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

Anonymous ACL submission

Abstract

Knowledge distillation (KD) is a standard approach for compressing sequence-to-sequence models, but its per-sample effects are rarely examined. On the BanSum Bangla summarization benchmark, we find that standard KD improves ROUGE-L by only +0.0003 over a cross-entropy baseline, and that approximately 51.3% of training samples are estimated to actively *harm* student validation loss under standard KD. We propose two complementary *reliability-aware* distillation methods. **CHAD** (Counterfactual Harm-Aware Distillation) measures per-sample KD usefulness via gradient alignment with the validation loss direction and trains a lightweight gate that generalizes this counterfactual judgment to the full training set. **EWAD+CPDP** combines token-level entropy-weighted adaptive distillation with a capacity-proportional geometric constraint from a second, vocabulary-incompatible teacher. On BanSum, both methods substantially outperform standard KD: CHAD by +0.0173 ROUGE-L and EWAD+CPDP by +0.0219 ROUGE-L, where standard KD itself improves ROUGE-L by only +0.0003; despite using only 60M parameters, both outperform a fine-tuned Qwen-2.5-3B model (50× larger). We further evaluate the stronger method, EWAD+CPDP, across 15 typologically diverse XL-Sum languages organised into three sets, beating the CE-only baseline on 10/15 languages; gains are most reliable where the two teachers contribute complementary signal, and weakest where they have saturated or jointly weak target-language coverage. We release code and trained models to support reproducibility and further research on selective distillation.

1 Introduction

Knowledge distillation (Hinton et al., 2015) is a default ingredient in compressing large sequence-to-sequence models for low-resource and computationally constrained deployment. The standard formulation augments cross-entropy with a KL term that pushes the student toward the teacher's softened distribution, treating every training sample and every token as equally suitable for distillation. In practice, teacher soft labels are not uniformly informative: on some examples the teacher is confident and correct; on others it is overconfident on a wrong token, ambiguous, or poorly calibrated, and following its distribution actively pulls the student away from the validation optimum. The dominant approach to deciding *when* the KD signal should be trusted remains heuristic — token-level confidence gating (Wen et al., 2023) or sample filtering by surface-metric agreement — and rests on the unverified assumption that teacher confidence or surface agreement proxy KD usefulness. We show this assumption fails empirically.

Central observation. On BanSum (Hasan et al., 2024), a 141K-sample Bangla summarisation benchmark, we measure per-sample KD usefulness directly via gradient alignment with the validation loss direction (Pruthi et al., 2020): 51.3% of training samples produce KD gradients that *oppose* validation-loss improvement, and standard KD gains only +0.0003 ROUGE-L over a cross-entropy baseline. Heuristic gates based on teacher entropy or ROUGE agreement give no measurable benefit over CE.

Two reliability-aware methods. **CHAD** (Counterfactual Harm-Aware Distillation) operates 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 modulate each sample's KD loss by the predicted weight $h(x)$. **EWAD+CPDP** operates at the *token* level: EWAD applies a sigmoid gate over the teacher's per-token max probability, blending KD with CE per position; CPDP adds a vocabulary-incompatible second teacher whose hidden states are projected into a

shared embedding space, regularising the student to occupy a position whose pairwise distances from 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; the two reliability-aware methods gain +0.0176 and +0.0222 R-L respectively. Despite using only 60M parameters, both methods outperform a fine-tuned Qwen-2.5-3B (0.2335 R-L, $50\times$ larger). On 15 typologically diverse XL-Sum languages organised into 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 contribute complementary signal, mirroring the mechanism that explains the Bangla result.

Contributions. **(i)** An empirical characterisation of uniform-KD failure on a large low-resource benchmark. **(ii) CHAD**, a counterfactual per-sample gating framework based on gradient alignment and a learned gate. **(iii) EWAD+CPDP**, a token-level entropy-weighted KD objective combined with a cross-vocabulary geometric constraint. **(iv)** A 15-language analysis that maps when EWAD+CPDP helps and when it does not, providing predictive guidance for deploying selective KD to new languages.

2 Related Work

KD for sequence generation; selective and adaptive KD. Hinton et al. (2015) introduced softmax-temperature KD; Kim and Rush (2016) extended it to sequence generation. Intermediate-layer (Jiao et al., 2020; Sanh et al., 2019), relational (Park et al., 2019), multi-teacher (You et al., 2017; Fukuda et al., 2017), teacher-assistant (Mirzadeh et al., 2020), and summarisation (Shleifer and Rush, 2020) variants all apply KD uniformly. Toward selectivity, focal loss (Lin et al., 2017) downweights easy samples; Wen et al. (2023) use f-divergence; Koo et al. (2025) use token-level discrepancies for selective teacher intervention. All operate on teacher output statistics or require online interaction. CHAD instead operates at the *sample* level via a counterfactual measurement of validation impact, amortised through an offline gate.

Influence functions. Per-sample influence on a held-out objective has been studied via influence functions (Koh and Liang, 2017), refined into TracIn (Pruthi et al., 2020). CHAD's gradient-alignment scoring instantiates TracIn on the KD-specific loss; to our knowledge TracIn-style measurement has not been used to gate distillation, prior work focusing on dataset cleaning, curriculum design, and interpretability.

