by its teacher-conditioned context distribution, and teacher and student predictions are compared through mixtures of these distributions. We combine this composition divergence with a standard probability-matching objective.

3.1 TEACHER-INDUCED CONTEXT MIXTURES

Rearranging Equation (4) gives the teacher-conditioned context distribution

$$P(C = c \mid I = i) = \frac{P_{\mathrm{ctx}}(c)\, P(I = i \mid C = c)}{P(I = i)}. \tag{5}$$

Figure 1(b) shows three examples: the distributions for *cat* and *kitten* place mass on largely the same contexts, whereas the distribution for *car* concentrates on different ones.

Its density ratio with respect to P_{ctx}:

$$h_i(c) \coloneqq \frac{P(C = c \mid I = i)}{P_{\mathrm{ctx}}(c)} = \frac{P(I = i \mid C = c)}{P(I = i)}. \tag{6}$$

Thus $h_i(c)$ gives the relative mass of context c under $P(C \mid I = i)$ compared with P_{ctx}.

At a prediction position, we define two mixtures over the teacher-conditioned context distributions $P(C \mid I = i)$ using the student and teacher next-token probabilities $\boldsymbol{p}$ and $\boldsymbol{q}$ as mixture weights:

$$\begin{aligned} Q_{\boldsymbol{p}}(c) &\coloneqq \sum_{i=1}^{m} p_i P(C = c \mid I = i) = P_{\mathrm{ctx}}(c) \sum_{i=1}^{m} p_i h_i(c), \\ Q_{\boldsymbol{q}}(c) &\coloneqq \sum_{i=1}^{m} q_i P(C = c \mid I = i) = P_{\mathrm{ctx}}(c) \sum_{i=1}^{m} q_i h_i(c). \end{aligned} \tag{7}$$

Both are distributions on the same context space, as illustrated in Figure 1(c).

Equation (7) also makes the effect of moving probability between tokens explicit.

Remark 1 (Effect of redistributing probability mass)**.** Suppose the student moves probability mass ε from token j to token i, leaving all other token probabilities unchanged; that is, $p'_j = p_j - \varepsilon$ and $p'_i = p_i + \varepsilon$. By Equation (7),

$$Q_{\boldsymbol{p}'}(c) - Q_{\boldsymbol{p}}(c) = \varepsilon \left[P(C = c \mid I = i) - P(C = c \mid I = j) \right]. \tag{8}$$

Thus the effect depends on the destination: moving mass from *cat* to *kitten* changes $Q_{\boldsymbol{p}}$ only slightly, because their teacher-conditioned context distributions are similar, whereas moving the same mass to *car* changes it substantially (Figure 1(c)).

3.2 COMPOSITION DIVERGENCE AND RCD

We compare the teacher and student context mixtures using the Cauchy–Schwarz divergence (Jenssen et al., 2006; Kampa et al., 2011). For functions on the context population, write

$$\langle f, g\rangle_{P_{\mathrm{ctx}}} \coloneqq \mathbb{E}_{C\sim P_{\mathrm{ctx}}}[f(C)g(C)], \qquad \|f\|^2_{P_{\mathrm{ctx}}} = \langle f, f\rangle_{P_{\mathrm{ctx}}}. \tag{9}$$

Definition 1 (Context-mixture composition divergence)**.** By Equation (7), the density ratios of the two context mixtures with respect to P_{ctx} are

$$\frac{Q_{\boldsymbol{p}}(c)}{P_{\mathrm{ctx}}(c)} = \sum_{i=1}^{m} p_i h_i(c), \qquad \frac{Q_{\boldsymbol{q}}(c)}{P_{\mathrm{ctx}}(c)} = \sum_{i=1}^{m} q_i h_i(c). \tag{10}$$

These ratios give the relative mass of each context under the two mixtures compared with P_{ctx}. We define their composition divergence as

$$D_{\mathrm{CS},P_{\mathrm{ctx}}}(Q_{\boldsymbol{p}}, Q_{\boldsymbol{q}}) \coloneqq -\log \frac{\langle Q_{\boldsymbol{p}}/P_{\mathrm{ctx}},\, Q_{\boldsymbol{q}}/P_{\mathrm{ctx}}\rangle^2_{P_{\mathrm{ctx}}}}{\|Q_{\boldsymbol{p}}/P_{\mathrm{ctx}}\|^2_{P_{\mathrm{ctx}}}\, \|Q_{\boldsymbol{q}}/P_{\mathrm{ctx}}\|^2_{P_{\mathrm{ctx}}}}. \tag{11}$$

By the Cauchy–Schwarz inequality, $D_{\mathrm{CS},P_{\mathrm{ctx}}} \geq 0$, with equality if and only if $Q_{\boldsymbol{p}} = Q_{\boldsymbol{q}}$.

By Equation (10), the two mixture density ratios are linear combinations of the token-specific density ratios h_i. We collect their pairwise inner products in the symmetric positive-semidefinite Gram matrix $\boldsymbol{K} \in \mathbb{R}^{m\times m}$ with entries

$$K_{ij} \coloneqq \langle h_i, h_j\rangle_{P_{\mathrm{ctx}}}. \tag{12}$$

With this definition, the inner products between the mixture density ratios reduce to

$$\left\langle \frac{Q_{\boldsymbol{p}}}{P_{\mathrm{ctx}}}, \frac{Q_{\boldsymbol{q}}}{P_{\mathrm{ctx}}} \right\rangle_{P_{\mathrm{ctx}}} = \boldsymbol{p}^\top \boldsymbol{K} \boldsymbol{q}, \tag{13}$$

and

$$\left\| \frac{Q_{\boldsymbol{p}}}{P_{\mathrm{ctx}}} \right\|^2_{P_{\mathrm{ctx}}} = \boldsymbol{p}^\top \boldsymbol{K} \boldsymbol{p}, \qquad \left\| \frac{Q_{\boldsymbol{q}}}{P_{\mathrm{ctx}}} \right\|^2_{P_{\mathrm{ctx}}} = \boldsymbol{q}^\top \boldsymbol{K} \boldsymbol{q}. \tag{14}$$

Substituting these expressions into Equation (11) yields

$$D_{\mathrm{CS},P_{\mathrm{ctx}}}(Q_{\boldsymbol{p}}, Q_{\boldsymbol{q}}) = -\log \frac{(\boldsymbol{p}^\top \boldsymbol{K} \boldsymbol{q})^2}{(\boldsymbol{p}^\top \boldsymbol{K} \boldsymbol{p})(\boldsymbol{q}^\top \boldsymbol{K} \boldsymbol{q})}. \tag{15}$$

Further properties of $\boldsymbol{K}$ are given in Section B.

At each prediction position, we define the composition loss as

$$\mathcal{L}_{\mathrm{comp}}(\boldsymbol{p}, \boldsymbol{q}) = D_{\mathrm{CS},P_{\mathrm{ctx}}}\big(Q_{\boldsymbol{p}}, Q_{\boldsymbol{q}}\big). \tag{16}$$

RCD combines this term with a probability-matching objective:

$$\mathcal{L}_{\mathrm{RCD}}(\boldsymbol{p}, \boldsymbol{q}) = (1-\eta)\mathcal{L}_{\mathrm{match}}(\boldsymbol{p}, \boldsymbol{q}) + \eta \mathcal{L}_{\mathrm{comp}}(\boldsymbol{p}, \boldsymbol{q}), \tag{17}$$

where $\mathcal{L}_{\mathrm{match}}$ is a token-level objective such as those in Section 2.2 and $\eta \in [0, 1]$ weights the composition term.

3.3 VOCABULARY-SCALE CONSTRUCTION OF THE TEACHER-DERIVED RELATION

The composition divergence in Equation (15) depends on the teacher-derived Gram matrix $\boldsymbol{K}$. We construct $\boldsymbol{K}$ from frozen teacher predictions on a set of contexts sampled from the training data, and then describe how the composition term is evaluated efficiently at vocabulary scale.

Let $\mathcal{C} = \{C_1, \ldots, C_N\}$ be prediction contexts sampled from the training data by selecting one valid response-token position per usable example and taking its preceding prefix. We evaluate the frozen teacher at these contexts; the observed response token itself is not used. On the sampled set, we define

$$P_{\mathcal{C}}(C = C_n) = \frac{1}{N}, \qquad n = 1, \ldots, N. \tag{18}$$

Teacher-conditioned context distributions can differ substantially in how spread out they are. Following SNE (Hinton & Roweis, 2002), we fix a global perplexity κ and use a token-specific temperature τ_i to normalize the teacher log-probabilities across contexts:

$$\pi_{ni}(\tau_i) = \frac{P(I = i \mid C = C_n)^{1/\tau_i}}{\sum_{r=1}^{N} P(I = i \mid C = C_r)^{1/\tau_i}}. \tag{19}$$

When $\tau_i = 1$, $\pi_{\cdot i}$ is exactly the conditional distribution over the sampled contexts given $I = i$ under Equation (18). We choose $\tau_i > 0$ so that

$$\exp H(\pi_{\cdot i}(\tau_i)) = \kappa,$$

where H is the Shannon entropy (Section C). With respect to $P_{\mathcal{C}}$, the corresponding density ratio is

$$h_i^{\mathcal{C}}(C_n) = \frac{\pi_{ni}}{1/N} = N\pi_{ni}. \tag{20}$$

Using these density ratios in Equation (12) gives

$$K_{ij}^{\mathcal{C}} = \frac{1}{N} \sum_{n=1}^{N} h_i^{\mathcal{C}}(C_n) h_j^{\mathcal{C}}(C_n) = N \sum_{n=1}^{N} \pi_{ni}\pi_{nj}. \tag{21}$$

The relation is defined over the full vocabulary. At each prediction position, we evaluate the composition term on the highest-probability vocabulary items under either the teacher or the student:

$$A = \mathrm{Top}_b(\boldsymbol{q}) \cup \mathrm{Top}_b(\boldsymbol{p}), \tag{22}$$

with a predefined b. We renormalize both distributions on this set:

$$\widetilde{p}_i = \frac{p_i}{\sum_{j \in A} p_j}, \qquad \widetilde{q}_i = \frac{q_i}{\sum_{j \in A} q_j}, \qquad i \in A. \tag{23}$$

During training, the composition loss is evaluated using $\widetilde{\boldsymbol{p}}$ and $\widetilde{\boldsymbol{q}}$ on A, while the probability-matching loss remains defined over the full vocabulary. Section G discusses the choice of b.

