Reliability-Gated Multi-Teacher Distillation for Low-Resource Abstractive Summarization

Anonymous ACL submission

Abstract

We study multi-teacher knowledge distillation for low-resource abstractive summarization from a *reliability-aware* perspective. We introduce **EWAD** (Entropy-Weighted Agreement-Aware Distillation), a token-level mechanism that routes supervision between teacher distillation and gold supervision based on inter-teacher agreement, and **CPDP** (Capacity-Proportional Divergence Preservation), a geometric constraint on the student's position relative to heterogeneous teachers. Across two Bangla datasets, 13 BanglaT5 ablations, and eight Qwen-2.5 experiments, we find that *logit-level KD provides the most reliable gains*, while more complex distillation improves semantic similarity for short summaries but degrades longer outputs. Cross-lingual pseudo-label KD across ten languages retains 71–122% of teacher ROUGE-L at 3.2× compression. A human-validated multi-judge LLM evaluation further reveals calibration bias in single-judge pipelines. Overall, our results show that reliability-aware distillation helps characterize when multi-teacher supervision improves summarization and when data scaling outweighs loss engineering.

1 Introduction

Large sequence-to-sequence models have greatly improved abstractive summarization (Lewis et al., 2020; Raffel et al., 2020; Xue et al., 2021), but their computational cost limits deployment in low-resource settings. Knowledge distillation (KD) (Hinton et al., 2015) addresses this by transferring knowledge from a large *teacher* to a smaller *student*. However, distillation for generative tasks is fragile: teacher predictions vary across tokens, and disagreement between teachers can introduce noisy supervision.

Most prior work relies on single-teacher KD or static aggregation of multiple teachers (Kim and Rush, 2016; You et al., 2017; Fukuda et al., 2017), assuming teacher predictions can be safely averaged. In practice, teachers often disagree on entities and phrasing, suggesting that supervision should be applied selectively.

We therefore study *reliability-aware multi-teacher distillation*. We introduce **EWAD** (Entropy-Weighted Agreement-Aware Distillation), a token-level objective that routes supervision between teacher KD and gold supervision based on teacher confidence and inter-teacher agreement, and **CPDP** (Capacity-Proportional Divergence Preservation), a geometric constraint regulating the student's divergence relative to heterogeneous teachers. Experiments on two Bangla summarization datasets show that *logit-level KD provides the most consistent gains*, while additional KD components improve semantic similarity for short summaries but degrade longer outputs. In Qwen-2.5 experiments, reliability-aware routing does not outperform direct fine-tuning, suggesting a capacity ceiling when student quality approaches the teacher. Cross-lingual pseudo-label KD transfers across ten languages with 71–122% teacher ROUGE-L retention. Rather than treating reliability-aware distillation purely as a performance optimization, we use it as a framework to study when multi-teacher supervision improves generation and when it introduces noise.

Our contributions are:

- We introduce **EWAD**, a reliability-gated distillation objective that dynamically switches between teacher supervision and gold supervision based on inter-teacher agreement.
- We introduce **CPDP**, a capacity-aware geometric regularizer constraining the student's divergence relative to heterogeneous teachers.
- Through extensive experiments, we show that logit-level KD dominates more complex KD objectives and that multi-component KD is strongly output-length dependent.

- We demonstrate cross-lingual pseudo-label distillation across ten languages and highlight calibration bias in single-judge LLM evaluation.

2 Related Work

KD for sequence generation. Hinton et al. (2015) introduced distillation via softened logits; Kim and Rush (2016) extended it to sequence-level KD. Multi-teacher KD (You et al., 2017; Fukuda et al., 2017) and teacher assistants (Mirzadeh et al., 2020) address capacity gaps, while Park et al. (2019) transfers inter-sample relationships. All prior methods use *fixed or globally learned* weights; to our knowledge, prior work does not explicitly gate distillation per token using inter-teacher agreement nor enforce capacity-proportional divergence.