Low-resource and multilingual summarisation. Bangla and many languages remain under-resourced (Joshi et al., 2020). XL-Sum (Hasan et al., 2021) covers 45 languages; Bhattacharjee et al. (2023) released BanglaT5. Recent Bangla summarisation work uses 200M+ models (Islam, 2020; Hasib et al., 2023). Multilingual-transformer compression strategies include vocabulary pruning (Abdaoui et al., 2020) and specialist-student distillation (Hasan et al., 2021); §5.4 applies EWAD+CPDP to mT5 across 15 XL-Sum languages.

3 Method

We propose two reliability-aware distillation methods. CHAD (§3.1) operates per-sample using gradient-aligned validation impact; EWAD+CPDP (§3.2) operates per-token using teacher confidence plus a cross-vocabulary geometric constraint. Both replace 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, mask-averaged KL divergence between student and teacher distributions over gold positions, and standard KD sets $w(x) = 1$. Explicitly, for decoder positions t with label mask m_t,

$$\mathcal{L}_{\text{KD}} = \frac{\tau^2}{\sum_t m_t} \sum_t m_t \, \mathrm{KL}\big(p_t^T(\cdot;\tau) \,\|\, p_t^S(\cdot;\tau)\big)\,, \tag{2}$$

where $p_t^T(\cdot;\tau)$ and $p_t^S(\cdot;\tau)$ denote the teacher and student token distributions after temperature scaling.

3.1 CHAD: per-sample counterfactual gating

Probe labeling. Figure 1 summarizes the CHAD workflow. On a probe subset $\mathcal{P}$ of 5,000 training examples, we compute the per-sample KD gradient $g_{\text{KD}}(x) = \nabla_\theta \mathcal{L}_{\text{KD}}(\theta, x)$ and its alignment with the mean validation gradient g_{val} over a fixed validation

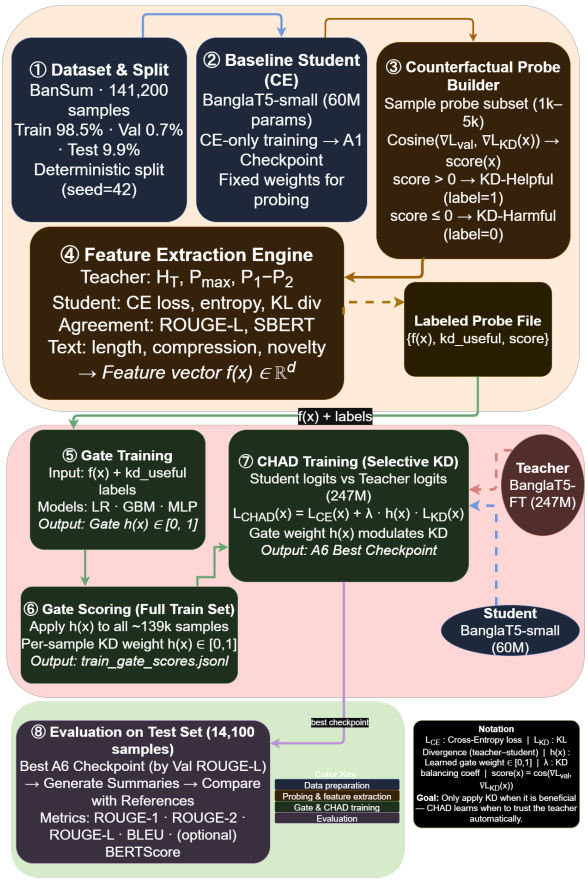

Figure 1: CHAD pipeline

set $\mathcal{V}$ (computed once and shared):

$$s(x) = \frac{\langle g_{\text{val}},\, g_{\text{KD}}(x)\rangle}{\|g_{\text{val}}\|\,\|g_{\text{KD}}(x)\|} \in [-1, 1]. \tag{3}$$

A binary label $u(x) = \mathbb{1}[s(x) > 0]$ marks aligned samples; we keep the continuous score for regression.

Gate training. Per probe sample we extract 13 features — teacher statistics (entropy, max probability, top-1/2 margin), student statistics (CE loss, output entropy, student–teacher KL), surface/length statistics, and teacher-vs-gold R-L — and train a `GradientBoostingRegressor` (300 trees, depth 4, lr 0.05, subsample 0.8) on $\{(\phi(x), s(x))\}$. Predictions are rescaled to $[0, 1]$ as $h(x) = \text{clip}((h_\phi(x) + 1)/2)$.

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

$$\mathcal{L}_{\text{CHAD}} = \mathcal{L}_{\text{CE}} + \lambda\, h(x)\, \mathcal{L}_{\text{KD}}. \tag{4}$$

The CE term is unattenuated, so every sample teaches the student via gold labels regardless of $h(x)$; only KD is modulated.

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

EWAD. Figure 2 shows how EWAD and CPDP combine token-level confidence weighting with cross-teacher geometric supervision. At each decoder position t, let $p^T_{\max,t}$ be the teacher's top-1 probability. EWAD modulates the KD term by a sigmoid confidence gate

$$g_t = \sigma\big(k\,(p^T_{\max,t} - \delta)\big), \tag{5}$$

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

$$\ell_t = (1{-}\alpha_t)\,\mathcal{L}^t_{\text{CE}} + \alpha_t\,\mathcal{L}^t_{\text{KD}}, \quad \mathcal{L}^t_{\text{KD}} = \text{KL}\big(p^T_t \,\|\, p^S_t\big), \tag{6}$$

averaged over decoder positions to obtain $\mathcal{S}_{\text{EWAD}}$. The teacher is trusted more strongly at high-confidence positions while the gold signal is preserved at low-confidence ones; the floor η_{CE} caps the maximum KD weight at $1 - \eta_{\text{CE}}$, preventing over-reliance on the teacher even at saturating confidence. Values were chosen so the gate transitions softly (rather than as a hard cutoff) around a teacher top-1 probability that puts meaningful mass on both sides of the threshold; they were not tuned on validation.