4 Experiments

Our experiments examine how RCD with the teacher-derived matrix $\boldsymbol{K}$ affects training, prediction, and generation. We ask three questions: (i) how $\boldsymbol{K}$ changes the gradient field; (ii) what information $\boldsymbol{K}$ captures from real teacher predictions; and (iii) whether our proposed composition term consistently changes training, prediction, and generation across model families and probability-matching objectives.

4.1 Experimental setup

Models and training. We use the Dolly instruction-distillation setting of MiniLLM/DistiLLM (Gu et al., 2024; Ko et al., 2024) with the CSD training and evaluation protocol (Kim et al., 2026) on Databricks Dolly 15K (Conover et al., 2023). We instantiate $\mathcal{L}_{\mathrm{match}}$ in Equation (17) with three probability-matching objectives: forward KL (FKL), skewed reverse KL (SRKL) (Ko et al., 2024), and the α–β divergence (α–β) (Wang et al., 2025). We denote RCD composed with X by RCD (X), where X is one of these three objectives. We use the composition weight $\eta = 0.1$ by default in both model families and evaluate sensitivity to $\eta \in \{0.1, 0.3, 0.5\}$ on GPT-2 (Section 4.5). Unless stated otherwise, $T_{\mathrm{KD}} = 1$, $b = 100$, and $\kappa = 32$.

We use two teacher–student families, GPT-2 XL to GPT-2 124M (Radford et al., 2019) and OPT-1.3B to OPT-125M (Zhang et al., 2022), with vocabulary sizes $m = 50{,}257$ and $m = 50{,}272$, respectively. Within each family, teacher and student share the same tokenizer. For ablation and sensitivity analyses, we use GPT-2. Further details are summarized in Section E.

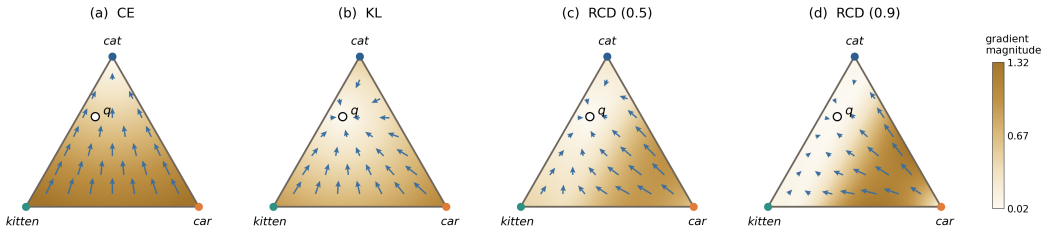


Figure 2: Gradient fields on the *cat–kitten–car* simplex. (a) The cross-entropy (CE) field points toward the one-hot *cat* target. (b) The FKL field points toward the teacher distribution $\boldsymbol{q}$. (c, d) RCD (FKL) with composition weights $\eta = 0.5$ and 0.9. As η increases, RCD produces larger gradients over much of the *car*-heavy region and smaller gradients along the *cat–kitten* direction; the gradient magnitude decreases again close to the *car* vertex. Arrow lengths are rescaled within each panel; shading shows gradient magnitude on a shared scale. Section D gives the analytic details.

Evaluation. We evaluate prediction quality, generation quality, and generation diversity. On the reference responses, we report response NLL and ECE, the teacher-forced calibration error of Kim et al. (2026) relative to the teacher probabilities. For generation quality, we report lexical and semantic agreement with the references: ROUGE-L, best-of-five ROUGE-L (Oracle@5), mean BERTScore-F1, and worst-of-five BERTScore-F1 (min@5). For diversity, we report Self-BLEU (S-BLEU) and Distinct-2 (Dist-2) across the five samples. Evaluation uses DollyEval (Conover et al., 2023), SelfInst (Wang et al., 2023), VicunaEval (Chiang et al., 2023), and S-NI (Wang et al., 2022). Detailed definitions and aggregation are given in Section E.5.

Baselines. We compare against supervised fine-tuning (SFT) and four token-level distillation objectives: FKL, SRKL (Ko et al., 2024), α–β (Wang et al., 2025), and CSD (Kim et al., 2026). We also pair each RCD (X) with X to examine the effect of the composition term.

4.2 Toy example: how RCD changes the gradient

We visualize how RCD changes the optimization field on a three-output simplex, where the field is the gradient with respect to the student logits. We continue with *cat*, *kitten*, and *car*, choose the teacher distribution $\boldsymbol{q} = (0.6, 0.3, 0.1)$, and, for clarity, use the discrete relation matrix $\boldsymbol{K} = \begin{pmatrix} 1 & 2/3 & 0 \\ 2/3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ constructed in Section D.

In this setting, $K_{\text{cat,kitten}} = K_{\text{kitten,cat}} = 2/3$ and all other off-diagonal entries are zero. Relative to FKL, RCD therefore flattens the field along the *cat–kitten* direction and strengthens it over much

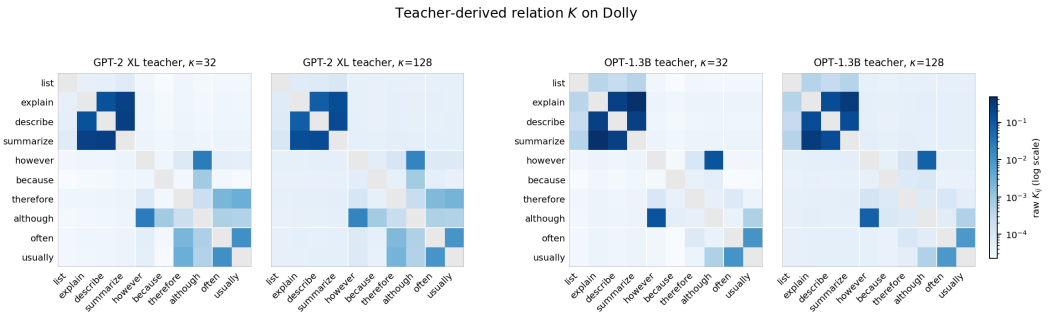


Figure 3: Submatrices of $\boldsymbol{K}^{\mathcal{C}}$ for selected tokens. Values are shown on a logarithmic color scale with the diagonal masked. The four panels correspond to the two teacher models and perplexities $\kappa \in \{32, 128\}$.

Table 1: Prediction quality, generation quality, and generation diversity for the two teacher–student families. Values are averaged over DollyEval, SelfInst, VicunaEval and S-NI within each seed and reported as mean±std over three seeds. Per-set results are in Section F; the best value in each column within a family is bold. Shaded rows mark the SRKL pair.

