When Does Knowledge Distillation Hurt? Reliability-Aware Distillation for Low-Resource Language Summarization

Dipto Sumit, Ankan Kumar Roy Srizon, Sadia Khair Rodela, Atia Haque Asha, Mourchona Afrin, Niloy Farhan, Farig Sadeque
Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
{dipto.sumit,ankan.kumar.roy1,sadia.khair.rodela}@g.bracu.ac.bd
{atia.haque.asha,mourchona.afrin}@g.bracu.ac.bd
{niloy.farhan,farig.sadeque}@bracu.ac.bd

Abstract

Knowledge distillation (KD) is a standard way to compress sequence-to-sequence models, but it is usually applied uniformly to every training sample. On the BanSum Bangla summarization benchmark, standard KD improves ROUGE-L by only +0.0003 over a cross-entropy (CE) baseline, and on a 5,000-sample probe, 51.3% of training samples have KD gradients that are negatively aligned with the validation-loss gradient. Gates based on teacher confidence or ROUGE agreement do no better than CE. We propose two reliability-aware distillation methods. **CHAD** (Counterfactual Harm-Aware Distillation) scores probe samples by this gradient alignment and fits a small gradient-boosted gate that predicts the score for the whole training set; it needs a single teacher. **EWAD+CPDP** needs a second teacher: it gates KD per token by teacher confidence (entropy-weighted adaptive distillation, EWAD) and adds a geometric constraint from a second, vocabulary-incompatible teacher (CPDP). With a 60M student and a 247M teacher, CHAD and EWAD+CPDP beat standard KD by +0.0173 and +0.0219 ROUGE-L, and both outperform a fine-tuned Qwen-2.5-3B model (50× larger). These gains are specific to this wide-gap, in-domain-teacher setting: with Qwen-2.5 teachers and a 3B student, no KD variant we tried beat direct fine-tuning on ROUGE. In an exploratory single-seed study on 15 XL-Sum languages, EWAD+CPDP beats the CE baseline on 10 of them, but the mean gain (+0.0044) is not significant; gains concentrate where the two teachers are complementary. We release code and trained models.

1 Introduction

Knowledge distillation (Hinton et al., 2015) is a default ingredient in compressing sequence-to-sequence models for low-resource and compute-constrained deployment. The standard objective adds to cross-entropy a KL term that pulls the student toward the teacher's softened distribution, and it treats every training sample and every token as equally worth distilling. In practice, teacher soft labels are not uniformly informative. On some examples the teacher is confident and correct; on others it is confidently wrong, ambiguous, or poorly calibrated, and following it pulls the student away from the validation optimum. Existing ways to decide *when* to trust the KD signal rely on output statistics: the student's cross-entropy (Wang et al., 2021), teacher–student disagreement (Koo et al., 2025), teacher confidence, or surface agreement between teacher outputs and references. These heuristics assume that such statistics predict whether distilling on a sample helps. We test this assumption directly, and on our benchmark it fails.

Central observation. On BanSum (Hasan et al., 2024), a 141K-sample Bangla summarization benchmark, we measure per-sample KD usefulness directly, as the alignment between a sample's KD gradient and the gold-reference validation gradient (Pruthi et al., 2020). Of 5,000 probed training samples, 51.3% have KD gradients that are negatively aligned with this direction, and standard KD gains only +0.0003 ROUGE-L over a cross-entropy baseline. Gates built on teacher entropy or ROUGE agreement give no measurable benefit over CE, and no cheap feature we tested explains more than 1% of the variance in the alignment label.

Two reliability-aware methods. **CHAD** (Counterfactual Harm-Aware Distillation) works at the *sample* level. On a probe subset we compute $s(x) = \cos(\nabla_\theta \mathcal{L}_{\text{val}}, \nabla_\theta \mathcal{L}_{\text{KD}}(x))$, train a gradient-boosted gate to predict $s(x)$ from cheap features (teacher statistics, student–teacher KL, ROUGE agreement, length), and scale each sample's KD loss by the predicted weight $h(x)$. CHAD needs only one teacher. **EWAD+CPDP** works at the *token* level and needs a second teacher. EWAD applies a sigmoid gate to the teacher's per-token top-1 probability, blending KD with CE at each posi-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 35–48
November 6-10, 2026 ©2026 Association for Computational Linguistics

tion. CPDP adds a second teacher with a different vocabulary: the encoder states of the student and both teachers are projected into a shared space, and the student is regularized so that its distances to the two teachers reflect the teachers' distance from each other.

Results. On BanSum's held-out test split, CHAD reaches R-L 0.2541 and EWAD+CPDP 0.2587, against 0.2368 for standard KD and 0.2365 for CE. Standard KD and the heuristic gates all sit within ± 0.0005 of CE, while the two reliability-aware methods gain +0.0176 and +0.0222 R-L. Despite using only 60M parameters, both outperform a fine-tuned Qwen-2.5-3B (0.2335 R-L, $50\times$ larger). In an exploratory, single-seed study on 15 XL-Sum languages in three sets, EWAD+CPDP beats the CE baseline on 10/15 (mean +0.0044, not significant at $n{=}15$); the gains concentrate in the set where the two teachers are complementary.

Scope. Our positive results hold in one regime: a wide teacher–student gap (247M$\rightarrow$60M, about $4\times$) with an in-domain specialist teacher, which is typical of on-device summarization for a low-resource language. They do not extend to a narrow gap: with Qwen-2.5 32B/14B teachers and a 3B student, no KD variant we tried, including EWAD+CPDP, beats direct fine-tuning on ROUGE (§5.3; CHAD was not tested there). We study one task, abstractive summarization, and treat the multilingual results as exploratory.

Contributions. **(i)** An empirical characterization of uniform-KD failure on a large low-resource benchmark, showing that cheap teacher- and student-side statistics barely predict per-sample KD usefulness. **(ii) CHAD**, a per-sample gating method based on validation-gradient alignment and a learned gate, which needs only one teacher. **(iii) EWAD+CPDP**, a token-level confidence-weighted KD objective combined with a cross-vocabulary geometric constraint. **(iv)** An exploratory 15-language analysis of where EWAD+CPDP helps and where it does not, which suggests a simple teacher-selection rule to test on new languages.

2 Related Work

Knowledge distillation. Hinton et al. (2015) introduced temperature-scaled KD, and Kim and Rush (2016) extended it to sequence generation. Later variants distill intermediate layers (Jiao et al., 2020; Sanh et al., 2019), relations between examples (Park et al., 2019), or several teachers (You et al., 2017; Fukuda et al., 2017); they add teacher assistants to bridge large capacity gaps (Mirzadeh et al., 2020), or target summarization directly (Shleifer and Rush, 2020). These methods apply the KD term uniformly across samples. Larger or more accurate teachers are not always better teachers (Cho and Hariharan, 2019; Mirzadeh et al., 2020), which is one reason we scope our claims by teacher–student gap.