Low-resource and multilingual summarization. XL-Sum (Hasan et al., 2021) covers 44 languages; BanglaT5 (Bhattacharjee et al., 2023) is the first Bangla seq2seq model. Cross-lingual transfer (Chi et al., 2021) and adapters (Pfeiffer et al., 2021) reduce adaptation cost but not inference cost. We show pseudo-label KD retains 71–122% ROUGE-L across ten languages at 3.2× compression.

Adaptive and token-level signals. Focal loss (Lin et al., 2017), TinyBERT (Jiao et al., 2020), and token-level weighting (Wen et al., 2023) move beyond uniform losses. EWAD uniquely decomposes gating into *confidence* and *agreement*; CPDP adds geometric constraints absent from prior adaptive KD.

3 Methodology

Our framework combines a *heterogeneous teacher ensemble* (same-vocabulary logit KD and cross-architecture pseudo-label supervision) with two reliability-aware training components: EWAD, which dynamically routes supervision between teacher KD and gold supervision based on inter-teacher agreement, and CPDP, which imposes a capacity-aware geometric constraint on the student's divergence relative to heterogeneous teachers.

3.1 Framework Overview

Given document $D = (x_1, \ldots, x_n)$, the goal is to generate summary $S = (y_1, \ldots, y_m)$ maximizing $P(S \mid D)$. Since transformers are bounded by context window L=1024, the pipeline begins with length-aware routing:

$$R(D) = \begin{cases} \text{DirectKD}(D) & \text{if } n \leq L \\ \text{MapReduce}(D) & \text{if } n > L \end{cases} \quad (1)$$

Short documents enter the KD branch; longer ones are routed to MapReduce (§3.5). Teachers are scored *offline* and cached; the student trains under a composite loss (CE + KL + optional MSE on hidden states). A five-stage ablation (A1–A5) and a separate eight-experiment EWAD+CPDP ablation isolate each component's contribution.

3.2 Multi-Teacher Knowledge Transfer

Same-Vocabulary Logit-Level KD. When teacher and student share the same vocabulary as with BanglaT5 (247.6M) → BanglaT5-small (109.9M), or within the Qwen-2.5 family (32B/14B → 3B) we perform *logit-level* distillation. The teacher is run in teacher-forced mode, producing p_T^t at each step; the student minimizes $\text{KL}(p_T^t \| p_S^t)$ softened by τ (§3.4). For Qwen-2.5, teachers use NF4 quantization and only top-k (k=50) log-probs are cached per position.

Cross-Architecture Pseudo-Label Transfer. When tokenizers are incompatible (mT5 250K vs. BanglaT5 32K), we use sequence-level distillation (Kim and Rush, 2016): each teacher generates pseudo-summaries (beam B=4), stored as text and re-tokenized at training time. Pseudo-labels replace gold summaries with probability p_{pseudo}=0.3 in A3–A5, using two mT5 teachers (∼580M each), creating a three-teacher heterogeneous ensemble.

3.3 Loss Functions

A fundamental tension in multi-teacher distillation for abstractive summarization is that teachers are not equally reliable at every generation step. Standard KD losses treat every token and every teacher identically, providing no mechanism for the student to distinguish confident consensus from noisy disagreement. We address this through a three-tier loss design: (i) a *standard distillation baseline* that establishes gold-anchored supervision (§3.3.1), (ii) EWAD, which introduces token-level reliability gating so the student selectively trusts teachers only when they are both confident and consistent (§3.3.2), and (iii) CPDP, which constrains the student's position in distribution space relative to each teacher, enforcing a geometric coherence absent from all prior KD objectives (§3.3.3).