	Prediction quality		Generation quality				Generation diversity	
					BERTScore-F1 ↑			
Method	NLL ↓	ECE ↓	ROUGE-L ↑	Oracle@5 ↑	mean	min@5	S-BLEU ↓	Dist-2 ↑
GPT-2								
SFT	$5.83_{\pm 0.023}$	$0.3029_{\pm 0.0021}$	$16.19_{\pm 0.06}$	$22.79_{\pm 0.22}$	$0.834_{\pm 0.0003}$	$0.817_{\pm 0.0006}$	$0.209_{\pm 0.010}$	$0.783_{\pm 0.007}$
FKL	$4.61_{\pm 0.008}$	$0.2522_{\pm 0.0007}$	$16.49_{\pm 0.12}$	$23.89_{\pm 0.22}$	$0.834_{\pm 0.0004}$	$0.815_{\pm 0.0009}$	$0.151_{\pm 0.002}$	$0.825_{\pm 0.002}$
SRKL	$6.52_{\pm 0.029}$	$0.3240_{\pm 0.0029}$	$19.35_{\pm 0.49}$	$25.40_{\pm 0.42}$	$0.838_{\pm 0.0008}$	$0.822_{\pm 0.0017}$	$0.424_{\pm 0.014}$	$0.555_{\pm 0.012}$
α–β	$8.69_{\pm 0.021}$	$0.3631_{\pm 0.0012}$	$\mathbf{20.05}_{\pm 0.13}$	$24.87_{\pm 0.11}$	$0.838_{\pm 0.0009}$	$0.823_{\pm 0.0016}$	$0.568_{\pm 0.014}$	$0.450_{\pm 0.005}$
CSD	$5.95_{\pm 0.043}$	$0.2967_{\pm 0.0016}$	$19.25_{\pm 0.30}$	$25.50_{\pm 0.32}$	$0.838_{\pm 0.0003}$	$0.820_{\pm 0.0002}$	$0.367_{\pm 0.010}$	$0.588_{\pm 0.006}$
RCD (FKL)	$4.49_{\pm 0.034}$	$0.2466_{\pm 0.0023}$	$16.37_{\pm 0.11}$	$23.82_{\pm 0.15}$	$0.834_{\pm 0.0004}$	$0.814_{\pm 0.0005}$	$\mathbf{0.148}_{\pm 0.001}$	$\mathbf{0.829}_{\pm 0.004}$
RCD (SRKL)	$\mathbf{4.41}_{\pm 0.858}$	$\mathbf{0.2445}_{\pm 0.0346}$	$18.01_{\pm 0.29}$	$25.22_{\pm 0.83}$	$0.837_{\pm 0.0011}$	$0.818_{\pm 0.0028}$	$0.213_{\pm 0.021}$	$0.755_{\pm 0.005}$
RCD (α–β)	$6.39_{\pm 0.108}$	$0.3278_{\pm 0.0001}$	$19.88_{\pm 0.26}$	$\mathbf{25.63}_{\pm 0.41}$	$\mathbf{0.840}_{\pm 0.0001}$	$\mathbf{0.824}_{\pm 0.0001}$	$0.439_{\pm 0.020}$	$0.547_{\pm 0.015}$
OPT								
SFT	$6.17_{\pm 0.011}$	$0.2940_{\pm 0.0012}$	$13.50_{\pm 0.08}$	$19.75_{\pm 0.28}$	$0.824_{\pm 0.0003}$	$0.807_{\pm 0.0002}$	$0.138_{\pm 0.002}$	$0.845_{\pm 0.001}$
FKL	$5.12_{\pm 0.016}$	$0.2492_{\pm 0.0023}$	$15.13_{\pm 0.14}$	$22.29_{\pm 0.17}$	$0.829_{\pm 0.0005}$	$0.809_{\pm 0.0003}$	$0.146_{\pm 0.007}$	$0.840_{\pm 0.004}$
SRKL	$6.13_{\pm 0.007}$	$0.3125_{\pm 0.0009}$	$17.20_{\pm 0.49}$	$23.96_{\pm 0.43}$	$0.832_{\pm 0.0008}$	$0.813_{\pm 0.0012}$	$0.285_{\pm 0.006}$	$0.654_{\pm 0.009}$
α–β	$10.05_{\pm 0.020}$	$0.4343_{\pm 0.0012}$	$\mathbf{17.86}_{\pm 0.51}$	$22.98_{\pm 0.72}$	$0.828_{\pm 0.0003}$	$0.811_{\pm 0.0013}$	$0.530_{\pm 0.015}$	$0.440_{\pm 0.009}$
CSD	$5.26_{\pm 0.011}$	$0.2558_{\pm 0.0014}$	$15.85_{\pm 0.39}$	$22.80_{\pm 0.31}$	$0.828_{\pm 0.0011}$	$0.806_{\pm 0.0017}$	$0.185_{\pm 0.003}$	$0.737_{\pm 0.005}$
RCD (FKL)	$5.04_{\pm 0.009}$	$0.2425_{\pm 0.0004}$	$15.02_{\pm 0.18}$	$22.43_{\pm 0.34}$	$0.828_{\pm 0.0004}$	$0.808_{\pm 0.0004}$	$0.146_{\pm 0.002}$	$0.844_{\pm 0.002}$
RCD (SRKL)	$\mathbf{4.97}_{\pm 0.194}$	$\mathbf{0.2334}_{\pm 0.0019}$	$14.74_{\pm 0.10}$	$21.88_{\pm 0.21}$	$0.826_{\pm 0.0005}$	$0.807_{\pm 0.0003}$	$\mathbf{0.135}_{\pm 0.003}$	$\mathbf{0.849}_{\pm 0.005}$
RCD (α–β)	$6.41_{\pm 0.122}$	$0.3507_{\pm 0.0070}$	$17.72_{\pm 0.18}$	$\mathbf{24.03}_{\pm 0.15}$	$\mathbf{0.833}_{\pm 0.0003}$	$\mathbf{0.816}_{\pm 0.0004}$	$0.337_{\pm 0.034}$	$0.600_{\pm 0.029}$

of the *car*-heavy region (Figure 2). Both effects grow with η, and the gradient magnitude decreases again close to the *car* vertex (Section D).

4.3 What the teacher-derived relation captures

Figure 3 visualizes submatrices of the matrix $\boldsymbol{K}^{\mathcal{C}}$ used for distillation, constructed separately from each teacher on the sampled Dolly contexts. Each submatrix is restricted to common single-token words from three interpretable groups: task verbs (*list*, *explain*, *describe*, *summarize*), discourse connectives (*however*, *because*, *therefore*, *although*), and frequency adverbs (*often*, *usually*).

The pairwise entries are clearly nonuniform. *Explain*, *describe*, and *summarize* have relatively large pairwise values, as do *however* and *although*, and *often* and *usually*, while most cross-group pairs have much smaller values. The submatrices show the same pattern for both teacher models and for both perplexities $\kappa = 32$ and 128.

4.4 Distillation results

Table 1 reports three groups of metrics for both teacher–student families: prediction quality, generation quality, and generation diversity. Together, they examine the trade-offs of RCD compared with the baselines.

For prediction quality, RCD (SRKL) attains the lowest NLL and ECE in both families, and relative to each X, RCD (X) lowers both metrics in all six family–objective pairs. For generation quality, the cost appears mainly in lexical agreement: mean ROUGE-L decreases in all six pairs, by amounts that vary across objectives, whereas Oracle@5 and the BERTScore metrics change little and without a consistent direction. For generation diversity, RCD (X) raises Dist-2 and lowers S-BLEU relative to X (for FKL on OPT, S-BLEU is equal at the reported precision); the gain is small for FKL and large for SRKL and α–β; for SRKL, Dist-2 rises from 0.555 to 0.755 on GPT-2 and from 0.654 to 0.849 on OPT. Across both teacher–student families, these aggregated results show the same pattern for all three metric groups.

Table 2 further analyzes these results through a seed-paired comparison between RCD (X) and X. All metrics of prediction quality and generation diversity favor RCD, with pooled CIs entirely above

zero, consistently across both families. Within generation quality, only mean ROUGE-L decreases clearly, while the CIs of Oracle@5 and both BERTScore metrics include zero. In summary, RCD improves prediction quality and generation diversity; in generation quality, its cost lies mainly in average lexical agreement, while best-of-five and semantic quality do not significantly worsen.

Table 2: Seed-paired comparison between RCD (X) and X for each probability-matching objective X. A positive sign indicates a result in favor of RCD. For each metric, the GPT-2 and OPT columns report one value per family, averaged over the three objectives; pooled 95% CIs use a cluster bootstrap over the six family–objective pairs. Bold pooled values have CIs entirely above zero; the shaded row has a CI entirely below zero.

	Δ by family and pooled, + favors RCD				
Metric	GPT-2	OPT	Pooled (95% CI)	Fav./6	Fav./18
Prediction quality					
NLL	+1.50	+1.63	**+1.57** [+0.61, +2.59]	6/6	18/18
ECE	+0.0401	+0.0565	**+0.0483** [+0.0206, +0.0735]	6/6	18/18
Generation quality					
ROUGE-L	−0.55	−0.90	−0.72 [−1.50, −0.13]	0/6	3/18
Oracle@5	+0.17	−0.30	−0.06 [−0.95, +0.61]	3/6	10/18
BERTScore-F1 mean	+0.0001	0.0000	+0.0000 [−0.0026, +0.0026]	2/6	8/18
BERTScore-F1 min@5	−0.0013	−0.0006	−0.0010 [−0.0036, +0.0018]	2/6	5/18
Generation diversity					
S-BLEU	+0.115	+0.114	**+0.114** [+0.047, +0.178]	6/6	17/18
Dist-2	+0.100	+0.120	**+0.110** [+0.046, +0.174]	6/6	17/18

4.5 Sensitivity to the composition weight

We analyze the sensitivity of RCD to the composition weight η. Table 3 varies $\eta \in \{0.1, 0.3, 0.5\}$ for RCD (SRKL) on GPT-2, and we keep all other hyperparameters and training settings as in the main experiments.

Table 3: Sensitivity analysis of the composition weight η for RCD (SRKL) on GPT-2. All other settings follow the main protocol.

	Prediction quality		Generation quality				Generation diversity	
					BERTScore-F1 ↑			
η	NLL ↓	ECE ↓	ROUGE-L ↑	Oracle@5 ↑	mean	min@5	S-BLEU ↓	Dist-2 ↑
0.1	$4.41_{\pm 0.858}$	$0.2445_{\pm 0.0346}$	$18.01_{\pm 0.29}$	$25.22_{\pm 0.83}$	$0.837_{\pm 0.0011}$	$0.818_{\pm 0.0028}$	$0.213_{\pm 0.021}$	$0.755_{\pm 0.005}$
0.3	$4.51_{\pm 0.016}$	$0.2293_{\pm 0.0005}$	$15.55_{\pm 0.10}$	$22.97_{\pm 0.05}$	$0.827_{\pm 0.0006}$	$0.805_{\pm 0.0007}$	$0.154_{\pm 0.005}$	$0.847_{\pm 0.003}$
0.5	$4.31_{\pm 0.007}$	$0.2129_{\pm 0.0015}$	$13.99_{\pm 0.14}$	$21.46_{\pm 0.15}$	$0.819_{\pm 0.0007}$	$0.793_{\pm 0.0010}$	$0.121_{\pm 0.005}$	$0.886_{\pm 0.005}$

Increasing η places more weight on the composition term. As η increases, every metric except NLL changes monotonically. For prediction quality, ECE continues to decrease, while NLL changes only slightly. For generation quality, all four metrics decrease; unlike in the main comparison, the decrease now extends beyond mean ROUGE-L to Oracle@5 and both BERTScore metrics. For generation diversity, S-BLEU decreases and Dist-2 increases. Larger values of η therefore shift the trade-off further toward diversity, at an increasingly clear cost in generation quality.

5 Conclusion

We asked whether the destination of probability mass redistributed across tokens relative to the teacher should matter when a student cannot reproduce the teacher distribution exactly, and argued that it should depend on how the teacher relates the tokens across contexts. Accordingly, RCD combines our proposed composition term with a standard probability-matching objective. Across model families and objectives, this term consistently improves prediction quality and generation diversity, with a trade-off in generation quality.

USE OF GENERATIVE AI

Generative AI tools were used to refine the language and presentation of the manuscript and to provide feedback on the manuscript. All AI-assisted outputs were read, checked, and revised by the authors, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All mathematical constructions and derivations are given in Section 3 and Sections A to D. The implementation follows these definitions exactly. Section E reports the data, model configurations, optimization settings, hyperparameters, evaluation protocol, and statistical aggregation used in all experiments, and Section F reports per-set results. Unless stated otherwise, experiments use the default settings specified in the paper. We will release the training and evaluation code upon acceptance.