Selective and adaptive KD. Several works show that the KD signal is not equally useful everywhere. Wang et al. (2021) find that distilling on some words hurts neural machine translation, observe that KD and gold cross-entropy gradients on the output distribution often point in different directions, and distill only the top-$r\%$ of words ranked by the student's cross-entropy, selected per batch or with a global queue. Zhou et al. (2021) identify "regularization samples" whose soft labels increase bias and re-weight soft labels per sample instead of discarding them. Koo et al. (2025) let the teacher intervene only where teacher and student token probabilities diverge, and Wen et al. (2023) change the divergence itself, using mode-seeking or mode-covering f-divergences. Outside KD, focal loss (Lin et al., 2017) down-weights easy examples. These methods select or weight by output statistics: student cross-entropy, teacher–student discrepancy, or confidence. CHAD instead scores each sample by whether its KD gradient points in the same direction as the gold-reference validation gradient, in parameter space, and our probe shows that such output statistics barely predict this score (§5.2).

Influence estimation and data selection. Per-sample influence on a held-out loss can be estimated with influence functions (Koh and Liang, 2017) or, more cheaply, by tracing gradient dot products (TracIn; Pruthi et al., 2020). Validation-gradient alignment has been used to re-weight noisy training examples (Ren et al., 2018) and to select instruction-tuning data (Xia et al., 2024). CHAD applies the same first-order idea to the KD term only, so the gold cross-entropy term is never down-weighted, and it amortizes the scores with a cheap gate, so per-sample gradients are needed only on a small probe. To our knowledge, validation-gradient alignment has not been used before to weight the KD term per sample.

Low-resource and multilingual summarization. Bangla and many other languages remain under-resourced (Joshi et al., 2020). XL-Sum (Hasan et al., 2021) provides BBC article–summary pairs in 44 languages and fine-tunes mT5 (Xue et al., 2021) on them; BanglaT5 (Bhattacharjee et al., 2023) is a Bangla-specific encoder–decoder. Prior Bangla summarization ranges from extractive systems (Islam et al., 2020) to hybrid pipelines built on BanglaT5 (Hasib et al., 2023). For multilingual models, vocabulary pruning (Abdaoui et al., 2020) is a common compression alternative to distillation. §5.4 applies EWAD+CPDP to mT5 across 15 XL-Sum languages.

3 Method

We propose two reliability-aware distillation methods. CHAD (§3.1) weights each sample by its measured validation impact; EWAD+CPDP (§3.2) weights each token by teacher confidence and adds a cross-vocabulary geometric constraint. Both modify the standard KD objective

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\, w(x)\, \mathcal{L}_{\text{KD}}, \tag{1}$$

where $\mathcal{L}_{\text{KD}}$ is the temperature-scaled KL divergence between teacher and student token distributions, averaged over target positions, and standard KD sets $w(x) = 1$. For decoder positions t with label mask m_t,

$$\mathcal{L}_{\text{KD}} = \frac{\tau^2}{\sum_t m_t} \sum_t m_t\, \mathrm{KL}\big(p^T_{t,\tau} \,\|\, p^S_{t,\tau}\big), \tag{2}$$

where $p^T_{t,\tau}$ and $p^S_{t,\tau}$ are the teacher and student token distributions at temperature τ.

3.1 CHAD: per-sample counterfactual gating

Probe labeling. Figure 1 summarizes the workflow. We first train a CE-only student (the A1 baseline) and freeze its parameters θ. On a probe subset $\mathcal{P}$ of 5,000 training examples, we compute each sample's KD gradient $g_{\text{KD}}(x) = \nabla_\theta \mathcal{L}_{\text{KD}}(\theta, x)$. The reference direction is the mean gold-reference gradient on a fixed validation set $\mathcal{V}$,

$$g_{\text{val}} = \frac{1}{|\mathcal{V}|} \sum_{(x',y')\in\mathcal{V}} \nabla_\theta \mathcal{L}_{\text{CE}}(\theta; x', y'), \tag{3}$$

where y' is the reference summary. g_{val} contains no teacher term, so the reference direction does not depend on how reliable the teacher is on $\mathcal{V}$. We score each probe sample by

$$s(x) = \frac{\langle g_{\text{val}},\, g_{\text{KD}}(x)\rangle}{\|g_{\text{val}}\|\,\|g_{\text{KD}}(x)\|} \in [-1, 1]. \tag{4}$$

If $s(x) > 0$, a gradient step on the KD loss of x also lowers the gold-reference validation loss, to first order; if $s(x) < 0$, it raises it. $\mathcal{V}$ is drawn once and shared by all probe samples, so its sampling noise affects every score in the same way. A binary label $u(x) = \mathbf{1}[s(x) > 0]$ marks aligned samples; the gate regresses on the continuous score.

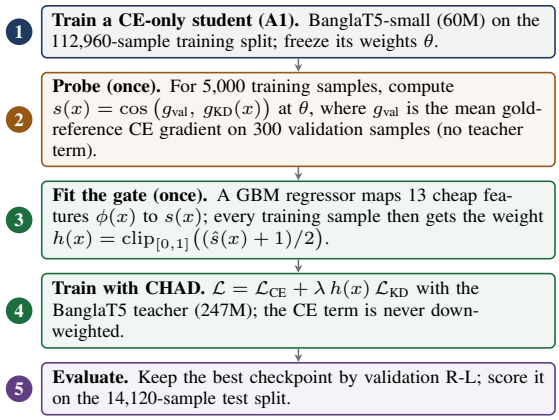


Figure 1: CHAD pipeline on the full BanSum data. Steps 2 and 3 are one-time costs: per-sample gradients are needed only on the probe, and the gate extends the score to the whole training set.

Gate training. For each probe sample we extract 13 cheap features $\phi(x)$: teacher statistics (entropy, top-1 probability, top-1/top-2 margin), student statistics (CE loss, output entropy, student–teacher KL), agreement between the teacher's output and the reference (ROUGE-L and Sentence-BERT similarity), and length and surface statistics (e.g., source length, compression and novelty ratios). We fit a `GradientBoostingRegressor` (Pedregosa et al., 2011) (300 trees, depth 4, learning rate 0.05, subsample 0.8) on $\{(\phi(x), s(x))\}$ and turn its prediction $\hat{s}(x)$ into a weight $h(x) = \text{clip}_{[0,1]}\big((\hat{s}(x)+1)/2\big)$ for every training sample.

Training objective. CHAD uses the gate score as a per-sample KD weight:

$$\mathcal{L}_{\text{CHAD}} = \mathcal{L}_{\text{CE}} + \lambda\, h(x)\, \mathcal{L}_{\text{KD}}. \tag{5}$$

The CE term is never down-weighted, so every sample still teaches the student through its reference summary; only the KD term is scaled. A mislabeled sample therefore loses at most part of its KD signal, never its gold supervision.