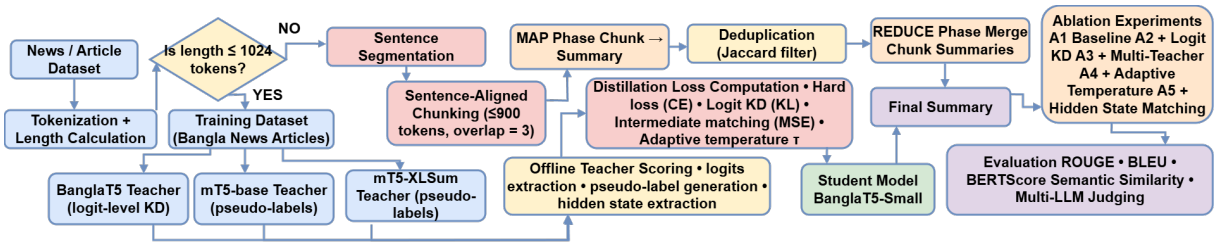

Figure 1: End-to-end framework. Documents are length-routed to the multi-teacher KD branch or MapReduce module. Three teachers provide logit and pseudo-label supervision across five ablation stages.

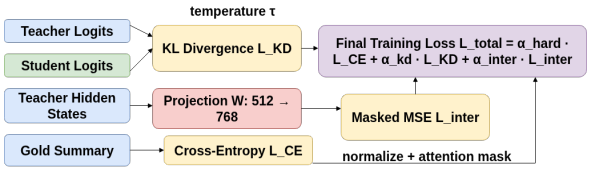

Figure 2: Standard distillation loss (Eq. 2): $\mathcal{L}_{\text{KD}}$ (softened KL), $\mathcal{L}_{\text{inter}}$ (projected MSE), and $\mathcal{L}_{\text{CE}}$ (gold cross-entropy).

3.3.1 Standard Distillation Loss

Before introducing our novel components, we establish a baseline objective that keeps gold supervision dominant while transferring teacher knowledge critical for maintaining summary quality when compressing models for deployment on resource-constrained devices. The base training objective for the BanglaT5 experiments (Figure 2) combines three loss terms:

$$\mathcal{L}_{\text{total}} = \alpha_{\text{hard}} \cdot \mathcal{L}_{\text{CE}} + \alpha_{\text{kd}} \cdot \mathcal{L}_{\text{KD}} + \alpha_{\text{inter}} \cdot \mathcal{L}_{\text{inter}} \quad (2)$$

where $\alpha_{\text{hard}} = 1 - \alpha_{\text{kd}} - \alpha_{\text{inter}}$ so the weights sum to one.

Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}$). Gold-label supervision anchors the student to human-written summaries, preventing the distribution drift that is particularly dangerous in low-resource settings where training data is scarce and every reference signal is valuable. Standard token-level negative log-likelihood on the gold summary:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T} \sum_{t=1}^{T} \log p_S^t(y_t^*) \quad (3)$$

where y_t^* is the gold token at position t and T is the summary length.

Logit-Level KD Loss ($\mathcal{L}_{\text{KD}}$). While hard labels encode only the argmax token, teacher logits carry a full distribution over the vocabulary capturing semantic proximity between candidate tokens (e.g., synonyms receiving similar probability mass) that is especially informative for abstractive generation, where multiple valid paraphrases exist. KL divergence between temperature-softened teacher and student distributions:

$$\mathcal{L}_{\text{KD}} = \tau^2 \cdot \text{KL}\left(\text{softmax}\left(\frac{\mathbf{z}_T}{\tau}\right) \,\middle\|\, \text{softmax}\left(\frac{\mathbf{z}_S}{\tau}\right)\right) \quad (4)$$

where $\mathbf{z}_T$ and $\mathbf{z}_S$ are teacher and student logits, respectively, and τ is the distillation temperature (fixed at 0.8 in A2–A3; adaptive in A4–A5). The τ^2 scaling maintains gradient magnitude parity with the cross-entropy term (Hinton et al., 2015).