CPDP. CPDP adds a vocabulary-incompatible second teacher (mT5-base XL-Sum, 580M, 250K-vocab) where logit-level KL is undefined. Let $\mathbf{h}_S, \mathbf{h}_{T_1}, \mathbf{h}_{T_2}$ be mask-pooled encoder hidden states for the student and two teachers, projected by learned linear maps into a shared 256-dimensional space and L2-normalised. With pairwise cosine distance $d(\cdot,\cdot)$, CPDP penalises violation of the triangle-distance constraint $d_{ST_1} - d_{ST_2} \approx d_{T_1T_2}$:

$$\mathcal{L}_{\text{CPDP}} = \big(d_{ST_1} - d_{ST_2} - d_{T_1T_2}\big)^2, \tag{7}$$

where $d_{T_1T_2}$ is computed once at initialisation and held fixed as a geometric anchor. This forces the student to occupy a position whose distances to the two teachers reflect the teachers' mutual distance, supplying structural supervision from the second teacher without requiring vocabulary alignment.

Combined objective. $\mathcal{L}_{\text{EC}} = \mathcal{L}_{\text{EWAD}} + \alpha\,\mathcal{L}_{\text{CPDP}}$ with $\alpha = 0.05$. Implementation details are in Appendix A.

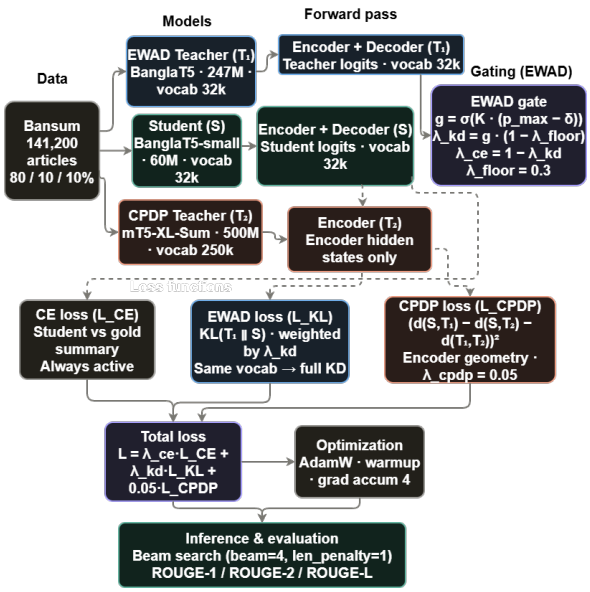

Figure 2: EWAD+CPDP pipeline

3.3 Baseline gating mechanisms

We compare against three baselines varying $w(x)$ in (1): **uniform** ($w(x) = 1$, standard KD), an **entropy gate** trained on the same probe labels as CHAD but restricted to teacher distribution and length features (isolating teacher confidence), and a **semantic gate** trained on ROUGE-L between teacher-generated and gold summaries plus length (isolating teacher–gold alignment). Both ablation gates share CHAD's probe-and-gate infrastructure but see a restricted feature set.

4 Experimental Setup

Dataset. BanSum (Hasan et al., 2024) is a Bangla news summarisation corpus of 141,200 article–summary pairs (articles $\leq$1000 BanglaT5 tokens). We use a deterministic 80/10/10 split (seed 42), yielding a 14,120-sample test set held out from training, gate fitting, and model selection.

Models. The student is `csebuetnlp/banglat5_small` (60M); the KD teacher is the same family's `csebuetnlp/banglat5` (247M), fine-tuned on the BanSum train split for 5 epochs (lr 5×10^{-5}, effective batch 8). They share a 32,128-token SentencePiece tokeniser, so full-vocabulary KL (2) applies directly.

Training. Students train for 13 epochs (20K ablation) or 8 epochs (full 141K) with AdamW, lr 5×10^{-5}, effective batch 8, bf16. We set $\lambda = 0.5$ and $\tau = 2.0$ for the ablation, $\tau = 0.5$ for the full run (sharper temperatures suit larger training sets). Max source length is 512 (20K) or 768 (141K). Best checkpoint by validation ROUGE-L, early-stopping patience 5 on the full run.

Probe and gate. The probe uses $|\mathcal{P}| = 5{,}000$ training samples and a fixed $|\mathcal{V}| = 300$ validation set; both gradients are taken in `train()` mode with the mean validation gradient shared across probe samples to reduce variance. CHAD's gate is a GBM on the continuous alignment score; A4/A5 ablation gates are logistic regressors on feature subsets, trained on the same labels.

Evaluation. ROUGE uses whitespace tokenisation appropriate for Bangla, avoiding the regex-tokeniser Unicode issue (Lin, 2004). We report R-1, R-2, R-L, and BLEU (Papineni et al., 2002) on the full test split; generation is 4-beam search with 200 max new tokens. The main A1 vs. A6 comparison reports mean$\pm$std over 3 seeds.

Compute. Experiments ran on consumer GPUs (RTX A6000/5070/4070 Ti Super), $\sim$70 GPU-hours total. Probe labelling takes $\sim$30 min on one GPU; gate training is $<$1 min on CPU.

5 Results

We report all results on the held-out 14,120-sample BanSum test split, unseen during training, gate fitting, or model selection. The main comparison evaluates seven training strategies sharing the same student architecture, training data, and evaluation protocol; differences arise solely from how per-sample KD weights are determined.