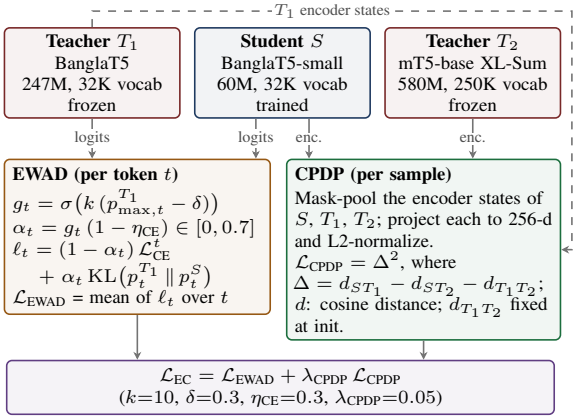


Figure 2: EWAD+CPDP on BanSum. EWAD blends CE and KD per token by teacher confidence; CPDP adds a geometric constraint from a second teacher whose vocabulary differs from the student's, so logit-level KL is undefined for it.

3.2 EWAD+CPDP: per-token confidence with cross-vocabulary geometry

EWAD. EWAD (entropy-weighted adaptive distillation) gates the KD term per token by teacher confidence (Figure 2). At each decoder position t, let $p^T_{\max,t}$ be the teacher's top-1 probability. EWAD computes a sigmoid confidence gate

$$g_t = \sigma\big(k\,(p^T_{\max,t} - \delta)\big), \tag{6}$$

with k=10 and δ=0.3. The effective KD weight is $\alpha_t = g_t\,(1 - \eta_{\text{CE}})$ with η_{CE}=0.3, so $\alpha_t \in [0, 0.7]$, and the per-token loss is

$$\ell_t = (1 - \alpha_t)\,\mathcal{L}^t_{\text{CE}} + \alpha_t\,\text{KL}\big(p^T_t \,\|\, p^S_t\big), \tag{7}$$

averaged over decoder positions to obtain $\mathcal{L}_{\text{EWAD}}$. The teacher is trusted more at high-confidence positions, while the gold signal is kept at low-confidence ones; the floor η_{CE} caps the KD weight at $1 - \eta_{\text{CE}}$, which prevents over-reliance on the teacher even when it is fully confident. We chose these values so that the gate changes softly rather than as a hard cutoff, around a top-1 probability that leaves meaningful mass on both sides of the threshold. They were not tuned on validation, and we did not sweep them (see Limitations).

CPDP. CPDP adds a second teacher whose vocabulary differs from the student's: mT5-base fine-tuned on XL-Sum (580M parameters, 250K-token vocabulary), for which logit-level KL with the student is undefined. Let $\mathbf{h}_S, \mathbf{h}_{T_1}, \mathbf{h}_{T_2}$ be mask-pooled encoder hidden states of the student and the two teachers, projected by learned linear maps into a shared 256-dimensional space and L2-normalized. With pairwise cosine distance $d(\cdot,\cdot)$, CPDP penalizes violations of the triangle-distance constraint $d_{ST_1} - d_{ST_2} \approx d_{T_1T_2}$:

$$\mathcal{L}_{\text{CPDP}} = \big(d_{ST_1} - d_{ST_2} - d_{T_1T_2}\big)^2, \tag{8}$$

where $d_{T_1T_2}$ is computed once at initialization and held fixed as a geometric anchor. The student is thus pushed to a position whose distances to the two teachers reflect the teachers' mutual distance, which gives structural supervision from the second teacher without vocabulary alignment.

Combined objective. The full EWAD+CPDP objective is

$$\mathcal{L}_{\text{EC}} = \mathcal{L}_{\text{EWAD}} + \lambda_{\text{CPDP}}\,\mathcal{L}_{\text{CPDP}}, \tag{9}$$

with $\lambda_{\text{CPDP}} = 0.05$. Implementation details are in Appendix A.

3.3 Baselines

We compare against a CE-only student (**A1**) and three choices of $w(x)$ in Eq. (1): **uniform** (A2, $w(x) = 1$, standard KD); an **entropy gate** (A4) trained on the same probe labels as CHAD but restricted to teacher-distribution and length features, which isolates teacher confidence; and a **ROUGE gate** (A5) trained on ROUGE-L between teacher-generated and reference summaries plus length, which isolates teacher–reference agreement. Both ablation gates use CHAD's probe-and-gate pipeline with a restricted feature set. CHAD itself is configuration A6.

4 Experimental Setup

Dataset. BanSum (Hasan et al., 2024) is a Bangla news summarization corpus whose reference summaries were generated with Gemini Pro. We use 141,200 article–summary pairs (articles $\leq$1,000 BanglaT5 tokens) and a deterministic 80/10/10 split (seed 42): 112,960 training, 14,120 validation, and 14,120 test samples. The test split is held out from training, gate fitting, and model selection. Ablations and the seed study also use a 20K-sample subset ("20K").

Models. The student is `csebuetnlp/banglat5_small` (60M) and the KD teacher is `csebuetnlp/banglat5` (247M) (Bhattacharjee et al., 2023), fine-tuned on the BanSum training split for 5 epochs (learning rate 5×10^{-5}, effective batch 8). Both use the same 32,128-token SentencePiece vocabulary, so full-vocabulary KL (Eq. 2) applies directly.

Training. The A1–A6 students train for 13 epochs (20K) or 8 epochs (full data) with AdamW (Loshchilov and Hutter, 2019), learning rate 5×10^{-5}, effective batch 8, and bf16. We set $\lambda = 0.5$, with $\tau = 2.0$ at 20K and $\tau = 0.5$ on the full data. The maximum source length is 512 (20K) or 768 (full data). We keep the checkpoint with the best validation ROUGE-L (early-stopping patience 5 on the full data). EWAD+CPDP uses its own settings (learning rate 3×10^{-4}, effective batch 32, patience 2), because it trains new projection layers jointly with the student (Appendix A, Table 6).

Probe and gate. The probe uses $|\mathcal{P}| = 5{,}000$ training samples and a fixed set $\mathcal{V}$ of 300 samples drawn once at random from the validation split. Both gradients are taken at the frozen A1 checkpoint in `train()` mode. Because $\mathcal{V}$ comes from the validation split, it overlaps with the data used for checkpoint selection; the test split is never used for probing, gate fitting, or model selection. CHAD's gate is a GBM regressor on the continuous score; the A4/A5 gates are logistic regressions on feature subsets, trained on the binary labels $u(x)$.

Evaluation. We compute ROUGE (Lin, 2004) on whitespace tokens, because the default tokenizer of common ROUGE implementations drops non-Latin characters and would discard Bangla text. We report R-1, R-2, R-L, BLEU (Papineni et al., 2002), and SemSim, the cosine similarity between Sentence-BERT (Reimers and Gurevych, 2019) embeddings of the generated and reference summaries, on the full test split. Decoding uses 4-beam search with at most 200 new tokens. Table 1 reports one run per method; we measure seed variance separately for A1 and CHAD on the 20K subset (§5.5).

Compute and cost. Experiments ran on single workstation and consumer GPUs (RTX A6000, 5070, and 4070 Ti Super), about 70 GPU-hours in total. CHAD has a multi-stage cost. Besides its own training run, it needs (1) a CE-only student, which is our A1 baseline; (2) one probe pass, about 30 minutes on one GPU for 5,000 samples; and (3) gate fitting, under one minute on CPU, plus one feature-extraction pass to score the training set. Stages (2) and (3) are one-time costs that can be reused across runs. EWAD+CPDP has no probe or gate stage. Its extra cost is a forward pass through a second teacher at every training step; all three models fit on one 24 GB GPU.