ETHICS STATEMENT

This work uses only publicly released datasets and publicly available model checkpoints, all of which are cited in the paper. It does not involve human subjects, crowdsourcing, or the collection of private or personally identifiable data.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Cristian Bucilă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, 2006. ISBN 1595933395. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks Blog, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkhorn distance minimization for knowledge distillation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14846–14858, May 2024. URL https://aclanthology.org/2024.lrec-main.1293/.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, volume 15, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. DIST+: Knowledge distillation from a stronger adaptive teacher. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7): 5571–5585, 2025. doi: 10.1109/TPAMI.2025.3554235. URL https://doi.org/10.1109/TPAMI.2025.3554235.

Robert Jenssen, Jose C. Príncipe, Deniz Erdogmus, and Torbjørn Eltoft. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006. doi: 10.1016/j.jfranklin.2006.03.018. URL https://doi.org/10.1016/j.jfranklin.2006.03.018.

Kittipat Kampa, Erion Hasanbelliu, and Jose C. Principe. Closed-form Cauchy–Schwarz PDF divergence for mixture of Gaussians. In *The 2011 International Joint Conference on Neural Networks*, pp. 2578–2585, 2011. doi: 10.1109/IJCNN.2011.6033555. URL https://doi.org/10.1109/IJCNN.2011.6033555.

Yeongmin Kim, Donghyeok Shin, Mina Kang, Byeonghu Na, and Il-Chul Moon. Distillation of large language models via concrete score matching. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bZBJFrxH1H.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Forty-first International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895, 2024. URL https://arxiv.org/abs/2402.03898.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=rc65N9xIrY.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, June 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, July 2004. URL https://aclanthology.org/W04-1013/.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals Kullback–Leibler divergence for knowledge distillation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 65445–65475, 2024. doi: 10.52202/079017-2090. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/78526d7ad4a2532bd91416e948b9644c-Paper-Conference.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, July 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Siyu Ren, Zhiyong Wu, and Kenny Q. Zhu. EMO: Earth mover distance optimization for auto-regressive language modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.04691.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, April 2021. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25/.

Gordei Verbii and Juho Lee. The distributional view of knowledge distillation, 2026. URL https://arxiv.org/abs/2608.15215.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vt65VjJakt.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, December 2022. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340/.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, July 2023. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

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 Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, January 2025. URL https://aclanthology.org/2025.coling-main.383/.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11325.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shujian Yu, Xi Yu, Sigurd Løkse, Robert Jenssen, and Jose C Príncipe. Cauchy–Schwarz divergence information bottleneck for regression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7wY67ZDQTE.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 1097–1100, 2018. ISBN 9781450356572. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

A SETTING AND NOTATION

For a tokenized sequence $W_{1:L}$, let $x = W_{<t}$ denote the prediction prefix at position t, and let $\boldsymbol{p}(x)$ and $\boldsymbol{q}(x)$ denote the student and teacher next-token distributions over the shared vocabulary $\mathcal{V} = \{1, \ldots, m\}$, abbreviated as $\boldsymbol{p}$ and $\boldsymbol{q}$.

The context $C \sim P_{\mathrm{ctx}}$ is drawn from the context population used to construct $\boldsymbol{K}$. Conditional on $C = c$, the auxiliary token I is drawn from the teacher distribution $P(I \mid C = c)$.

B POPULATION CONSTRUCTION AND PROPERTIES

Let C be a discrete context variable with distribution P_{ctx}, and restrict attention to contexts with $P_{\mathrm{ctx}}(c) > 0$. For every token with $P(I = i) > 0$, Bayes' rule gives

$$h_i(c) = \frac{P(C = c \mid I = i)}{P_{\mathrm{ctx}}(c)} = \frac{P(I = i \mid C = c)}{P(I = i)}. \tag{B.1}$$

Each h_i is nonnegative and satisfies

$$\sum_c h_i(c)\, P_{\mathrm{ctx}}(c) = 1. \tag{B.2}$$

Moreover,

$$0 \le h_i(c) \le \frac{1}{P(I = i)},$$

so $h_i \in L^2(P_{\mathrm{ctx}})$ and all inner products $\langle h_i, h_j \rangle_{P_{\mathrm{ctx}}}$ are finite.

The context-mixture density ratios satisfy

$$\frac{Q_{\boldsymbol{p}}}{P_{\mathrm{ctx}}} = \sum_{i=1}^{m} p_i h_i, \qquad \frac{Q_{\boldsymbol{q}}}{P_{\mathrm{ctx}}} = \sum_{i=1}^{m} q_i h_i. \tag{B.3}$$

By Equation (B.2) and $\boldsymbol{p}, \boldsymbol{q} \in \Delta^{m-1}$, both mixture density ratios have mean one under P_{ctx}:

$$\sum_c \frac{Q_{\boldsymbol{p}}(c)}{P_{\mathrm{ctx}}(c)}\, P_{\mathrm{ctx}}(c) = \sum_c \frac{Q_{\boldsymbol{q}}(c)}{P_{\mathrm{ctx}}(c)}\, P_{\mathrm{ctx}}(c) = 1. \tag{B.4}$$

Using

$$K_{ij} = \langle h_i, h_j \rangle_{P_{\mathrm{ctx}}},$$

we obtain

$$\left\langle \frac{Q_{\boldsymbol{p}}}{P_{\mathrm{ctx}}}, \frac{Q_{\boldsymbol{q}}}{P_{\mathrm{ctx}}} \right\rangle_{P_{\mathrm{ctx}}} = \sum_{i,j} p_i q_j \langle h_i, h_j \rangle_{P_{\mathrm{ctx}}} = \boldsymbol{p}^\top \boldsymbol{K} \boldsymbol{q}, \tag{B.5}$$

$$\left\| \frac{Q_{\boldsymbol{p}}}{P_{\mathrm{ctx}}} \right\|_{P_{\mathrm{ctx}}}^2 = \boldsymbol{p}^\top \boldsymbol{K} \boldsymbol{p}, \tag{B.6}$$

$$\left\| \frac{Q_{\boldsymbol{q}}}{P_{\mathrm{ctx}}} \right\|_{P_{\mathrm{ctx}}}^2 = \boldsymbol{q}^\top \boldsymbol{K} \boldsymbol{q}. \tag{B.7}$$

Substitution gives Equation (15).

For any $\boldsymbol{u} \in \mathbb{R}^m$,

$$\boldsymbol{u}^\top \boldsymbol{K} \boldsymbol{u} = \left\| \sum_i u_i h_i \right\|_{P_{\mathrm{ctx}}}^2 \ge 0, \tag{B.8}$$

so $\boldsymbol{K}$ is symmetric positive semidefinite.

By the equality condition of the Cauchy–Schwarz inequality and Equation (B.4),

$$D_{\mathrm{CS},P_{\mathrm{ctx}}}(Q_{\boldsymbol{p}}, Q_{\boldsymbol{q}}) = 0 \iff Q_{\boldsymbol{p}} = Q_{\boldsymbol{q}} \iff \boldsymbol{K}(\boldsymbol{p} - \boldsymbol{q}) = \boldsymbol{0}. \tag{B.9}$$

If $\boldsymbol{K}$ is positive definite, the composition term has the unique minimizer $\boldsymbol{p} = \boldsymbol{q}$. If $\boldsymbol{K}$ is only positive semidefinite, differences $\boldsymbol{p} - \boldsymbol{q}$ in its null space leave the divergence at zero, and the probability-matching objective in Equation (17) is needed to identify $\boldsymbol{q}$.

C SAMPLED-CONTEXT DISTRIBUTIONS AND TEMPERATURE SELECTION

Let $\mathcal{C} = \{C_1, \ldots, C_N\}$ be the sampled context set, and place the uniform distribution on these contexts:

$$P_{\mathcal{C}}(C = C_n) = \frac{1}{N}, \qquad n = 1, \ldots, N. \tag{C.1}$$

Evaluating the frozen teacher at each context C_n gives the next-token probabilities

$$s_{ni} \coloneqq P(I = i \mid C = C_n). \tag{C.2}$$

For each token i, we define

$$\pi_{ni}(\tau_i) = \frac{s_{ni}^{1/\tau_i}}{\sum_{r=1}^{N} s_{ri}^{1/\tau_i}} = \frac{\exp(\log s_{ni}/\tau_i)}{\sum_{r=1}^{N} \exp(\log s_{ri}/\tau_i)}. \tag{C.3}$$

At $\tau_i = 1$, Bayes' rule under $P_{\mathcal{C}}$ gives

$$\begin{aligned} P_{\mathcal{C}}(C = C_n \mid I = i) &= \frac{P_{\mathcal{C}}(C = C_n)\, s_{ni}}{\sum_{r=1}^{N} P_{\mathcal{C}}(C = C_r)\, s_{ri}} \\ &= \frac{N^{-1} s_{ni}}{N^{-1} \sum_{r=1}^{N} s_{ri}} = \pi_{ni}(1). \end{aligned} \tag{C.4}$$

The token-specific temperature $\tau_i > 0$ is chosen to satisfy

$$\exp H(\pi_{\cdot i}(\tau_i)) = \kappa, \tag{C.5}$$

where $H(\pi) = -\sum_n \pi_n \log \pi_n$. The temperature is determined independently for each token by solving this one-dimensional equation.