Intermediate Matching Loss ($\mathcal{L}_{\text{inter}}$). Logit-level and label-level losses supervise only the output layer; yet for summarization, the encoder must learn to identify salient content, resolve coreference, and compress discourse structure capacities encoded in intermediate representations. By aligning encoder hidden states, we transfer these structural competencies directly, stabilizing learning especially under aggressive compression. We align encoder hidden states via a learned projection $\mathbf{W} \in \mathbb{R}^{512\times768}$ (student d=512 → teacher d=768):

$$\mathcal{L}_{\text{inter}} = \frac{1}{|\mathcal{M}|} \sum_{t\in\mathcal{M}} \left\| \overline{\mathbf{h}}_S^t - \overline{\mathbf{h}}_T^t \right\|^2, \quad \overline{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|_2} \quad (5)$$

where $\mathcal{M}$ is the set of non-padding positions and $\overline{\mathbf{h}}_S^t = \overline{\mathbf{W}\,\mathbf{h}_S^t}$. Active only in A5 ($\alpha_{\text{inter}}$=0.1).

3.3.2 EWAD: Entropy-Weighted Agreement-Aware Distillation

The standard loss above treats every teacher signal as equally trustworthy at every token position an assumption that breaks down in multi-teacher summarization. Consider a position where one teacher confidently predicts a factual entity while the other spreads probability across unrelated tokens: blindly averaging their logits injects noise that, in autore-

gressive generation, propagates through all subsequent tokens. EWAD addresses this with a two-axis reliability decomposition unique to our framework: *confidence* (which teacher to trust, based on entropy) and *agreement* (whether to trust teachers *at all*, based on distributional divergence). When both teachers are confident and agree, the student receives rich soft-label supervision; when they conflict, the loss automatically falls back to gold labels, preventing the student from learning from contradictory signals. This token-level gating enables reliability-aware distillation by dynamically routing supervision between teacher KD and gold supervision based on inter-teacher agreement. We describe the four-step computation below for teachers T_1, T_2 and student S (Figure 3).

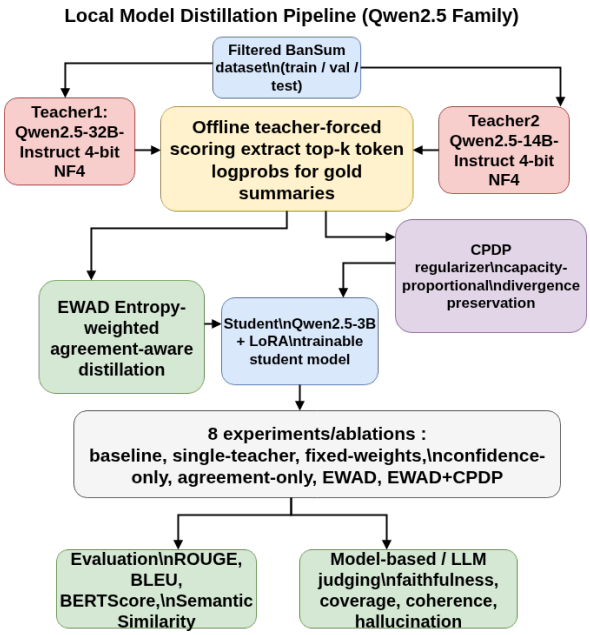

Figure 3: Dual-teacher EWAD+CPDP with Qwen-2.5 (32B + 14B → 3B + LoRA). Eight ablation experiments isolate each component.

Step 1: Teacher Confidence. The first axis measures how decisive each teacher is at each generation step. A teacher that concentrates probability on a few tokens carries a stronger, more informative signal than one with a flat distribution a distinction especially important in summarization, where content selection requires the model to commit to specific entities and phrasings. At each summary token position t, we compute each teacher's confidence as the complement of its normalized entropy:

$$C_i^t = 1 - \frac{H(p_i^t)}{\log |V|} \tag{6}$$

where $H(p_i^t) = -\sum_v p_i^t(v) \log p_i^t(v)$ and $|V|$ is the vocabulary size. $C_i^t \in [0, 1]$; values near 1 indicate the teacher concentrates probability mass on a few tokens.