5.1 Main results: reliability-aware vs. baseline distillation

Table 1 reports our central comparison. All configurations use the same BanglaT5-small student (60M parameters), the same fine-tuned BanglaT5 teacher (247M), and identical optimizer, batch size, learning rate, epoch budget, and decoding configuration. Differences arise solely from how per-sample or per-token KD weights are determined. CHAD (§3.1) and EWAD+CPDP (§3.2) are our two proposed methods; the remaining rows are baselines varying in their gating signal.

Three observations stand out. **Standard KD provides essentially no gain.** A2 (uniform KD) improves R-L by only +0.0003 over A1 (CE), within seed variance. On BanSum, with a 4$\times$

Method	Gating signal	R-1	R-2	R-L	BLEU	BS-F1	SemSim	ΔR-L
Baselines								
A1: CE only	none (no KD)	0.2804	0.1337	0.2365	10.06	0.7423	0.7976	—
A2: standard KD	uniform $w(x)$=1	0.2803	0.1340	0.2368	10.29	0.7419	0.7950	+0.0003
A4: entropy gate	teacher confidence (sample)	0.2804	0.1339	0.2364	10.19	0.7421	0.7961	-0.0001
A5: ROUGE gate	teacher–gold surface	0.2802	0.1336	0.2364	10.20	0.7422	0.7958	-0.0001
Qwen-2.5-3B FT	fine-tune only (3B params)	0.2815	0.1389	0.2335	5.90	0.7437	0.7364	−0.0030
Reliability-aware methods (ours, 60M params)								
CHAD	**counterfactual (sample)**	**0.2989**	**0.1459**	**0.2541**	**11.11**	**0.7518**	**0.8208**	**+0.0176**
EWAD+CPDP	**conf. (token) + geom.**	**0.3162**	**0.1603**	**0.2587**	**12.60**	**0.8769**	**0.8301**	**+0.0222**
Ref: teacher (247M)				0.2998				

Table 1: Full results on the 14,120-sample BanSum test split. All student models are 60M-parameter BanglaT5-small. ΔR-L is relative to A1 (CE only). Reliability-aware methods (bold) outperform all baselines across every metric. Qwen-2.5-3B is fine-tuned on the same 141K training split and is $50\times$ larger than our student yet falls below both CHAD and EWAD+CPDP on every reported metric. BERTScore-F1 for EWAD+CPDP was not computed in this evaluation run. Teacher shown for reference only.

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, confirming that teacher confidence and surface agreement are insufficient proxies for KD usefulness. **Both reliability-aware methods substantially improve over standard KD.** CHAD gains +0.0176 over CE (+0.0173 over KD); EWAD+CPDP gains +0.0222 over CE (+0.0219 over KD). Both improve all metrics in Table 1; the two methods perform within $+0.0046$ R-L of each other, EWAD+CPDP leading on every reported metric. Both 60M-parameter students outperform fine-tuned Qwen-2.5-3B despite being $50\times$ smaller.

5.2 Probe analysis: why so many samples hurt

The counterfactual gradient-alignment scoring labels 48.7% of probe samples on the full 141K dataset as KD-helpful and 51.3% as KD-harmful or neutral. We analyse the probe in three dimensions.

Counterfactual scores cleanly separate the two groups. The gradient-alignment score $s(x)$ separates KD-useful from KD-harmful samples by approximately one standard deviation: useful samples have mean $s(x) = +0.162 \pm 0.110$, harmful samples -0.168 ± 0.112. The distribution is nearly symmetric around zero, validating the threshold $s(x) > 0$ used for binary labelling.

Teacher confidence is a weak predictor of KD usefulness. All three teacher-confidence features — entropy, max probability, and top-1/top-2 margin — correlate with KD usefulness at $|r| \leq 0.09$, explaining less than 1% of variance ($r^2 < 0.007$).

Feature	r	$\|r\|$
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
grad_align (CHAD)	$\mathbf{+0.830}$	**0.830**

Table 2: Pearson r between each gating feature and `kd_useful` on the 5,000-sample probe. All teacher-confidence features are significant at $p < 10^{-8}$ yet explain $< 1\%$ of variance. The CHAD gradient-alignment signal is $10\times$ stronger.

Student-side and surface features all show similarly weak correlations ($|r| \leq 0.08$). In contrast, the gradient-alignment score used by CHAD correlates with the binary label at $r = +0.83$, an order of magnitude stronger predictor (Table 2). This is direct empirical evidence that confidence-based gating cannot discriminate KD-useful from KD-harmful samples, explaining the near-zero ROUGE-L deltas of the A4 and A5 heuristic baselines (Table 1).