5 Results

All results are on the held-out 14,120-sample BanSum test split, unseen during training, gate fitting, and model selection.

5.1 Main results: reliability-aware vs. baseline distillation

Table 1 reports our central comparison. All BanglaT5-small students (60M) share the same fine-tuned BanglaT5 teacher (247M), training data, and decoding. The A1–A6 rows also share the optimizer, batch size, learning rate, and epoch budget, so they differ only in how per-sample KD weights are set. EWAD+CPDP uses its own optimization settings (§4). Qwen-2.5-3B is a larger decoder-only reference fine-tuned on the same training split.

Three observations stand out. **Standard KD gives essentially no gain.** A2 improves R-L by only +0.0003 over A1, well within the seed variation we measure in §5.5. With a $4\times$ capacity gap and an in-domain teacher, the uniform KD signal is indistinguishable from zero. **Heuristic gates do not help either.** A4 and A5 sit within 0.0001 R-L of CE, so teacher confidence and surface agreement are poor proxies for KD usefulness (§5.2). **Both reliability-aware methods improve substantially over standard KD.** CHAD gains +0.0176 over CE (+0.0173 over KD), and EWAD+CPDP gains +0.0222 over CE (+0.0219 over KD). Both improve every metric in Table 1. The two methods are within 0.0046 R-L of each other, with EWAD+CPDP ahead on every metric. Both 60M students outperform the fine-tuned Qwen-2.5-3B, which is $50\times$ larger (Figure 3).

5.2 Probe analysis: why so many samples hurt

On the full data, the alignment probe labels 48.7% of the 5,000 probe samples as KD-helpful ($s(x) > 0$) and 51.3% as KD-harmful or neutral ($s(x) \leq 0$).

Alignment is spread evenly around zero. Scores are nearly symmetric around zero: samples with $s(x) > 0$ average $+0.162 \pm 0.110$, and those with $s(x) \leq 0$ average -0.168 ± 0.112. The harmful half is therefore not a thin tail of near-zero scores; many samples have clearly negative alignment. Because many scores also lie near zero, the binary label is noisy near the threshold, which is one reason CHAD's gate regresses on the continuous score rather than on the label.

Method	Gating signal	R-1	R-2	R-L	BLEU	SemSim	ΔR-L
Baselines							
A1: CE only	none (no KD)	0.2804	0.1337	0.2365	10.06	0.7976	—
A2: standard KD	uniform $w(x)$=1	0.2803	0.1340	0.2368	10.29	0.7950	+0.0003
A4: entropy gate	teacher confidence (sample)	0.2804	0.1339	0.2364	10.19	0.7961	−0.0001
A5: ROUGE gate	teacher–reference surface	0.2802	0.1336	0.2364	10.20	0.7958	−0.0001
Qwen-2.5-3B FT	fine-tune only (3B params)	0.2815	0.1389	0.2335	5.90	0.7364	−0.0030
Reliability-aware methods (ours, 60M params)							
A6: CHAD	**counterfactual (sample)**	**0.2989**	**0.1459**	**0.2541**	**11.11**	**0.8208**	**+0.0176**
EWAD+CPDP	**conf. (token) + geom.**	**0.3162**	**0.1603**	**0.2587**	**12.60**	**0.8301**	**+0.0222**
Ref: teacher (247M)				0.2998			

Table 1: Results on the 14,120-sample BanSum test split (one run per method). All students are 60M-parameter BanglaT5-small; ΔR-L is relative to A1 (CE only). Both reliability-aware methods (bold) outperform every baseline on every metric, including Qwen-2.5-3B, which is fine-tuned on the same training split and is 50× larger. SemSim is the Sentence-BERT cosine similarity to the reference. The teacher row is for reference only.

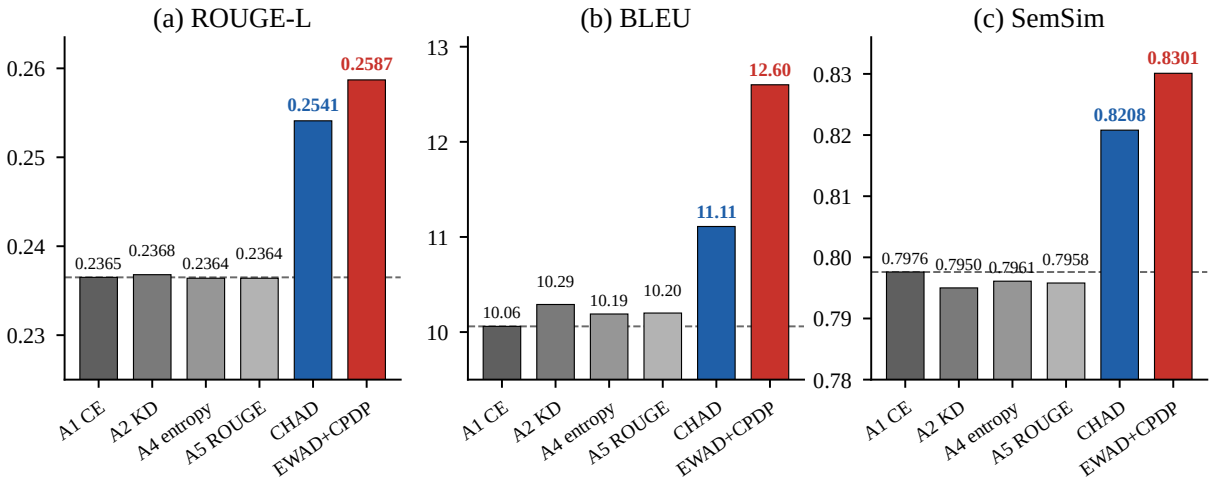


Figure 3: BanSum test results (14,120 samples; values from Table 1). Standard KD and the entropy- and ROUGE-gated variants stay at the CE baseline (dashed line), while CHAD (blue) and EWAD+CPDP (red) separate clearly on ROUGE-L, BLEU, and SemSim.

Cheap features barely predict KD usefulness. Table 2 correlates each gate feature with the binary label. The three teacher-confidence features (entropy, top-1 probability, and top-1/top-2 margin) reach only $|r| \leq 0.084$. Each is statistically significant ($p < 10^{-7}$ at $n = 5{,}000$), yet each explains less than 1% of the variance. Student-side and surface features are just as weak ($|r| \leq 0.073$). For reference, the alignment score itself correlates with the label at $r = +0.83$; this is by construction, since the label is its sign, so it is a ceiling rather than an independent predictor. The gap between this ceiling and every cheap feature helps explain why the A4 and A5 heuristic gates stay at the CE baseline (Table 1), and why CHAD measures alignment directly on a probe and then fits its gate on all features jointly.

Length has little effect. The helpful ratio is 50.1% for medium-length and 47.0% for long articles, so the split is not driven by article length.