Step 2: Confidence-Proportional Weights. Given per-token confidence scores, we must convert them into a weighting scheme that lets the more certain teacher dominate without hard-switching, which would discard useful information from the less confident teacher. Each teacher's influence is determined by a softmax over their confidence scores:

$$\begin{aligned} w_{T_1}^t &= \frac{\exp\big(C_{T_1}^t/\tau_w\big)}{\exp\big(C_{T_1}^t/\tau_w\big) + \exp\big(C_{T_2}^t/\tau_w\big)}, \\ w_{T_2}^t &= 1 - w_{T_1}^t \end{aligned} \tag{7}$$

with τ_w=1.0.

Step 3: Agreement Gate. Confidence alone is insufficient: two teachers can each be highly confident yet predict entirely different tokens a failure mode where naive confidence weighting would still inject harmful supervision. The second axis of EWAD measures inter-teacher *agreement* via Jensen–Shannon divergence (JSD), a symmetric and bounded divergence that naturally captures distributional overlap:

$$A_t = 1 - \frac{\mathrm{JSD}(p_{T_1}^t \| p_{T_2}^t)}{\log 2} \tag{8}$$

where $M = \frac{1}{2}(p_{T_1}^t + p_{T_2}^t)$ is the implicit midpoint distribution in the Jensen-Shannon divergence: $\mathrm{JSD}(p\|q) = \frac{1}{2}\mathrm{KL}(p\|M) + \frac{1}{2}\mathrm{KL}(q\|M)$. The score is passed through a sigmoid gate that creates a smooth but decisive transition between trust and distrust:

$$\lambda_t = \sigma\big(k \cdot (A_t - \delta)\big) \tag{9}$$

with k=5.0 and δ=0.5; $\lambda_t \approx 1$ when teachers agree, $\lambda_t \approx 0$ otherwise.

Step 4: EWAD Loss. The agreement gate enables the core novelty of EWAD: a *dynamic routing* between two supervision regimes. When teachers agree strongly ($\lambda_t \approx 1$), the student receives confidence-weighted soft-label KD that transfers rich distributional knowledge; when they conflict ($\lambda_t \approx 0$), the loss falls back to gold-label cross-entropy, treating the position as if no teacher supervision were available rather than polluting the

student with contradictory signals. The weighted KD loss and gold-label CE loss are blended via the agreement gate:

$$\mathcal{L}_{\text{EWAD}} = \frac{1}{T}\sum_{t=1}^{T}\left[\lambda_t\,\mathcal{L}_{\text{KD}}^t + (1-\lambda_t)\,\mathcal{L}_{\text{CE}}^t\right] \quad (10)$$

where the per-token distillation and gold-label terms are

$$\mathcal{L}_{\text{KD}}^t = w_{T_1}^t \cdot \text{KL}(p_{T_1}^t \| p_S^t) + w_{T_2}^t \cdot \text{KL}(p_{T_2}^t \| p_S^t), \quad (11)$$

$$\mathcal{L}_{\text{CE}}^t = -\log p_S^t(y_t^*). \quad (12)$$

3.3.3 CPDP: Capacity-Proportional Divergence Preservation

While EWAD determines when teacher supervision should be trusted, CPDP constrains where the student should lie in distribution space relative to teachers of different capacities. EWAD governs *when* to trust teachers; CPDP addresses a complementary question: *where should the student sit in distribution space relative to its teachers?* In a multi-teacher setup with heterogeneous capacities (e.g., 32B and 14B distilling into 3B), the student should not be equidistant from both teachers it should be closer to the smaller teacher whose representational capacity it more closely matches, and further from the larger teacher. Moreover, the gap between the student's divergences to each teacher should reflect the teachers' own mutual divergence. Without this constraint, the student can collapse toward one teacher while ignoring the other, or occupy an arbitrary region of distribution space that is geometrically inconsistent with the capacity hierarchy.