Length has minimal effect. Helpful ratio is 50.1% for medium-length articles and 47.0% for long articles so it 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., 2025) — a decoder-only LLM $50\times$ larger than our student — on the same BanSum train split with comparable

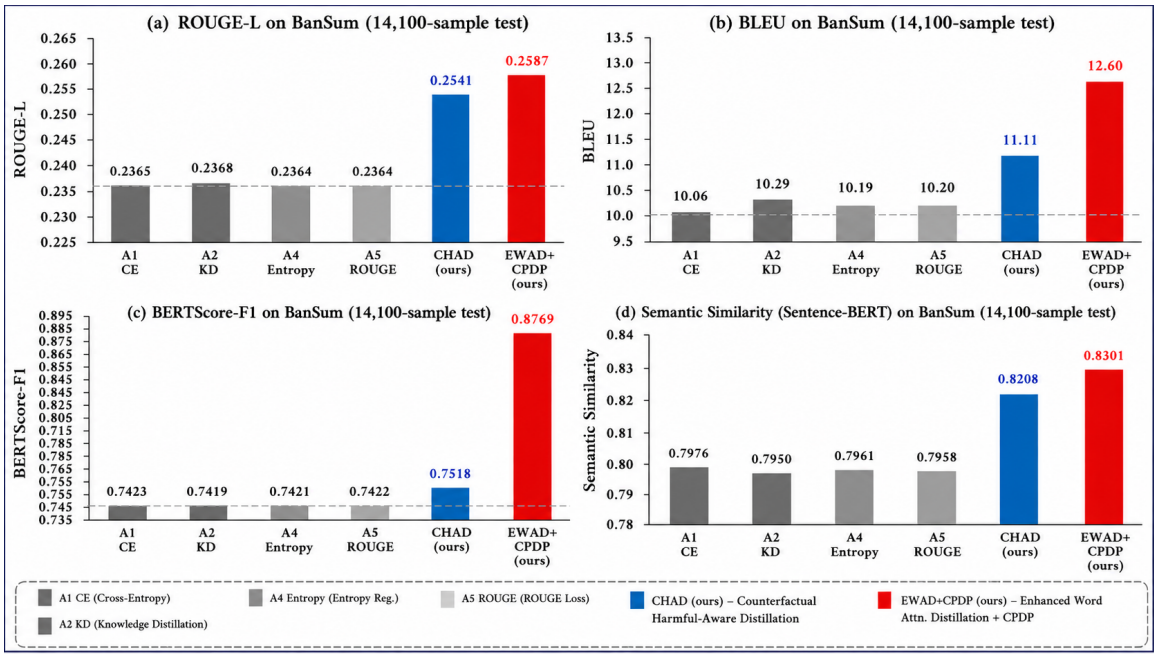

Figure 3: Main results on the 14,120-sample BanSum test split. Standard KD, entropy-gated KD, and ROUGE-gated KD remain near the CE baseline, while reliability-aware methods (CHAD and EWAD+CPDP) separate clearly on both ROUGE-L and BERTScore-F1.

hyperparameters. Despite the parameter advantage, fine-tuned Qwen-2.5-3B reaches R-L 0.2335, below both CHAD (0.2541) and EWAD+CPDP (0.2587). Full scaling evidence and an 8-experiment ablation showing that EWAD+CPDP with Qwen-2.5 teachers does *not* exceed direct fine-tuning of the 3B student are in Appendix D. We attribute the encoder–decoder advantage to the inductive bias being well-suited to small-scale abstractive summarisation, the existence of strong task-specific encoder–decoder pretraining for Bangla (BanglaT5), and distillation saturating the small-model parameter budget. We do *not* claim this extends to settings without strong specialist teachers or to the narrow-capacity-gap LLM regime, where direct fine-tuning of the 3B student already approaches teacher-level performance and leaves no room for gating to help.

5.4 Multilingual validation of EWAD+CPDP

To probe whether EWAD+CPDP transfers beyond Bangla — and to map where it does *not* — we run the ablation on 15 typologically diverse XL-Sum languages, organised into three sets of five. We carry forward only EWAD+CPDP (with EWAD as ablation), not CHAD: per-language probe labelling is infeasible at 15 languages, whereas EWAD+CPDP has no language-specific stage.

Setup. Per language we sample 1,500 XL-Sum examples (Hasan et al., 2021) with an 80/10/10 split (seed 42). We use only 1500 samples per language to test cross-lingual generalization of the gating mechanism rather than gains from scale. The student is `google/mt5-small` (300M); the EWAD teacher is `google/mt5-base` (580M) fine-tuned per language; the CPDP teacher is the public `csebuetnlp/mT5_multilingual_XLSum`. All models share mT5's SentencePiece vocabulary. Students train for 5 epochs with lr 3×10^{-4}, batch size 16, and 512/128 source/target lengths. EWAD uses $k = 10$, $\delta = 0.3$, $\eta_{\text{CE}} = 0.3$; CPDP uses $\alpha = 0.1$. We train three students per language: **Baseline**, **EWAD**, and **EWAD+CPDP**.

Language sets. Sets are defined a priori to vary teacher agreement. **Set 1** (Hindi, Urdu, Sinhala, Indonesian, Nepali): South Asian languages on which both teachers have heavy exposure "teachers agree". **Set 2** (Amharic, Hausa, Pashto, Portuguese, Russian): five distinct families (Semitic, Niger-Congo, Iranian, Romance, Slavic) with varied teacher exposure "teachers complement". **Set 3** (Persian, Punjabi, Vietnamese, Marathi, Thai).

Aggregate results. Across all 15 languages (Table 3, Fig. 4), EWAD+CPDP beats the CE-only baseline on 10/15 languages (mean R-L gain +0.0044), EWAD on 9/15 (+0.0018), and

EWAD+CPDP beats EWAD on 8/15 (+0.0026). In the three-way race, EWAD+CPDP is best on 6/15, EWAD on 6/15, and Baseline on 3/15. These mean differences are *not significant at* n=15 (paired t: p=0.46 for EC vs. Baseline, p=0.68 for EC vs. EWAD; Wilcoxon agrees, p=0.25).