5.3 Comparison to a fine-tuned large language model

We fine-tune Qwen-2.5-3B (Yang et al., 2024), a decoder-only LLM 50× larger than our student, on the same BanSum training split. It reaches R-L 0.2335, below both CHAD (0.2541) and EWAD+CPDP (0.2587). Appendix D gives the data-scaling results and an 8-experiment ablation with Qwen-2.5 teachers. In that narrow-gap setting (32B/14B teachers, 3B student), no KD variant, including EWAD and EWAD+CPDP, exceeds direct fine-tuning of the 3B student on ROUGE. Confidence-only weighting even collapses to R-L 0.1169, against 0.2160 without KD, so confidence-

Feature	r	$\lvert r\rvert$
teacher_entropy	$+0.076$	0.076
teacher_max_prob	-0.081	0.081
teacher_margin	-0.084	0.084
student_ce_loss	$+0.071$	0.071
student_entropy	$+0.073$	0.073
student_teacher_kl	$+0.066$	0.066
novelty_ratio	$+0.070$	0.070
source_words	-0.039	0.039
compression_ratio	$+0.010$	0.010
$s(x)$, *label source*	$+0.830$	0.830

Table 2: Pearson r between each gate feature and the binary label `kd_useful` on the 5,000-sample probe. The teacher-confidence features are significant ($p < 10^{-7}$) yet explain $< 1\%$ of the variance. The last row is the alignment score from which the label is derived; its r is high by construction and serves only as a reference ceiling.

gated distillation can be actively harmful when the gap is narrow. We attribute the encoder–decoder advantage on BanSum to an inductive bias suited to small-scale abstractive summarization, strong task-specific Bangla pretraining (BanglaT5), and distillation that makes full use of the small model's capacity. We do *not* claim this extends to settings without a strong specialist teacher or to the narrow-gap LLM regime. A plausible explanation for the latter is that the 3B student already learns the task from the references alone, which leaves the teacher little useful signal to add and little harmful signal to suppress.

5.4 Exploratory multilingual study of EWAD+CPDP

To probe whether EWAD+CPDP transfers beyond Bangla, and to map where it does *not*, we run it on 15 typologically diverse XL-Sum languages in three sets of five. **All results in this section come from a single seed, and the aggregate gain is not significant; we treat them as exploratory.** We carry forward only EWAD+CPDP (with EWAD as an ablation), not CHAD: per-language probe labeling would require 15 separate probe stages, whereas EWAD+CPDP has no language-specific stage.

Setup. Per language we sample 1,500 XL-Sum examples (Hasan et al., 2021) and split them 80/10/10 (seed 42; 1,200/150/150). The small size tests whether the gating mechanism generalizes across languages, not gains from scale. This is a separate regime from BanSum, not a scaled-up version of it. The student is `google/mt5-small` (300M) (Xue et al., 2021); the EWAD teacher is `google/mt5-base` (580M), fine-tuned per language, so the teacher–student gap is only about $2\times$; and the CPDP teacher is the public `csebuetnlp/mT5_multilingual_XLSum`. All three models share mT5's vocabulary, so here CPDP acts as a geometric constraint even though logit-level KL would be possible. Students train for 5 epochs with learning rate 3×10^{-4}, batch size 16, and 512/128 source/target lengths. EWAD uses $k = 10$, $\delta = 0.3$, and $\eta_{\text{CE}} = 0.3$; CPDP uses $\lambda_{\text{CPDP}} = 0.1$. For each language we train three students: **Baseline** (CE only), **EWAD**, and **EWAD+CPDP**.

Language sets. We defined the sets before training, by expected teacher agreement. **Set 1** (Hindi, Urdu, Sinhala, Indonesian, Nepali): four South Asian languages and Indonesian, on which we expected both teachers to be strong ("teachers agree"). **Set 2** (Amharic, Hausa, Pashto, Portuguese, Russian): five branches (Semitic, Chadic, Iranian, Romance, Slavic) with varied teacher exposure ("teachers complement"). **Set 3** (Persian, Punjabi, Vietnamese, Marathi, Thai): languages on which we expected both teachers to be weak ("teachers struggle together").

Aggregate results. Across all 15 languages (per-language scores in Appendix E, Table 9), EWAD+CPDP beats the CE-only baseline on 10/15 languages (mean R-L gain +0.0044), and EWAD beats it on 9/15 (+0.0018); EWAD+CPDP beats EWAD on 8/15 (+0.0026). EWAD+CPDP is the best of the three on 6/15 languages, EWAD on 6/15, and the baseline on 3/15. None of these mean differences is significant at n=15 (paired t: p=0.46 for EWAD+CPDP vs. Baseline and p=0.68 for EWAD+CPDP vs. EWAD; Wilcoxon agrees, p=0.25).

Per-set picture. Table 3 and Figure 4 average R-1/R-2/R-L within each set. On **Set 2**, EWAD+CPDP improves over the baseline on all three metrics (R-L $+0.0162$, R-1 $+0.0182$, R-2 $+0.0099$) and also beats EWAD on all three. On **Set 1**, EWAD+CPDP gains on R-1 and R-2 but only $+0.0018$ on R-L, and EWAD is best on R-2 and R-L. On **Set 3**, EWAD+CPDP wins R-2 but not R-1 or R-L. This pattern is consistent with EWAD+CPDP helping most when the two teachers occupy distinct regions of representation space, being redundant

Set	Metric	Base	EWAD	EC	$\Delta_{\text{EC-Base}}$
Set 1	R-1	0.1978	0.1964	**0.2005**	+0.0027
	R-2	0.0552	**0.0671**	0.0629	+0.0077
	R-L	0.1709	**0.1744**	0.1727	+0.0018
Set 2	R-1	0.1863	0.1970	**0.2045**	+0.0182
	R-2	0.0461	0.0495	**0.0560**	+0.0099
	R-L	0.1601	0.1673	**0.1763**	+0.0162
Set 3	R-1	**0.1476**	0.1506	0.1438	−0.0038
	R-2	0.0423	0.0463	**0.0487**	+0.0064
	R-L	**0.1293**	0.1241	0.1247	−0.0046

Table 3: Per-set means over five languages (mT5-small student). **Single seed; exploratory.** EC = EWAD+CPDP; bold marks the best of the three per row.

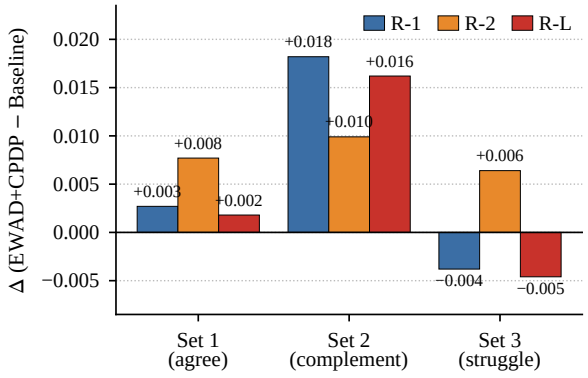


Figure 4: Per-set Δ(EWAD+CPDP − Baseline) on R-1, R-2, and R-L (**single seed; exploratory**). Set 2 (teachers complement) improves on all three metrics; Set 1 (teachers agree) shows small mixed gains; Set 3 (teachers struggle) is negative on R-1 and R-L.

when they overlap, and adding little when both are weak. With five languages per set and one seed, it is a hypothesis to test rather than an established effect.