With respect to $P_{\mathcal{C}}$, the density ratio corresponding to $\pi_{\cdot i}$ is

$$h_i^{\mathcal{C}}(C_n) = \frac{\pi_{ni}}{P_{\mathcal{C}}(C = C_n)} = N\pi_{ni}, \tag{C.6}$$

which has mean one:

$$\mathbb{E}_{C \sim P_{\mathcal{C}}}[h_i^{\mathcal{C}}(C)] = \frac{1}{N} \sum_{n=1}^{N} N\pi_{ni} = 1. \tag{C.7}$$

The corresponding finite Gram matrix is

$$K_{ij}^{\mathcal{C}} = \left\langle h_i^{\mathcal{C}}, h_j^{\mathcal{C}} \right\rangle_{P_{\mathcal{C}}} = \frac{1}{N} \sum_{n=1}^{N} h_i^{\mathcal{C}}(C_n) h_j^{\mathcal{C}}(C_n) = N \sum_{n=1}^{N} \pi_{ni}\pi_{nj}. \tag{C.8}$$

Because Equation (15) is invariant to a common positive scaling of $\boldsymbol{K}^{\mathcal{C}}$, we drop the factor N and use $K_{ij}^{\mathcal{C}} = \sum_{n=1}^{N} \pi_{ni}\pi_{nj}$ in computation.

D ANALYTIC GRADIENTS IN THE SIMPLEX EXAMPLE

We derive the logit gradients underlying Figure 2 and examine two representative redistributions. Using the ordering (*cat*, *kitten*, *car*), the teacher distribution is

$$\boldsymbol{q} = (0.6, 0.3, 0.1). \tag{D.1}$$

For this illustration only, we binarize each column of Table 4 by keeping its three highest-probability contexts and set K_{ij} to the fraction of these contexts shared by tokens i and j. The sets for *cat* and *kitten* share two of the three contexts, whereas the set for *car* shares none with either, giving

$$\boldsymbol{K} = \begin{pmatrix} 1 & 2/3 & 0 \\ 2/3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{D.2}$$

Table 4: $P(C \mid I = i)$ for the eight contexts of Figure 1(b), from GPT-2 124M (for illustration). Each column sums to one.

Context C_n	*cat*	*kitten*	*car*
She adopted a	0.094	0.112	0.001
The vet examined the	0.307	0.289	0.009
A ball of yarn amused the	0.102	0.155	0.000
She fed the	0.257	0.289	0.002
He parked the	0.004	0.001	0.714
The mechanic repaired the	0.022	0.003	0.051
They bought a new	0.025	0.019	0.184
He took it for a walk, then leashed the	0.188	0.132	0.039

Table 5: Gradient norms along $\boldsymbol{p}(t) = (1-t)\boldsymbol{q} + t\boldsymbol{e}_{\text{car}}$, with ratio $\|\boldsymbol{g}_{\text{RCD}}\|_2/\|\boldsymbol{g}_{\text{FKL}}\|_2$ (orange above one, blue below one).

t	p_{cat}	p_{kitten}	p_{car}	$\|\boldsymbol{g}_{\text{FKL}}\|_2$	$\|\boldsymbol{g}_{\text{RCD}}\|_2$	ratio
0.10	0.540	0.270	0.190	0.1122	0.0950	0.846
0.20	0.480	0.240	0.280	0.2245	0.2409	1.073
0.30	0.420	0.210	0.370	0.3367	0.4409	1.309
0.40	0.360	0.180	0.460	0.4490	0.6764	1.507
0.50	0.300	0.150	0.550	0.5612	0.9075	1.617
0.60	0.240	0.120	0.640	0.6735	1.0893	1.617
0.70	0.180	0.090	0.730	0.7857	1.1899	1.514
0.80	0.120	0.060	0.820	0.8980	1.1870	1.322
0.95	0.030	0.015	0.955	1.0664	0.8639	0.810
1.00	0.000	0.000	1.000	1.1225	0.5612	0.500

Let $D_{\boldsymbol{K}}(\boldsymbol{p}, \boldsymbol{q})$ denote the right-hand side of Equation (15). Differentiation with respect to the student probabilities gives

$$\frac{\partial D_{\boldsymbol{K}}}{\partial p_i} = 2\left[\frac{(\boldsymbol{K}\boldsymbol{p})_i}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{p}} - \frac{(\boldsymbol{K}\boldsymbol{q})_i}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{q}}\right]. \tag{D.3}$$

Since $\boldsymbol{p}^\top \nabla_{\boldsymbol{p}} D_{\boldsymbol{K}} = 0$, the centering term of the softmax Jacobian vanishes, giving

$$\frac{\partial D_{\boldsymbol{K}}}{\partial z_i} = 2p_i\left[\frac{(\boldsymbol{K}\boldsymbol{p})_i}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{p}} - \frac{(\boldsymbol{K}\boldsymbol{q})_i}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{q}}\right]. \tag{D.4}$$

For forward KL,

$$\nabla_{\boldsymbol{z}} \mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}) = \boldsymbol{p} - \boldsymbol{q}. \tag{D.5}$$

Hence, for the $\eta = 0.5$ field in Figure 2,

$$\boldsymbol{g}_{\text{RCD}} = \tfrac{1}{2}(\boldsymbol{p} - \boldsymbol{q}) + \tfrac{1}{2}\nabla_{\boldsymbol{z}} D_{\boldsymbol{K}}(\boldsymbol{p}, \boldsymbol{q}). \tag{D.6}$$

The composition term changes the gradient magnitude across redistribution directions. Along the *cat–kitten* direction, consider $\boldsymbol{p} = (0.5, 0.4, 0.1)$. At this point,

$$\|\boldsymbol{g}_{\text{FKL}}\|_2 = 0.1414, \qquad \|\boldsymbol{g}_{\text{RCD}}\|_2 = 0.1010. \tag{D.7}$$

The RCD gradient norm is smaller than the FKL gradient norm, consistent with the flatter field in Figure 2.

We next move from the teacher distribution toward the *car* vertex along

$$\boldsymbol{p}(t) = (1-t)\boldsymbol{q} + t\boldsymbol{e}_{\text{car}} = \big(0.6(1-t),\, 0.3(1-t),\, 0.1 + 0.9t\big). \tag{D.8}$$

Table 5 reports the corresponding logit-gradient magnitudes. Near the teacher distribution, the RCD gradient is slightly smaller than the FKL gradient. As excess mass accumulates on *car*, the RCD gradient becomes larger over a broad interior region. The ratio peaks around $t = 0.5$–0.6 and decreases again toward the vertex.

The non-monotone behavior follows directly from the composition gradient. Since $K_{\mathrm{car,cat}} = K_{\mathrm{car,kitten}} = 0$, we have

$$(\boldsymbol{K}\boldsymbol{p})_{\mathrm{car}} = p_{\mathrm{car}}, \qquad (\boldsymbol{K}\boldsymbol{q})_{\mathrm{car}} = 0.1, \tag{D.9}$$

and therefore

$$\frac{\partial D_{\boldsymbol{K}}}{\partial z_{\mathrm{car}}} = 2p_{\mathrm{car}} \left[\frac{p_{\mathrm{car}}}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{p}} - \frac{0.1}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{q}} \right]. \tag{D.10}$$

As p_{car} increases, the bracket grows over much of the simplex interior, increasing the *car*-coordinate correction.

At the vertex, $\boldsymbol{p} \to \boldsymbol{e}_{\mathrm{car}}$ gives

$$\frac{p_{\mathrm{car}}}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{p}} \to 1, \qquad \frac{0.1}{\boldsymbol{p}^\top \boldsymbol{K}\boldsymbol{q}} \to 1, \tag{D.11}$$

so the *car* bracket vanishes. The remaining coordinates are suppressed by $p_i \to 0$, and hence

$$\nabla_{\boldsymbol{z}} D_{\boldsymbol{K}}(\boldsymbol{p}, \boldsymbol{q}) \to 0. \tag{D.12}$$

E Experimental details

E.1 Data and models

We use the Databricks Dolly 15K data in the MiniLLM format, with maximum sequence and prompt lengths of 512 and 256 tokens. Examples whose instruction/input key matches one in DollyEval or in the 1,000-example development split of Gu et al. (2024) are removed. DollyEval and S-NI follow the splits of Gu et al. (2024), with S-NI restricted to responses of more than 10 tokens. The teachers are GPT-2 XL (Radford et al., 2019) and OPT-1.3B (Zhang et al., 2022), fine-tuned on the training split; the students are initialized from the pretrained GPT-2 124M and OPT-125M checkpoints without supervised fine-tuning.

E.2 Training

Losses are averaged over response-token prediction positions, with $T_{\mathrm{KD}} = 1$. Students are trained with AdamW, cosine learning-rate decay, no learning-rate warmup, weight decay 10^{-2}, gradient clipping at 1.0, and are trained for 20 epochs on a single A100; the teacher runs in fp16. Distillation runs use the ablation configuration of CSD (learning rate 10^{-4}, effective batch size 8), and SFT uses learning rate $5{\times}10^{-4}$ at effective batch size 32. For RCD, η is increased linearly from 0 to its target value over the first epoch. Checkpoints are selected by development ROUGE-L. We use training seeds $\{10, 20, 30\}$, which vary both dropout and data order, and decoding seeds $\{10, 20, 30, 40, 50\}$.

E.3 Context set and active set

For each training example with at least one response-token position, we sample one such position and add its prefix to $\mathcal{C}$. The matrix $\boldsymbol{K}^{\mathcal{C}}$ is constructed once per family from the frozen teacher and kept fixed. The composition term is evaluated on the active set A of Equation (22), which is held fixed during differentiation; the probability-matching objective uses the full vocabulary.

E.4 Compared objectives

The probability-matching objectives are FKL, SRKL with skew parameter 0.1 (Ko et al., 2024), and α–β with $(\alpha, \beta) = (0.2, 0.7)$ (Wang et al., 2025). Setting $\eta = 0$ in Equation (17) recovers X from RCD (X). SFT and CSD (Kim et al., 2026) are trained separately with the same data, epochs, and evaluation protocol.

E.5 Evaluation metrics

Prediction-quality metrics are computed with teacher forcing on the reference-response positions $\mathcal{H}$ of each evaluation set. For generation, we sample $K = 5$ responses per input at temperature 1 and top-$p = 1$, one per decoding seed; all generation metrics use these five responses, and inputs with several references use the best-matching one.

Table 6: Evaluation metrics.