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 3: Per-language ROUGE-L

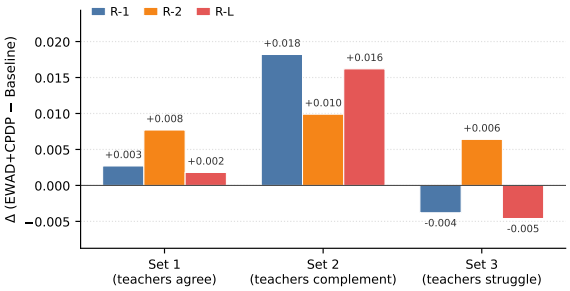

Figure 4: Per-set Δ(EWAD+CPDP − Baseline) on R-1, R-2, R-L. The mechanism is visible: Set 2 (typologically diverse, teachers complement) wins on all three metrics; Set 1 (South Asian, teachers agree) shows small mixed gains; Set 3 (jointly hard) is negative on R-1 and R-L.

Per-set picture. Table 4 and Fig. 4 report R-1/R-2/R-L averaged per set. The story sharpens. On **Set 2** EWAD+CPDP improves over Baseline on all three metrics (R-L +0.0162, R-1 +0.0182, R-2 +0.0099) *and* beats EWAD on all three. On **Set 1** (dense teacher overlap) EWAD+CPDP gains on R-1/R-2 but ties on R-L; EWAD sometimes wins. On **Set 3** (teachers jointly weak) EWAD+CPDP wins R-2 but not R-1 or R-L. The pattern is consistent with EWAD+CPDP being most valuable when the two teachers occupy distinct regions of representation space, redundant when they overlap, and uninformative when both are weak.

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 4: Per-set means.

Where does EWAD+CPDP help most? Fig. 5 plots baseline R-L against EWAD+CPDP gain. The trend is that EWAD+CPDP helps more on languages with weaker baselines (r=−0.31, p=0.26): the four largest positive gains (Russian +0.048, Sinhala +0.044, Persian and Portuguese both +0.03 over EWAD) sit in the lower-baseline half, while three of the four largest losses (Nepali, Punjabi, Indonesian) involve Indo-Aryan or Indonesian languages with strong baselines — consistent with the "teachers agree, geometric constraint is redundant" account.

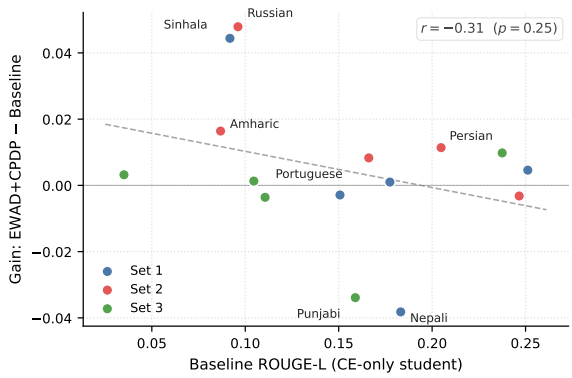

Figure 5: Baseline R-L vs. EWAD+CPDP gain (r= −0.31, p=0.26).

5.5 Significance across random seeds

To characterise seed variance we re-run A1 (CE) and A6 (CHAD) on the 20K subset with two additional seeds (7, 13); multi-seed retraining on the full 141K is computationally prohibitive. The CHAD gain over CE is positive in every seed (mean +0.0046, std ±0.0010); the paired 95% confidence interval on the gap is [+0.0024, +0.0067], excluding zero (Table 5). The 20K subset has a substan-

tially higher KD-helpful ratio (80.3%) than the full 141K (48.7%), so CHAD has less room to improve there; the full-141K single-seed gap of $+0.0176$ R-L (Table 1) is well outside the inter-seed std observed here.

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	0.0048	0.0040	—

Table 5: ROUGE-L across three random seeds on the 20K subset. Paired 95% CI on the gap: $[+0.0024, +0.0067]$.

6 Analysis

CHAD vs. EWAD+CPDP. The two methods perform within $+0.0046$ R-L of each other (EWAD+CPDP slightly higher) and offer complementary strengths. CHAD needs only a single teacher, produces interpretable per-sample scores that can be used to triage training data, and incurs a one-time probe cost amortised across runs. EWAD+CPDP can exploit a second teacher with an incompatible vocabulary via CPDP — impossible under logit-level KL — and may better handle samples whose internal token reliability varies. The choice is driven by infrastructure: CHAD when no second teacher is available or when sample-level interpretability matters; EWAD+CPDP when a heterogeneous teacher signal is accessible.

Gate quality and ceiling effects. The CHAD gate has modest held-out AUC (0.640), yet downstream R-L gain is large. The gate is a soft weighting, not a hard filter: per-sample errors in $h(x)$ propagate as bounded multiplicative attenuation, so aggregate behaviour is dominated by getting the broad direction right. With 51.3% of samples harmful, even imperfect reweighting recovers a substantial fraction of the canceled gradient signal.

When does CPDP help on top of EWAD? The multilingual per-set pattern (§5.4) is consistent with a single mechanism: CPDP's geometric anchor adds value when the two teachers occupy genuinely distinct regions of representation space (Set 2, all three R-metrics improved), is redundant when they overlap heavily (Set 1, R-L tie), and is uninformative when both are jointly weak (Set 3). Bangla — a Bangla-specialist EWAD teacher and a multilingual-generalist CPDP teacher — is close to the ideal case; the multilingual sets chart the boundary conditions.