Baseline difficulty. Gains tend to be larger for languages with weaker baselines, but the trend is not significant (r=−0.31, p=0.25; Figure 5 in Appendix E), so we read it only as a direction. The three largest gains (Russian +0.048, Sinhala +0.044, Amharic +0.016) are on languages with baseline R-L below 0.10. The two large losses (Nepali −0.038, Punjabi −0.034) are Indo-Aryan languages with mid-range baselines; every other loss is smaller than 0.004.

5.5 Seed variance

To measure seed variance, we rerun A1 (CE) and A6 (CHAD) on the 20K subset with two more seeds (7 and 13); multi-seed retraining on the full data was beyond our compute budget. CHAD

Seed	A1 (CE)	A6 (CHAD)	Δ
42	0.2211	0.2244	+0.0034
7	0.2095	0.2146	+0.0051
13	0.2137	0.2190	+0.0053
mean	0.2148	0.2193	+0.0046
std	0.0059	0.0049	0.0010

Table 4: ROUGE-L across three seeds on the 20K subset (std: sample standard deviation, $n = 3$). Scores are rounded, so Δ may differ in the last digit from the difference of the rounded values. Paired 95% t-interval on the gap: $[+0.0020, +0.0072]$.

beats CE in every seed (mean gap +0.0046, std 0.0010), and the paired 95% t-interval on the gap, $[+0.0020, +0.0072]$, excludes zero (Table 4). The 20K subset has a much higher KD-helpful ratio (80.3%) than the full data (48.7%), so CHAD has less room to improve there. The full-data single-seed gap of +0.0176 R-L (Table 1) is about three times the largest per-method seed std we observe (0.0059). EWAD+CPDP was not part of this seed study.

6 Analysis

Which method to use. The two methods are within 0.0046 R-L of each other on BanSum and suit different situations. CHAD needs only one teacher, gives interpretable per-sample scores that can be used to triage training data, and pays a one-time probe cost (§4). EWAD+CPDP has no probe or gate stage and was slightly stronger, but it needs a second teacher; through CPDP it can use a teacher with an incompatible vocabulary, which logit-level KL cannot. We therefore recommend CHAD when only one teacher is available or when per-sample interpretability matters, and EWAD+CPDP when a heterogeneous second teacher is at hand.

Gate quality and ceiling effects. The CHAD gate has only modest held-out AUC (0.640), yet its downstream R-L gain is large. The gate weights samples softly rather than filtering them, so errors in $h(x)$ cause bounded attenuation of the KD term, and the aggregate effect depends mainly on getting the broad direction right. With 51.3% of probed samples negatively aligned, even imperfect reweighting can plausibly recover much of the KD signal that uniform weighting cancels out. We did not isolate the gate's contribution from that of the alignment signal (see Limitations).

When does CPDP help on top of EWAD? The multilingual per-set pattern (§5.4) suggests one mechanism: CPDP's geometric anchor adds value when the two teachers occupy distinct regions of representation space (Set 2), is redundant when they overlap (Set 1), and adds little when both are weak (Set 3). This suggests a simple rule to test on a new language. Before training, score both teachers on a small target-language dev set and compare their outputs (e.g., by ROUGE between them). Skip CPDP if both teachers are weak, expect little from it if both are strong and agree, and try it when both are reasonable but differ. We expect BanSum, which pairs a Bangla-specialist EWAD teacher with a multilingual generalist CPDP teacher, to be close to the favorable case. However, Table 1 has no EWAD-only row, so CPDP's separate contribution on BanSum is not measured.

7 Conclusion

This paper argues that the central question in low-resource sequence-to-sequence distillation is not only how much teacher signal to use, but *when* to trust it. On BanSum, standard KD gives only a marginal gain over CE (+0.0003 R-L), and gradient-alignment probing shows that 51.3% of probed training samples have KD gradients that do not support validation-loss improvement. Common reliability proxies, such as teacher confidence and ROUGE agreement, fail to identify these samples, which motivates estimating distillation usefulness more directly.

We introduced two reliability-aware approaches. CHAD measures sample-level KD usefulness through validation-gradient alignment, amortizes it with a gradient-boosted gate, and needs only one teacher. EWAD+CPDP combines token-level confidence weighting with a cross-vocabulary geometric constraint from a second teacher. Both improve substantially over standard KD (+0.0173 and +0.0219 R-L), and both do so with a 60M student that outperforms a fine-tuned Qwen-2.5-3B. These results hold for a wide teacher–student gap with an in-domain teacher; with a narrow gap, no KD variant we tried beat direct fine-tuning. The exploratory multilingual study suggests that EWAD+CPDP is most useful when the teachers are complementary and less reliable when they are redundant or jointly weak; across 15 languages it beats CE on 10/15, with a small, non-significant mean gain of +0.0044.

Overall, the results suggest a practical principle for selective KD: route teacher supervision by measured reliability, not assumed quality. Future work should make reliability signals cheaper to estimate, test them beyond summarization and across more seeds, and study how teacher diversity, calibration, and student capacity jointly decide when distillation helps rather than hurts.

Limitations

Scope and capacity gap. Our positive results come from one task, one main benchmark, and a wide teacher–student gap (247M→60M) with an in-domain teacher. In a narrow-gap LLM setting (Qwen-2.5 32B/14B → 3B), no KD variant we tried beat direct fine-tuning (Appendix D), so we expect gating-based KD to help mainly when the student is much smaller than the teacher. Other tasks, domains, and gap sizes are untested.

Seeds and statistical power. Table 1 reports one run per method. We measured seed variance only for A1 and CHAD on the 20K subset (three seeds), so EWAD+CPDP's BanSum result is a single run. The multilingual study uses one seed and 150 test samples per language, its aggregate gain is not significant (paired t, $p = 0.46$), and its per-set patterns rest on five languages per set. A multi-seed study, with three or more seeds per language and method, is the main follow-up we plan.

Ablations we did not run. We did not compare the learned gate with simpler uses of the alignment signal, such as using $s(x)$ directly as the KD weight or dropping samples with $s(x) < 0$, so we cannot yet say how much of CHAD's gain comes from the gate rather than from the signal. We did not run the batch- and global-level selection of Wang et al. (2021) as a baseline, and we did not sweep EWAD's k, δ, and η_{CE}, which were set by design. On BanSum we report EWAD+CPDP but not EWAD alone, so CPDP's own contribution there is not isolated.

Probe design. The gate reaches only moderate AUC (0.640). The alignment label is a first-order estimate computed once, at the CE-trained checkpoint; we did not test whether labels change during training, so a sample that is harmful early may help later. CHAD's soft weighting and unattenuated CE term bound the cost of such stale labels. The validation set $\mathcal{V}$ overlaps the split used for checkpoint selection; the test split is untouched.