CPDP enforces this structure by penalizing deviations from a capacity-proportional divergence ratio. Let the two teachers have distributions p_{T_1} (larger) and p_{T_2} (smaller), and the student p_S. We define the target divergence gap as the inter-teacher divergence $\Delta^* = \text{KL}(p_{T_1} \| p_{T_2})$, and regularize the student's *normalized* divergence difference to match it:

$$\mathcal{L}_{\text{CPDP}} = \left(\frac{\text{KL}(p_{T_1}\|p_S)}{H(p_S)} - \frac{\text{KL}(p_{T_2}\|p_S)}{H(p_S)} - \Delta^*\right)^2 \quad (13)$$

The normalization by student entropy $H(p_S)$ converts raw KL values into capacity-relative units, ensuring the constraint is meaningful regardless of the student's absolute entropy level. Critically, $H(p_S)$ is **detached** during backpropagation to prevent the trivial solution of minimizing the loss by inflating student entropy. The inter-teacher divergence Δ^* is computed once and held fixed throughout training, serving as a geometric anchor. Weighted by μ=0.05; per-token values clamped at 100.0 for numerical stability. This regularizer is, to our knowledge, the first to impose a *capacity-aware geometric constraint* on the student's position in a multi-teacher KD framework moving beyond scalar loss weighting to enforce structural relationships in distribution space.

3.3.4 Combined Training Objective

EWAD and CPDP address orthogonal failure modes: EWAD prevents the student from learning from *unreliable* teacher signals (a temporal problem which tokens to trust), while CPDP prevents the student from occupying a *geometrically incoherent* position in distribution space (a structural problem where to sit relative to teachers). Their combination yields a training objective that is both reliability-gated and capacity-aware:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{EWAD}} + \mu \cdot \mathcal{L}_{\text{CPDP}} \quad (14)$$

BanglaT5 ablations use Eq. 2 with α_{kd}=0.01, $\alpha_{\text{inter}} \in \{0, 0.1\}$.

3.4 Confidence-Adaptive Temperature Scaling

In A4–A5, we adapt τ per sample so confident teachers provide sharper supervision while uncertain teachers are softened, allowing the model to self-adjust distillation strength based on input difficulty. The temperature is adapted *per sample* based on the teacher's mean token-level entropy:

$$\bar{H} = \frac{1}{|\mathcal{M}|}\sum_{t\in\mathcal{M}} H(p_T^t), \quad (15)$$

$$\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \sigma\big(\bar{H} - \bar{H}_{\text{batch}}\big) \quad (16)$$

where $\tau_{\min}$=0.5, $\tau_{\max}$=2.0, $\bar{H}_{\text{batch}}$ is the batch mean entropy, and σ is the sigmoid function. Low-entropy samples get sharper supervision ($\tau \to 0.5$); high-entropy samples get smoother distributions ($\tau \to 2.0$).

3.5 Long-Document Handling via MapReduce

Documents exceeding the 1 024-token context window (the "NO" branch in Figure1 1) are processed by a two-stage MapReduce pipeline rather than being truncated.

Sentence-Aligned Chunking. Sentences $D = (s_1, \ldots, s_k)$ are accumulated into chunks up to C=900 tokens:

$$c_i = (s_{a_i}, \ldots, s_{b_i}), \quad |c_i| \leq C \tag{17}$$

Consecutive chunks overlap by o=3 sentences:

$$c_{i+1} = (s_{b_i - o + 1}, \ldots, s_{b_{i+1}}) \tag{18}$$

MAP Phase. Each chunk is independently summarized by a BanglaT5 MAP model fine-tuned on full-document inputs: $\hat{s}_i = f_\theta(c_i)$.

Deduplication. Because overlapping chunks may produce redundant content, we remove duplicate sentences across chunk summaries using Jaccard similarity:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{19}$$

Sentence pairs with $J > 0.75$ are deduplicated.

REDUCE Phase. Deduplicated summaries are concatenated and passed to a REDUCE model (transfer-learned from MAP): $S = f_\phi(\text{concat}(\hat{S}_{\text{map}}))$, applied recursively.