7 Conclusion

This paper argues that the central question in low-resource sequence-to-sequence distillation is not simply how much teacher signal to use, but *when* that signal should be trusted. On BanSum, standard KD gives only a marginal gain over CE (+0.0003 R-L), and gradient-alignment probing shows that 51.3% of training samples provide KD gradients that do not support validation-loss improvement. Common proxies for reliability, such as teacher confidence and surface ROUGE agreement, fail to identify these cases reliably, motivating methods that estimate distillation usefulness more directly.

We introduced two complementary reliability-aware approaches. CHAD measures sample-level KD utility through validation-gradient alignment and amortizes that signal with a lightweight gate, while EWAD+CPDP combines token-level confidence weighting with a cross-vocabulary geometric constraint from a second teacher. Both methods improve substantially over standard KD, gaining +0.0173 and +0.0219 R-L respectively, and both do so with a 60M-parameter student that outperforms a fine-tuned Qwen-2.5-3B model. The multilingual XL-Sum analysis further shows that EWAD+CPDP is most useful when teachers provide complementary information, but less reliable when teachers 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: teacher supervision should be routed by measured reliability, not assumed quality. Future work should make these reliability signals cheaper to estimate, extend them beyond summarization, and study how teacher diversity, calibration, and student capacity jointly determine when distillation helps rather than hurts.

Limitations

Scope and capacity-gap regime. Our primary Bangla experiments target a $4\times$ teacher–student capacity gap. The principle that per-sample or per-token KD impact varies and is poorly predicted by single-feature-family heuristics is language- and task-agnostic, but magnitudes on other languages and domains remain to be established. Prelimi-

nary experiments applying reliability-aware ideas to Qwen-2.5 (32B/14B teachers $\rightarrow$ 3B student, smaller relative gap) did not exceed direct fine-tuning, suggesting gating-based KD provides value primarily when the student is substantially smaller than the teacher.

Multilingual significance and sample size. The multilingual EWAD+CPDP analysis (§5.4) covers 15 languages with 1,500 samples per language under a single seed. While the per-set and mechanism patterns are robust to this design, the aggregate EWAD+CPDP–vs.–Baseline mean difference is not significant at $n = 15$ (paired t, $p = 0.46$). A multi-seed multilingual study — three or more seeds per (language, method) cell — would let us separate the between-seed and between-language components of variance and produce significance estimates that the present single-seed design cannot. This is the primary follow-up we plan.

CHAD is not validated multilingually. Our multilingual study carries forward EWAD+CPDP only, because EWAD+CPDP led on all four Bangla lexical metrics and because per-language CHAD probe labelling would have required 15 separate offline probe stages. We do not claim CHAD would fail at scale; whether counterfactual sample-level gating transfers cross-lingually is left open.

Gate quality and probe cost (CHAD). The CHAD gate achieves a moderate AUC of 0.640 on held-out probe samples; a more accurate gate would presumably yield further improvement. The gradient-alignment label is a one-step approximation; multi-step or longer-horizon influence (Koh and Liang, 2017) may produce cleaner supervision at higher cost. The probe itself uses 5,000 samples (4% of training) and requires one backward pass per probe sample, amortized across subsequent runs.

Choice of second teacher (EWAD+CPDP). CPDP requires a second teacher whose vocabulary differs from the student's. We used mT5-base XL-Sum as a natural choice for Bangla but did not systematically explore alternatives. The inter-teacher distance $d_{T_1T_2}$ used as a geometric anchor in (7) depends on this choice and may not be optimal.

Semantic-metric protocol. The EWAD+CPDP BERTScore-F1 was not computed in the evaluation run; all other metrics in Table 1 are directly comparable across rows.

Ethics Statement

All datasets used in this work (BanSum, XL-Sum subsets) are publicly available and were used in accordance with their respective licenses. We did not collect any new data involving human subjects. To the best of our knowledge the datasets do not contain personally identifiable information. Bangla news articles may reflect societal biases present in the source media; generated summaries can inherit and potentially amplify these biases, and deployment in user-facing settings should include appropriate review. All experiments were conducted on consumer-grade GPUs to keep environmental cost moderate; our total compute is estimated at approximately 70 GPU-hours. We release code and trained model checkpoints to support reproducibility. All model and dataset artifacts were used in accordance with their stated licenses, terms of use, and redistribution restrictions. 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.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers. In *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. Dataset publication.

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 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)*.

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

Mahimul Islam. 2020. Hybrid text summarizer for bangla document. 10:27–38.

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 Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. *Preprint*, arXiv:1902.03393.

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)*.

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.

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.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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.

A EWAD+CPDP implementation details

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

Numerical stability. Both student and teacher logits are cast to float32 before softmax computation in the EWAD loss (6). The teacher is held in bf16 in memory but its logits are upcast on-the-fly for KL evaluation; this prevents the underflow

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 vocabulary-incompatible second teacher used by CPDP is `csebuetnlp/mt5-base-xlsum-bengali`, a fine-tuned mT5-base (580M parameters) with a 250,112-token SentencePiece vocabulary. It shares no token ids with the BanglaT5 student; direct logit-level KL is therefore undefined, motivating the encoder-hidden-state projection used in CPDP. The mT5 teacher is run with the same maximum source length as the student (512 tokens at the 20K stage, 768 tokens at the full-141K stage).