Optimization settings. EWAD+CPDP used a higher learning rate and a larger batch than A1–A6 (Table 6), and we did not retrain the CE baseline with those settings, so part of its margin over A1 may come from optimization rather than from the objective.

Teachers and languages. CPDP needs a second teacher with a different vocabulary. We did not explore alternatives to the mT5-base XL-Sum teacher, and the anchor distance $d_{T_1T_2}$ depends on this choice. CHAD was not tested multilingually.

Data and metrics. BanSum's references were generated with Gemini Pro, so ROUGE measures agreement with that system rather than with human-written summaries, and we ran no human evaluation. We omit BERTScore (Zhang et al., 2020) from Table 1 because the only EWAD+CPDP value we have used a different backbone (`xlm-roberta-base`) from the other rows (`bert-base-multilingual-cased`). On the comparable rows, CHAD reaches 0.7518 against 0.7419–0.7423 for A1–A5 and 0.7437 for Qwen-2.5-3B.

Ethics Statement

All datasets used in this work (BanSum and XL-Sum subsets) are publicly available, and all model and dataset artifacts were used in accordance with their licenses, terms of use, and redistribution restrictions. We collected no new data involving human subjects. To the best of our knowledge, the datasets contain no personally identifiable information beyond what appears in published news articles. News text may reflect biases of the source media, and BanSum's references were produced by an LLM; generated summaries can inherit and amplify such biases, so user-facing deployment should include appropriate review. All experiments ran on single GPUs, about 70 GPU-hours in total. We release code and trained model checkpoints to support reproducibility. We used AI assistants (Claude) for writing assistance and coding support during the preparation of this work.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of multilingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4794–4802.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient knowledge distillation from an ensemble of teachers. In *Proceedings of Interspeech 2017*, pages 3697–3701.

Mahmudul Hasan, Abdullah Ibne Hanif Arean, and Md. Mosaddek Khan. 2024. BanSum: A dataset for Bangla abstractive article summarization with multiple sentences. Mendeley Data, Version 1.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Khan Md Hasib, Md. Atiqur Rahman, Mustavi Ibne Masum, Friso De Boer, Sami Azam, and Asif Karim. 2023. Bengali news abstractive summarization: T5 transformer and hybrid approach. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 539–545. IEEE.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mahimul Islam, Fariha Nuzhat Majumdar, Asadullahhil Galib, and Md Moinul Hoque. 2020. Hybrid text summarizer for Bangla document. *International Journal of Computer Vision and Signal Processing*, 10(1):27–38.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020.

TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Jahyun Koo, Yerin Hwang, Yongil Kim, Taegwan Kang, Hyunkyung Bae, and Kyomin Jung. 2025. SWITCH: Studying with teacher for knowledge distillation of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3733–3746, Albuquerque, New Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4334–4343. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation. *Preprint*, arXiv:2010.13002.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466, Online. Association for Computational Linguistics.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834, Toronto, Canada. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning.

In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54104–54132. PMLR.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1285–1294, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. 2021. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In *International Conference on Learning Representations*.

A EWAD+CPDP implementation details

This appendix supplements §3.2 with implementation details that are not essential to the main exposition.

Numerical stability. Student and teacher logits are cast to float32 before the softmax in the EWAD loss (Eq. 7). The teacher is held in bf16 in memory, but its logits are upcast on the fly for the KL term; this prevents the underflow we observed when computing $\log p_t^T$ from bf16 logits with very sharp distributions ($p_{\max,t}^T > 0.99$). The CPDP projection layers are kept in float32 throughout.

CPDP teacher details. The second teacher used by CPDP on BanSum is the public `csebuetnlp/mT5_multilingual_XLSum` checkpoint (Hasan et al., 2021), an mT5-base model (580M parameters) fine-tuned on all XL-Sum languages, including Bangla, with a 250,112-token SentencePiece vocabulary. Its vocabulary differs from BanglaT5's, so token ids do not correspond and logit-level KL is undefined; this motivates the encoder-state projection used in CPDP. The mT5 teacher uses the same maximum source length as the student (512 tokens at 20K, 768 tokens on the full data).

Projection layer initialization. The projection maps W_S, W_{T_1}, W_{T_2} use Xavier-uniform initialization. The anchor distance $d_{T_1T_2}$ in Eq. (8) is computed once after initialization and held fixed (detached) throughout training. In our runs, recomputing $d_{T_1T_2}$ during training let both projections adapt until the CPDP loss drifted toward zero, which removed the geometric supervision; the detach is therefore necessary.

Training hyperparameters. EWAD+CPDP is trained with learning rate 3×10^{-4}, batch size 8 with gradient accumulation 4 (effective batch 32), bf16 mixed precision, at most 8 epochs with early-stopping patience 2 on validation ROUGE-L, warmup ratio 0.06, and gradient clipping at max-norm 1.0. The maximum source length is 512 tokens at 20K and 768 tokens on the full data; the maximum target length is 200 tokens throughout. Both teachers are loaded in bf16 and frozen in eval mode. All three models (the student and two teachers) fit on a single 24 GB GPU.

Relationship to CHAD. EWAD+CPDP and CHAD address the same broad question—when should teacher supervision dominate over gold supervision?—at different granularities and with different signals. EWAD+CPDP works per token with teacher confidence (EWAD) and per sample with inter-teacher geometry (CPDP); CHAD works per sample with counterfactual validation impact. On BanSum (Table 1) the two methods are within 0.0046 R-L of each other (EWAD+CPDP slightly higher), which suggests that both signals capture a comparable amount of reliability information on this benchmark; §6 discusses when each is preferable.

B Probe analysis details

Helpful ratio by dataset scale. On the 20K subset, the gradient-alignment probe labels 80.3% of its 1,000 probe samples as KD-helpful. On the full data, the same procedure labels only 48.7% of 5,000 samples as KD-helpful and 51.3% as KD-harmful or neutral. One plausible reason is the broader topic and length distribution of the full data,

which adds more ambiguous articles on which the teacher's soft labels are less reliable. We did not isolate the cause: the two probes also differ in τ, $|\mathcal{P}|$, and $|\mathcal{V}|$ (Table 6).

Gate quality. Table 5 reports held-out AUC for all three gate configurations on the probe set.

Gate	Features	AUC	Type
A4: entropy	teacher conf. + len	0.619	logistic
A5: ROUGE	teacher–ref. + len	0.554	logistic
A6: CHAD	all 13 features	0.640	GBM regressor

Table 5: Gate AUC on held-out probe samples. CHAD's AUC of 0.640 exceeds that of the heuristic gates. The downstream ROUGE-L gap between CHAD and the heuristic gates (Table 1) is much larger than the AUC gap, which suggests that modest gains in gate accuracy can have an outsized effect on training.

C Hyperparameters and training details

Table 6 lists all hyperparameters for the BanSum experiments.