Metric	Direction	Measures
Prediction quality		
NLL	↓	Fit to observed response tokens
ECE	↓	Calibration error relative to teacher probabilities
Generation quality		
ROUGE-L	↑	Average lexical reference quality
Best-of-5 ROUGE-L (Oracle@5)	↑	Best reference match among five samples
BERTScore-F1 mean	↑	Average semantic reference quality
Worst-of-5 BERTScore-F1 (min@5)	↑	Lowest semantic reference quality among five samples
Generation diversity		
Self-BLEU	↓	Similarity among samples
Distinct-2	↑	Bigram diversity among samples

E.5.1 NLL

The mean negative log-likelihood of the observed response tokens under the student.

E.5.2 ECE

We use the calibration error of Kim et al. (2026, App. D.1), averaged over the teacher-forced positions $\mathcal{H}$:

$$\mathrm{ECE} = \frac{1}{|\mathcal{H}|} \sum_{x \in \mathcal{H}} \sum_{y} p_y(x) \left| q_y(x) - p_y(x) \right|. \tag{E.1}$$

E.5.3 ROUGE-L AND ORACLE@5

ROUGE-L (Lin, 2004) uses the `rouge_score` implementation with Porter stemming, scaled by 100 (Gu et al., 2024; Ko et al., 2024). We report the mean over the five samples and the best-of-five value (Oracle@5).

E.5.4 BERTSCORE-F1 AND MIN@5

BERTScore-F1 (Zhang et al., 2020) uses the reference implementation with its default English model, without IDF weighting or baseline rescaling. We report the mean over the five samples and the worst-of-five value (min@5).

E.5.5 SELF-BLEU

Each sample is scored with BLEU (Papineni et al., 2002) against the other four,

$$\mathrm{SelfBLEU}(x) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{BLEU}\big(y^{(k)}, \{y^{(j)}\}_{j \neq k}\big), \tag{E.2}$$

following Zhu et al. (2018) and Tevet & Berant (2021).

E.5.6 DISTINCT-2

The fraction of distinct bigrams among all bigrams of the five samples.[1]

E.6 STATISTICAL AGGREGATION

Each metric is averaged over the inputs of an evaluation set within a training seed, and Table 1 reports the mean and standard deviation over training seeds of the average over the four evaluation sets. For Table 2, RCD (X) and X are paired by training seed in each of the six family–objective pairs, and the

[1] Li et al. (2016) instead normalize by the number of generated tokens.

three seed differences are averaged. The pooled effect is the mean over the six pairs, with a percentile cluster-bootstrap interval from 10,000 resamples of the six pairs. Fav./6 counts the pairs whose mean difference favors RCD, and Fav./18 the seed-level differences that favor RCD.

F PER-SET RESULTS

Every cell below is computed within a training seed (generation metrics from the five decoding samples) and reported as the mean and sample standard deviation over the three training seeds.

Tables 7 and 8 (ROUGE-L and prediction quality) and Tables 9 and 10 (generation metrics) break Table 1 down by evaluation set; each main-text cell is the unweighted mean of the four per-set values within a training seed, aggregated over seeds as above. Tables 11 to 14 report the same quantities for RCD (FKL) and RCD (α–β).

Table 7: Per-set ROUGE-L and prediction-quality metrics, GPT-2 family.

Method	ROUGE-L ↑	NLL ↓	ECE ↓
	DollyEval		
SFT	$22.70_{\pm 0.05}$	$5.61_{\pm 0.025}$	$0.2968_{\pm 0.0011}$
FKL	$22.38_{\pm 0.28}$	$4.43_{\pm 0.002}$	$0.2453_{\pm 0.0008}$
SRKL	$24.85_{\pm 0.41}$	$6.48_{\pm 0.010}$	$0.3274_{\pm 0.0017}$
α–β	$25.06_{\pm 0.52}$	$8.73_{\pm 0.003}$	$0.3710_{\pm 0.0007}$
CSD	$24.59_{\pm 0.20}$	$5.80_{\pm 0.035}$	$0.2952_{\pm 0.0016}$
RCD (SRKL)	$23.45_{\pm 0.51}$	$4.24_{\pm 0.827}$	$0.2369_{\pm 0.0357}$
	SelfInst		
SFT	$10.12_{\pm 0.24}$	$6.45_{\pm 0.053}$	$0.3228_{\pm 0.0034}$
FKL	$10.25_{\pm 0.15}$	$5.11_{\pm 0.021}$	$0.2660_{\pm 0.0027}$
SRKL	$12.30_{\pm 0.55}$	$7.02_{\pm 0.015}$	$0.3498_{\pm 0.0037}$
α–β	$12.60_{\pm 0.41}$	$9.44_{\pm 0.028}$	$0.3964_{\pm 0.0015}$
CSD	$12.33_{\pm 0.24}$	$6.25_{\pm 0.019}$	$0.3130_{\pm 0.0012}$
RCD (SRKL)	$11.42_{\pm 0.47}$	$4.78_{\pm 0.940}$	$0.2548_{\pm 0.0410}$
	VicunaEval		
SFT	$15.06_{\pm 0.14}$	$4.52_{\pm 0.032}$	$0.3030_{\pm 0.0022}$
FKL	$14.95_{\pm 0.44}$	$3.41_{\pm 0.022}$	$0.2532_{\pm 0.0013}$
SRKL	$15.25_{\pm 0.81}$	$4.72_{\pm 0.060}$	$0.3199_{\pm 0.0031}$
α–β	$15.03_{\pm 0.60}$	$6.39_{\pm 0.013}$	$0.3585_{\pm 0.0015}$
CSD	$16.18_{\pm 0.46}$	$4.30_{\pm 0.014}$	$0.2955_{\pm 0.0006}$
RCD (SRKL)	$15.39_{\pm 0.29}$	$3.24_{\pm 0.618}$	$0.2453_{\pm 0.0328}$
	S-NI		
SFT	$16.87_{\pm 0.30}$	$6.73_{\pm 0.019}$	$0.2889_{\pm 0.0026}$
FKL	$18.37_{\pm 0.37}$	$5.47_{\pm 0.008}$	$0.2443_{\pm 0.0014}$
SRKL	$25.01_{\pm 0.73}$	$7.84_{\pm 0.062}$	$0.2990_{\pm 0.0033}$
α–β	$27.52_{\pm 0.92}$	$10.18_{\pm 0.072}$	$0.3265_{\pm 0.0041}$
CSD	$23.92_{\pm 0.92}$	$7.46_{\pm 0.121}$	$0.2829_{\pm 0.0030}$
RCD (SRKL)	$21.78_{\pm 1.40}$	$5.39_{\pm 1.045}$	$0.2408_{\pm 0.0290}$

Table 8: Per-set ROUGE-L and prediction-quality metrics, OPT family.

Method	ROUGE-L ↑	NLL ↓	ECE ↓
	DollyEval		
SFT	$20.82_{\pm 0.63}$	$5.84_{\pm 0.024}$	$0.2906_{\pm 0.0014}$
FKL	$22.44_{\pm 0.40}$	$4.97_{\pm 0.007}$	$0.2462_{\pm 0.0005}$
SRKL	$23.76_{\pm 0.35}$	$6.04_{\pm 0.020}$	$0.3118_{\pm 0.0013}$
α–β	$24.08_{\pm 0.27}$	$10.13_{\pm 0.056}$	$0.4394_{\pm 0.0002}$
CSD	$22.75_{\pm 0.29}$	$5.16_{\pm 0.014}$	$0.2550_{\pm 0.0018}$

Method	ROUGE-L ↑	NLL ↓	ECE ↓
RCD (SRKL)	$21.64_{\pm 0.24}$	$4.83_{\pm 0.190}$	$0.2301_{\pm 0.0013}$
SelfInst			
SFT	$7.64_{\pm 0.25}$	$6.86_{\pm 0.009}$	$0.2988_{\pm 0.0024}$
FKL	$8.82_{\pm 0.20}$	$5.59_{\pm 0.019}$	$0.2529_{\pm 0.0037}$
SRKL	$10.19_{\pm 0.51}$	$6.58_{\pm 0.027}$	$0.3276_{\pm 0.0008}$
α–β	$11.00_{\pm 0.21}$	$10.87_{\pm 0.025}$	$0.4704_{\pm 0.0023}$
CSD	$8.99_{\pm 0.50}$	$5.62_{\pm 0.028}$	$0.2586_{\pm 0.0010}$
RCD (SRKL)	$8.26_{\pm 0.14}$	$5.36_{\pm 0.179}$	$0.2358_{\pm 0.0027}$
VicunaEval			
SFT	$13.80_{\pm 0.07}$	$5.23_{\pm 0.015}$	$0.2977_{\pm 0.0008}$
FKL	$14.64_{\pm 0.08}$	$4.18_{\pm 0.007}$	$0.2521_{\pm 0.0029}$
SRKL	$14.88_{\pm 0.14}$	$5.09_{\pm 0.014}$	$0.3149_{\pm 0.0025}$
α–β	$15.22_{\pm 0.26}$	$8.66_{\pm 0.106}$	$0.4349_{\pm 0.0045}$
CSD	$15.36_{\pm 0.30}$	$4.34_{\pm 0.030}$	$0.2620_{\pm 0.0019}$
RCD (SRKL)	$14.08_{\pm 0.18}$	$4.13_{\pm 0.197}$	$0.2383_{\pm 0.0028}$
S-NI			
SFT	$11.73_{\pm 0.75}$	$6.75_{\pm 0.029}$	$0.2890_{\pm 0.0052}$
FKL	$14.60_{\pm 0.84}$	$5.73_{\pm 0.042}$	$0.2456_{\pm 0.0036}$
SRKL	$19.96_{\pm 1.30}$	$6.82_{\pm 0.055}$	$0.2957_{\pm 0.0016}$
α–β	$21.15_{\pm 2.22}$	$10.54_{\pm 0.103}$	$0.3927_{\pm 0.0039}$
CSD	$16.31_{\pm 0.81}$	$5.93_{\pm 0.031}$	$0.2477_{\pm 0.0034}$
RCD (SRKL)	$14.97_{\pm 0.23}$	$5.56_{\pm 0.209}$	$0.2295_{\pm 0.0028}$

Table 9: Per-set generation metrics, GPT-2 family. Table 1 averages the four sets within each seed.