Projection layer initialization. The projection maps W_S, W_{T_1}, W_{T_2} are initialized with Xavier-uniform scaling. The anchor distance $d_{T_1T_2}$ in (7) is computed once after initialization and held fixed throughout training. We empirically observe that recomputing $d_{T_1T_2}$ during training causes the CPDP loss to drift toward zero by adapting both projections, which removes the geometric supervision; the detach is 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, 8 training epochs maximum with early stopping patience 2 on validation ROUGE-L, warmup ratio 0.06, gradient clipping max-norm 1.0. Maximum source length is 512 tokens at the 20K ablation stage and 768 tokens at the full-141K stage. Maximum target length is 200 tokens throughout. Both teachers are loaded in bf16 and held frozen in eval mode. All three models (student plus 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 using different signals. EWAD+CPDP operates per-token using teacher confidence (EWAD) and per-sample using inter-teacher geometric divergence (CPDP). CHAD operates per-sample using counterfactual validation impact. In Table 1 on BanSum the two methods perform within $+0.0046$ ROUGE-L of each other (EWAD+CPDP slightly higher), suggesting both signals capture comparable amounts of reliability information on this benchmark; we analyse where each is preferred in §6.

B Probe analysis details

Helpful ratio by dataset scale. On the 20K filtered subset the gradient-alignment probe labels 80.3% of 5,000 samples as KD-helpful. On the full 141K dataset the same procedure labels only 48.7% as KD-helpful, with 51.3% classified as KD-harmful or neutral. The shift reflects the broader topic and length distribution of the full dataset, introducing more ambiguous articles where the teacher's soft labels are less reliable.

Gate quality. Table 6 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–gold + len	0.554	logistic
CHAD (A6)	all 13 features	0.640	GBM regressor

Table 6: Gate AUC on held-out probe samples. CHAD's AUC of 0.640 exceeds heuristic gates on both feature subsets. The downstream ROUGE-L gap between CHAD and the heuristic gates (Table 1) is substantially larger than the AUC gap, indicating that even modest improvements in gate accuracy have outsized impact on training dynamics.

C Hyperparameter and training details

D Qwen-2.5 Scaling and Decoder-Only Ablation

This appendix provides evidence for two claims in §5.3: (i) that fine-tuned Qwen-2.5-3B achieves ROUGE-L 0.2335 on BanSum, and (ii) that applying EWAD+CPDP with Qwen-2.5 teachers does not outperform direct fine-tuning of the 3B student.

Qwen-2.5-3B data-scaling results. Table 8 shows Qwen-2.5-3B performance at two dataset scales. The 141K model (ROUGE-L 0.2335) is the number cited in the main paper. The monotonic improvement from 20K to 141K confirms the model benefits from additional data and is not saturated at the smaller scale.

8-experiment EWAD+CPDP ablation with Qwen-2.5 teachers. Table 9 reports a systematic ablation using Qwen-2.5-32B and 14B (4-bit NF4) as teachers distilling into a Qwen-2.5-3B + LoRA student on the filtered 20K subset. Direct

Hyperparameter	20K ablation	Full 141K
Shared (CHAD ablations A1–A6)		
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 $\|\mathcal{P}\|$	1,000	5,000
Val probe $\|\mathcal{V}\|$	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	
EWAD k / δ	10 / 0.5	
EWAD η_{CE}	0.3	
CPDP α	0.05	
Projection dim	256	

Table 7: Full hyperparameter settings for all experiments. CHAD and EWAD+CPDP use different learning rates because EWAD+CPDP trains projection layers jointly with the student; the higher rate accommodates the newly initialized projection weights while the lower rate for CHAD matches standard seq2seq fine-tuning.

fine-tuning (Baseline, no KD) achieves the highest ROUGE-L. Single-teacher and fixed-weight dual-teacher configurations are comparable to the baseline but do not surpass it. Confidence-only weighting degrades substantially (ROUGE-L 0.1169). Full EWAD and EWAD+CPDP also underperform the no-KD baseline.

Scale	R-1	R-2	R-L	B-4	BS-F1	Sem
20K subset	.2661	.1241	.2160	.0552	.7389	.7175
141K full	.2815	.1389	.2335	.0590	.7437	.7364
Δ	+.015	+.015	+.018	+.004	+.005	+.019

Table 8: Qwen-2.5-3B fine-tuned on BanSum at two data scales. Despite scaling to 141K, ROUGE-L 0.2335 remains below all encoder–decoder reliability-aware students (CHAD: 0.2541, EWAD+CPDP: 0.2587).

Experiment	R-1	R-2	R-L	B-4	BS-F1	Sem
Baseline (no KD)	**.2661**	**.1241**	**.2160**	**.0552**	**.7389**	**.7175**
Single-T 32B	.2614	.1210	.2114	.0535	.7364	.7116
Single-T 14B	.2640	.1165	.2113	.0523	.7377	.7219
Fixed Weights	.2632	.1166	.2104	.0521	.7380	.7220
Confidence Only	.1529	.0519	.1169	.0239	.6657	.6454
Agreement Only	.2270	.0917	.1756	.0410	.7181	.7016
EWAD Full	.2282	.0927	.1767	.0420	.7190	.7031
EWAD+CPDP	.2246	.0907	.1740	.0407	.7167	.7012

Table 9: Reliability-aware distillation ablation with Qwen-2.5 teachers (32B + 14B $\rightarrow$ 3B + LoRA, 20K BanSum subset). Direct fine-tuning dominates across all metrics. This contrasts with our encoder–decoder setting (Table 1) where a 4$\times$ teacher–student capacity gap allows both CHAD and EWAD+CPDP to substantially improve over the no-KD baseline. The failure of reliability-aware distillation here is attributed to the capacity-gap ceiling: when the student already approaches teacher-level performance, gating-based routing introduces variance without providing cleaner supervision.