D Qwen-2.5 scaling and decoder-only ablation

This appendix supports two claims in §5.3: (i) fine-tuned Qwen-2.5-3B reaches ROUGE-L 0.2335 on BanSum, and (ii) EWAD+CPDP with Qwen-2.5 teachers does not outperform direct fine-tuning of the 3B student. BLEU is on a 0–100 scale as in Table 1; BS-F1 is BERTScore-F1 (Zhang et al., 2020) with the shared `bert-base-multilingual-cased` backbone.

Qwen-2.5-3B data-scaling results. Table 7 shows Qwen-2.5-3B at two data scales. The full-data model (ROUGE-L 0.2335) is the one cited in the main paper. It improves on every metric over the 20K model, so Qwen-2.5-3B still benefits from more data and is not saturated at the smaller scale.

8-experiment ablation with Qwen-2.5 teachers. Table 8 reports an ablation that uses Qwen-2.5-32B and 14B, loaded in 4-bit NF4 (Dettmers et al., 2023), as teachers for a Qwen-2.5-3B student trained with LoRA (Hu et al., 2022) on the 20K subset. Direct fine-tuning (Baseline, no KD) gives the highest ROUGE-L. Single-teacher and fixed-weight dual-teacher KD come close to the baseline but do not surpass it on ROUGE or BLEU. Confidence-only weighting degrades sharply (ROUGE-L 0.1169), and full EWAD and EWAD+CPDP also fall below the no-KD baseline.

Hyperparameter	20K	Full data
Shared by A1–A6 (CE, KD, gates, CHAD)		
Optimizer	AdamW	
Learning rate	5×10^{-5}	
Per-device batch	4	
Gradient accum.	2 (eff. batch = 8)	
Precision	bf16	
λ (KD wt.)	0.5	
Temperature τ	2.0	0.5
Max source tokens	512	768
Max target tokens	128	200
Beam width	4	
Epochs	13	8
Early stop patience	—	5
CHAD-specific		
Probe size $\lvert\mathcal{P}\rvert$	1,000	5,000
Val. probe $\lvert\mathcal{V}\rvert$	200	300
Gate type	logistic	GBM regressor
GBM trees / depth	—	300 / 4
GBM LR / subsample	—	0.05 / 0.8
EWAD+CPDP-specific		
Learning rate	3×10^{-4}	
Effective batch	32 (bs=8, accum=4)	
Warmup ratio	0.06	
Grad clip norm	1.0	
Early stop patience	2	
EWAD k / δ	10 / 0.3	
EWAD η_{CE}	0.3	
CPDP λ_{CPDP}	0.05	
Projection dim	256	

Table 6: Hyperparameters for the BanSum experiments. CHAD and EWAD+CPDP use different learning rates because EWAD+CPDP trains new projection layers jointly with the student; the higher rate suits the newly initialized projections, while the lower rate for CHAD matches standard seq2seq fine-tuning.

E Per-language multilingual results (exploratory)

Table 9 gives the per-language ROUGE-L scores behind §5.4, and Figure 5 plots each language's gain against its baseline score. **All numbers come from a single seed with 150 test samples per language; the aggregate gain is not significant (paired t, $p = 0.46$), so these results are exploratory.**

Scale	R-1	R-2	R-L	BLEU	BS-F1	Sem
20K	.2661	.1241	.2160	5.52	.7389	.7175
Full	.2815	.1389	.2335	5.90	.7437	.7364
Δ	+.015	+.015	+.018	+0.38	+.005	+.019

Table 7: Qwen-2.5-3B fine-tuned on BanSum at two data scales. Even with the full data, its ROUGE-L of 0.2335 remains below both encoder–decoder reliability-aware students (CHAD: 0.2541; EWAD+CPDP: 0.2587).

Experiment	R-1	R-2	R-L	BLEU	BS-F1	Sem
Baseline (no KD)	**.2661**	**.1241**	**.2160**	**5.52**	**.7389**	.7175
Single-T 32B	.2614	.1210	.2114	5.35	.7364	.7116
Single-T 14B	.2640	.1165	.2113	5.23	.7377	.7219
Fixed weights	.2632	.1166	.2104	5.21	.7380	**.7220**
Confidence only	.1529	.0519	.1169	2.39	.6657	.6454
Agreement only	.2270	.0917	.1756	4.10	.7181	.7016
EWAD full	.2282	.0927	.1767	4.20	.7190	.7031
EWAD+CPDP	.2246	.0907	.1740	4.07	.7167	.7012

Table 8: Reliability-aware distillation with Qwen-2.5 teachers (32B + 14B → 3B + LoRA, 20K BanSum subset). Direct fine-tuning is best on every metric except SemSim, where two KD variants are marginally higher. This contrasts with our encoder–decoder setting (Table 1), where a 4× teacher–student gap lets both CHAD and EWAD+CPDP improve substantially over the no-KD baseline. We tentatively attribute the difference to the narrow gap: when the student can already learn the task from the references alone, gating-based routing adds variance without giving cleaner supervision.

Language	Baseline	EWAD	EWAD+CPDP
Set 1 – teachers agree			
Hindi	0.1774	0.1712	**0.1784**
Urdu	0.2512	**0.2746**	0.2558
Sinhala	0.0918	0.1187	**0.1362**
Indonesian	**0.1507**	0.1342	0.1478
Nepali	**0.1832**	0.1730	0.1450
Set 2 – teachers complement			
Amharic	0.0868	**0.1306**	0.1032
Hausa	0.2048	0.2128	**0.2162**
Pashto	0.2466	**0.2514**	0.2434
Portuguese	0.1661	0.1448	**0.1744**
Russian	0.0961	0.0969	**0.1440**
Set 3 – teachers struggle together			
Persian	0.2375	0.2054	**0.2473**
Punjabi	**0.1589**	0.1245	0.1250
Vietnamese	0.1046	**0.1190**	0.1059
Marathi	0.1106	**0.1321**	0.1070
Thai	0.0351	**0.0396**	0.0383
Mean	0.1534	0.1553	**0.1579**

Table 9: Per-language ROUGE-L on XL-Sum (mT5-small student). **Single seed; exploratory.** Bold marks the best of the three per language.

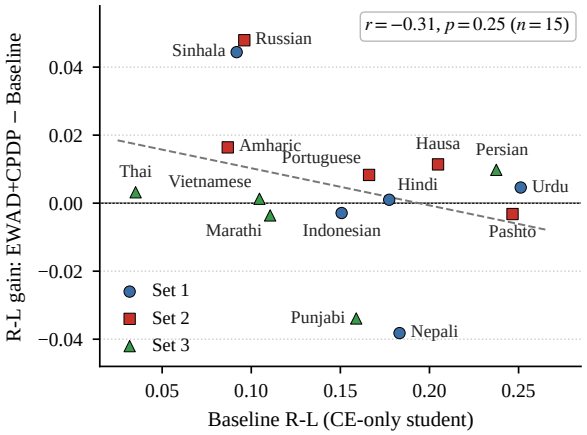


Figure 5: Baseline R-L vs. EWAD+CPDP gain per language (r=−0.31, p=0.25, n=15; single seed). The trend is directional only and not significant.