	BERTScore-F1 ↑				
Method	mean	min@5	S-BLEU ↓	Oracle@5 ↑	Dist-2 ↑
DollyEval					
SFT	$0.855_{\pm 0.0005}$	$0.837_{\pm 0.0008}$	$0.308_{\pm 0.011}$	$29.65_{\pm 0.30}$	$0.716_{\pm 0.006}$
FKL	$0.855_{\pm 0.0007}$	$0.835_{\pm 0.0006}$	$0.212_{\pm 0.002}$	$30.00_{\pm 0.51}$	$0.786_{\pm 0.001}$
SRKL	$0.858_{\pm 0.0008}$	$0.843_{\pm 0.0015}$	$0.463_{\pm 0.016}$	$30.70_{\pm 0.36}$	$0.536_{\pm 0.013}$
α–β	$0.856_{\pm 0.0006}$	$0.842_{\pm 0.0009}$	$0.602_{\pm 0.017}$	$29.94_{\pm 0.44}$	$0.432_{\pm 0.004}$
CSD	$0.858_{\pm 0.0000}$	$0.840_{\pm 0.0003}$	$0.412_{\pm 0.005}$	$30.57_{\pm 0.35}$	$0.568_{\pm 0.006}$
RCD (SRKL)	$0.858_{\pm 0.0019}$	$0.837_{\pm 0.0035}$	$0.261_{\pm 0.033}$	$30.67_{\pm 0.34}$	$0.731_{\pm 0.013}$
SelfInst					
SFT	$0.816_{\pm 0.0006}$	$0.799_{\pm 0.0009}$	$0.179_{\pm 0.015}$	$15.39_{\pm 0.71}$	$0.805_{\pm 0.012}$
FKL	$0.816_{\pm 0.0004}$	$0.798_{\pm 0.0005}$	$0.134_{\pm 0.003}$	$16.23_{\pm 0.23}$	$0.841_{\pm 0.003}$
SRKL	$0.820_{\pm 0.0002}$	$0.803_{\pm 0.0010}$	$0.412_{\pm 0.016}$	$17.67_{\pm 0.73}$	$0.554_{\pm 0.008}$
α–β	$0.819_{\pm 0.0013}$	$0.805_{\pm 0.0025}$	$0.567_{\pm 0.004}$	$16.54_{\pm 0.56}$	$0.438_{\pm 0.005}$
CSD	$0.819_{\pm 0.0011}$	$0.801_{\pm 0.0029}$	$0.347_{\pm 0.024}$	$17.56_{\pm 0.35}$	$0.588_{\pm 0.013}$
RCD (SRKL)	$0.819_{\pm 0.0015}$	$0.800_{\pm 0.0032}$	$0.202_{\pm 0.013}$	$17.25_{\pm 0.78}$	$0.763_{\pm 0.004}$
VicunaEval					
SFT	$0.829_{\pm 0.0008}$	$0.816_{\pm 0.0004}$	$0.135_{\pm 0.005}$	$18.69_{\pm 0.14}$	$0.817_{\pm 0.004}$
FKL	$0.830_{\pm 0.0010}$	$0.817_{\pm 0.0020}$	$0.081_{\pm 0.004}$	$18.59_{\pm 0.17}$	$0.855_{\pm 0.003}$
SRKL	$0.832_{\pm 0.0012}$	$0.817_{\pm 0.0024}$	$0.317_{\pm 0.026}$	$18.93_{\pm 0.53}$	$0.570_{\pm 0.023}$
α–β	$0.829_{\pm 0.0011}$	$0.816_{\pm 0.0015}$	$0.452_{\pm 0.024}$	$18.37_{\pm 0.57}$	$0.462_{\pm 0.007}$
CSD	$0.833_{\pm 0.0007}$	$0.818_{\pm 0.0012}$	$0.253_{\pm 0.011}$	$20.23_{\pm 0.46}$	$0.608_{\pm 0.004}$
RCD (SRKL)	$0.833_{\pm 0.0022}$	$0.819_{\pm 0.0028}$	$0.122_{\pm 0.016}$	$19.46_{\pm 0.10}$	$0.787_{\pm 0.004}$
S-NI					
SFT	$0.834_{\pm 0.0007}$	$0.814_{\pm 0.0011}$	$0.212_{\pm 0.012}$	$27.42_{\pm 0.33}$	$0.793_{\pm 0.009}$
FKL	$0.836_{\pm 0.0003}$	$0.811_{\pm 0.0007}$	$0.176_{\pm 0.003}$	$30.74_{\pm 0.52}$	$0.818_{\pm 0.003}$
SRKL	$0.844_{\pm 0.0010}$	$0.824_{\pm 0.0023}$	$0.505_{\pm 0.009}$	$34.29_{\pm 0.87}$	$0.562_{\pm 0.008}$

Method	BERTScore-F1 ↑ mean	BERTScore-F1 ↑ min@5	S-BLEU ↓	Oracle@5 ↑	Dist-2 ↑
α–β	$0.847_{\pm 0.0014}$	$0.831_{\pm 0.0020}$	$0.651_{\pm 0.018}$	$34.64_{\pm 0.50}$	$0.471_{\pm 0.006}$
CSD	$0.843_{\pm 0.0026}$	$0.820_{\pm 0.0037}$	$0.457_{\pm 0.017}$	$33.63_{\pm 0.94}$	$0.589_{\pm 0.007}$
RCD (SRKL)	$0.841_{\pm 0.0008}$	$0.816_{\pm 0.0023}$	$0.268_{\pm 0.022}$	$33.50_{\pm 2.38}$	$0.740_{\pm 0.004}$

Table 10: Per-set generation metrics, OPT family. Table 1 averages the four sets within each seed.

Method	BERTScore-F1 ↑ mean	BERTScore-F1 ↑ min@5	S-BLEU ↓	Oracle@5 ↑	Dist-2 ↑
DollyEval					
SFT	$0.847_{\pm 0.0011}$	$0.828_{\pm 0.0016}$	$0.250_{\pm 0.002}$	$27.92_{\pm 0.57}$	$0.764_{\pm 0.003}$
FKL	$0.853_{\pm 0.0012}$	$0.832_{\pm 0.0019}$	$0.253_{\pm 0.006}$	$29.71_{\pm 0.49}$	$0.761_{\pm 0.004}$
SRKL	$0.853_{\pm 0.0008}$	$0.834_{\pm 0.0021}$	$0.354_{\pm 0.009}$	$29.92_{\pm 0.08}$	$0.612_{\pm 0.001}$
α–β	$0.847_{\pm 0.0015}$	$0.830_{\pm 0.0014}$	$0.567_{\pm 0.004}$	$29.19_{\pm 0.10}$	$0.411_{\pm 0.003}$
CSD	$0.851_{\pm 0.0003}$	$0.829_{\pm 0.0011}$	$0.270_{\pm 0.004}$	$29.61_{\pm 0.33}$	$0.694_{\pm 0.003}$
RCD (SRKL)	$0.850_{\pm 0.0007}$	$0.829_{\pm 0.0016}$	$0.213_{\pm 0.010}$	$28.99_{\pm 0.52}$	$0.792_{\pm 0.002}$
SelfInst					
SFT	$0.808_{\pm 0.0003}$	$0.792_{\pm 0.0002}$	$0.108_{\pm 0.008}$	$12.34_{\pm 0.65}$	$0.866_{\pm 0.004}$
FKL	$0.811_{\pm 0.0006}$	$0.792_{\pm 0.0005}$	$0.118_{\pm 0.006}$	$14.63_{\pm 0.75}$	$0.863_{\pm 0.006}$
SRKL	$0.813_{\pm 0.0012}$	$0.794_{\pm 0.0015}$	$0.263_{\pm 0.031}$	$16.02_{\pm 0.44}$	$0.647_{\pm 0.016}$
α–β	$0.810_{\pm 0.0006}$	$0.794_{\pm 0.0006}$	$0.517_{\pm 0.010}$	$15.21_{\pm 0.28}$	$0.440_{\pm 0.009}$
CSD	$0.808_{\pm 0.0016}$	$0.787_{\pm 0.0026}$	$0.149_{\pm 0.002}$	$14.69_{\pm 0.64}$	$0.751_{\pm 0.007}$
RCD (SRKL)	$0.809_{\pm 0.0011}$	$0.790_{\pm 0.0010}$	$0.100_{\pm 0.009}$	$14.11_{\pm 0.40}$	$0.874_{\pm 0.004}$
VicunaEval					
SFT	$0.819_{\pm 0.0006}$	$0.805_{\pm 0.0008}$	$0.104_{\pm 0.006}$	$17.76_{\pm 0.16}$	$0.857_{\pm 0.003}$
FKL	$0.824_{\pm 0.0005}$	$0.810_{\pm 0.0005}$	$0.083_{\pm 0.010}$	$18.42_{\pm 0.13}$	$0.864_{\pm 0.004}$
SRKL	$0.828_{\pm 0.0009}$	$0.815_{\pm 0.0018}$	$0.247_{\pm 0.027}$	$18.75_{\pm 0.20}$	$0.653_{\pm 0.006}$
α–β	$0.820_{\pm 0.0017}$	$0.806_{\pm 0.0032}$	$0.486_{\pm 0.022}$	$18.67_{\pm 0.34}$	$0.410_{\pm 0.009}$
CSD	$0.826_{\pm 0.0006}$	$0.810_{\pm 0.0009}$	$0.148_{\pm 0.014}$	$19.36_{\pm 0.15}$	$0.740_{\pm 0.008}$
RCD (SRKL)	$0.821_{\pm 0.0013}$	$0.807_{\pm 0.0021}$	$0.098_{\pm 0.009}$	$17.82_{\pm 0.16}$	$0.865_{\pm 0.010}$
S-NI					
SFT	$0.822_{\pm 0.0005}$	$0.803_{\pm 0.0006}$	$0.090_{\pm 0.004}$	$21.01_{\pm 1.02}$	$0.890_{\pm 0.006}$
FKL	$0.826_{\pm 0.0026}$	$0.802_{\pm 0.0027}$	$0.131_{\pm 0.016}$	$26.40_{\pm 0.96}$	$0.872_{\pm 0.013}$
SRKL	$0.833_{\pm 0.0021}$	$0.808_{\pm 0.0022}$	$0.274_{\pm 0.023}$	$31.14_{\pm 1.41}$	$0.703_{\pm 0.024}$
α–β	$0.834_{\pm 0.0018}$	$0.813_{\pm 0.0011}$	$0.549_{\pm 0.043}$	$28.83_{\pm 2.81}$	$0.499_{\pm 0.036}$
CSD	$0.826_{\pm 0.0031}$	$0.799_{\pm 0.0037}$	$0.175_{\pm 0.008}$	$27.55_{\pm 0.83}$	$0.763_{\pm 0.010}$
RCD (SRKL)	$0.826_{\pm 0.0003}$	$0.801_{\pm 0.0011}$	$0.128_{\pm 0.001}$	$26.59_{\pm 0.61}$	$0.864_{\pm 0.006}$

Table 11: Per-set results for RCD (FKL) and RCD (α–β), GPT-2 family.

Method	ROUGE-L ↑	NLL ↓	ECE ↓
DollyEval			
RCD (FKL)	$22.19_{\pm 0.28}$	$4.31_{\pm 0.042}$	$0.2385_{\pm 0.0016}$
RCD (α–β)	$25.02_{\pm 0.35}$	$6.25_{\pm 0.114}$	$0.3298_{\pm 0.0013}$
SelfInst			
RCD (FKL)	$10.21_{\pm 0.18}$	$4.98_{\pm 0.041}$	$0.2594_{\pm 0.0033}$
RCD (α–β)	$12.66_{\pm 0.13}$	$7.05_{\pm 0.090}$	$0.3561_{\pm 0.0010}$
VicunaEval			
RCD (FKL)	$14.88_{\pm 0.18}$	$3.34_{\pm 0.021}$	$0.2478_{\pm 0.0015}$

Method	ROUGE-L ↑	NLL ↓	ECE ↓
RCD (α–β)	$15.73_{\pm 0.10}$	$4.57_{\pm 0.067}$	$0.3242_{\pm 0.0008}$
S-NI			
RCD (FKL)	$18.18_{\pm 0.22}$	$5.34_{\pm 0.045}$	$0.2409_{\pm 0.0038}$
RCD (α–β)	$26.13_{\pm 1.46}$	$7.67_{\pm 0.163}$	$0.3011_{\pm 0.0024}$

Table 12: Per-set generation metrics for RCD (FKL) and RCD (α–β), GPT-2 family.

	BERTScore-F1 ↑				
Method	mean	min@5	S-BLEU ↓	Oracle@5 ↑	Dist-2 ↑
DollyEval					
RCD (FKL)	$0.854_{\pm 0.0005}$	$0.833_{\pm 0.0006}$	$0.206_{\pm 0.006}$	$30.07_{\pm 0.20}$	$0.789_{\pm 0.005}$
RCD (α–β)	$0.859_{\pm 0.0009}$	$0.843_{\pm 0.0010}$	$0.471_{\pm 0.013}$	$30.80_{\pm 0.34}$	$0.530_{\pm 0.012}$
SelfInst					
RCD (FKL)	$0.815_{\pm 0.0007}$	$0.796_{\pm 0.0004}$	$0.134_{\pm 0.002}$	$16.43_{\pm 0.57}$	$0.845_{\pm 0.007}$
RCD (α–β)	$0.820_{\pm 0.0011}$	$0.805_{\pm 0.0013}$	$0.431_{\pm 0.016}$	$17.13_{\pm 0.12}$	$0.537_{\pm 0.008}$
VicunaEval					
RCD (FKL)	$0.830_{\pm 0.0006}$	$0.817_{\pm 0.0004}$	$0.080_{\pm 0.005}$	$18.39_{\pm 0.02}$	$0.860_{\pm 0.009}$
RCD (α–β)	$0.834_{\pm 0.0003}$	$0.819_{\pm 0.0008}$	$0.329_{\pm 0.037}$	$19.65_{\pm 0.32}$	$0.570_{\pm 0.031}$
S-NI					
RCD (FKL)	$0.835_{\pm 0.0003}$	$0.811_{\pm 0.0010}$	$0.171_{\pm 0.003}$	$30.39_{\pm 0.26}$	$0.823_{\pm 0.003}$
RCD (α–β)	$0.846_{\pm 0.0026}$	$0.827_{\pm 0.0028}$	$0.524_{\pm 0.024}$	$34.93_{\pm 1.32}$	$0.552_{\pm 0.013}$

Table 13: Per-set results for RCD (FKL) and RCD (α–β), OPT family.

Method	ROUGE-L ↑	NLL ↓	ECE ↓
DollyEval			
RCD (FKL)	$22.16_{\pm 0.42}$	$4.88_{\pm 0.024}$	$0.2406_{\pm 0.0018}$
RCD (α–β)	$24.36_{\pm 0.14}$	$6.18_{\pm 0.129}$	$0.3474_{\pm 0.0080}$
SelfInst			
RCD (FKL)	$8.87_{\pm 0.31}$	$5.50_{\pm 0.019}$	$0.2456_{\pm 0.0007}$
RCD (α–β)	$10.98_{\pm 0.07}$	$7.02_{\pm 0.108}$	$0.3771_{\pm 0.0097}$
VicunaEval			
RCD (FKL)	$14.15_{\pm 0.42}$	$4.13_{\pm 0.008}$	$0.2458_{\pm 0.0018}$
RCD (α–β)	$15.77_{\pm 0.81}$	$5.24_{\pm 0.075}$	$0.3480_{\pm 0.0066}$
S-NI			
RCD (FKL)	$14.91_{\pm 0.45}$	$5.66_{\pm 0.031}$	$0.2378_{\pm 0.0015}$
RCD (α–β)	$19.78_{\pm 1.25}$	$7.19_{\pm 0.181}$	$0.3301_{\pm 0.0051}$

Table 14: Per-set generation metrics for RCD (FKL) and RCD (α–β), OPT family.

	BERTScore-F1 ↑				
Method	mean	min@5	S-BLEU ↓	Oracle@5 ↑	Dist-2 ↑
DollyEval					

Method	BERTScore-F1 ↑ mean	BERTScore-F1 ↑ min@5	S-BLEU ↓	Oracle@5 ↑	Dist-2 ↑
RCD (FKL)	$0.852_{\pm 0.0007}$	$0.830_{\pm 0.0006}$	$0.246_{\pm 0.005}$	$29.89_{\pm 0.50}$	$0.770_{\pm 0.001}$
RCD (α–β)	$0.856_{\pm 0.0004}$	$0.839_{\pm 0.0007}$	$0.399_{\pm 0.029}$	$30.28_{\pm 0.04}$	$0.569_{\pm 0.028}$
SelfInst					
RCD (FKL)	$0.811_{\pm 0.0008}$	$0.791_{\pm 0.0005}$	$0.113_{\pm 0.003}$	$14.50_{\pm 0.78}$	$0.873_{\pm 0.003}$
RCD (α–β)	$0.815_{\pm 0.0003}$	$0.797_{\pm 0.0003}$	$0.325_{\pm 0.033}$	$16.46_{\pm 0.47}$	$0.589_{\pm 0.026}$
VicunaEval					
RCD (FKL)	$0.823_{\pm 0.0010}$	$0.808_{\pm 0.0022}$	$0.089_{\pm 0.013}$	$18.29_{\pm 0.46}$	$0.867_{\pm 0.007}$
RCD (α–β)	$0.830_{\pm 0.0007}$	$0.818_{\pm 0.0009}$	$0.276_{\pm 0.037}$	$19.63_{\pm 0.90}$	$0.603_{\pm 0.034}$
S-NI					
RCD (FKL)	$0.826_{\pm 0.0008}$	$0.802_{\pm 0.0011}$	$0.134_{\pm 0.011}$	$27.03_{\pm 0.66}$	$0.867_{\pm 0.005}$
RCD (α–β)	$0.833_{\pm 0.0021}$	$0.810_{\pm 0.0024}$	$0.349_{\pm 0.041}$	$29.73_{\pm 1.05}$	$0.639_{\pm 0.034}$

G ROLE OF ACTIVE-SET RENORMALIZATION

Not every vocabulary item is a plausible destination for redistributed probability mass at a given prefix. After "*She adopted a*", for example, tokens such as *cat* or *kitten* are plausible under the teacher or the student, whereas most of the vocabulary receives negligible probability from both. We see no reason for these low-probability tokens to receive additional gradient from the composition term. We therefore evaluate the composition term on the active set of Equation (22), the union of the top-b teacher and student predictions, and renormalize both distributions on this set. Table 15 compares $b \in \{20, 100, m\}$, where $b = m$ recovers the full-vocabulary composition term, with all other settings fixed.

Table 15: Sensitivity to the active-set size b for RCD (SRKL) on GPT-2 (seed 10, averaged over the four evaluation sets); $b = m$ is the full vocabulary.

	Prediction quality		Generation quality				Generation diversity	
					BERTScore-F1 ↑			
b	NLL ↓	ECE ↓	ROUGE-L ↑	Oracle@5 ↑	mean	min@5	S-BLEU ↓	Dist-2 ↑
20	4.84	0.2628	18.08	25.06	0.838	0.820	0.233	0.745
100	3.42	0.2045	18.20	26.11	0.836	0.815	0.190	0.760
m	4.85	0.2626	17.84	24.83	0.837	0.818	0.215	0.759