3.6 Cross-Lingual Extension

We apply offline pseudo-label KD to ten languages (Hindi, Urdu, Persian, Amharic, Hausa, Nepali, Pashto, Indonesian, Turkish, Swahili): a fine-tuned mT5-XLSum teacher (966M) generates pseudo-summaries, and mT5-small (300M, 3.2× compression) trains on them. Shared tokenizers eliminate vocabulary mismatch; offline caching makes training 100× faster than online KD. Notably, the entire framework including EWAD, CPDP, and the MapReduce pipeline is *language-agnostic*: it relies on no language-specific heuristics beyond tokenization, and the cross-lingual results confirm that it transfers without modification across typologically diverse languages and scripts.

4 Experimental Setup : Datasets, Models, Metrics, and Compute

We use three datasets: BTS (68K Bangla news article–headline pairs, ∼8-token targets) (Moni, 2019), BanSum (141K document–summary pairs, ∼106–248-token targets; a 20K filtered subset used for EWAD+CPDP) (Hasan et al., 2024), and XL-Sum subsets covering ten languages (512 source tokens) (Hasan et al., 2021). Table 1 lists all models. For BanglaT5 experiments, the teacher shares the student vocabulary while two mT5 models generate pseudo-labels. For EWAD+CPDP, Qwen-2.5 teachers (NF4) distill into Qwen-2.5-3B using LoRA ($r = 64$, $\alpha = 128$). We report ROUGE-1/2/L (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore F1 (Zhang et al., 2020), semantic similarity, and LLM-judge scores. All experiments were conducted on consumer GPUs (RTX 5090/5080/4070 Ti Super), totaling approximately 320 GPU-hours.

Role	Model	Params
BanglaT5 Ablation (A1–A5)		
Teacher (logit)	BanglaT5	247.6M
Teacher (pseudo)	mT5-base / mT5-XLSum	∼580M
Student	BanglaT5-small	109.9M
EWAD+CPDP *(Qwen-2.5)*		
Teachers	Qwen2.5-32B / 14B-Inst.	32B / 14B
Student	Qwen2.5-3B + LoRA	3B
Cross-Lingual		
Teacher / Student	mT5-XLSum / mT5-small	966M / 300M

Table 1: Teacher and student model configurations.

5 Results

5.1 BanglaT5 Ablation: Bengali Text Summarization

Table 2 presents the five-stage ablation on the BTS dataset (8,033 test samples). A2 (single-teacher logit KD) achieves the highest ROUGE scores, retaining 93.6% of teacher ROUGE-L. Adding further components (A3–A5) slightly reduces n-gram overlap; however, semantic similarity increases monotonically from A1 to A5 (0.868 → 0.870), indicating that each component injects meaningful representational knowledge even when surface overlap saturates. A5 achieves the highest semantic similarity (0.8695), showing that encoder matching transfers structure beyond ROUGE.

Config	R-1	R-2	R-L	BLEU	BS F1	Sem
A1 Baseline	.3943	.2321	.3794	15.79	.7863	.8678
A2 +LogitKD	**.3945**	**.2321**	**.3797**	15.49	.7867	.8683
A3 +Pseudo	.3924	.2301	.3777	15.20	.7855	.8689
A4 +Adapt. τ	.3914	.2296	.3767	15.02	.7846	.8693
A5 +InterMatch	.3914	.2299	.3769	15.08	.7848	**.8695**
Teacher	*.4234*	*.2512*	*.4058*	*16.47*	*.7917*	*.8722*

Table 2: BTS ablation results (BanglaT5-small student, 109.9M). R = ROUGE, BS = BERTScore, Sem = Semantic Similarity. Best student scores in **bold**.

5.2 BanglaT5 Ablation: BanSum

Table 3 shows results on the BanSum dataset (14,120 test samples). A2 dominates across *all* metrics ROUGE, BLEU, BERTScore, and semantic similarity achieving 95.9% of teacher ROUGE-L. Unlike BTS, additional components (A3–A5) consistently degrade performance. The contrasting behavior is explained by BanSum's substantially longer outputs ($\sim$106–248 tokens vs. $\sim$8 tokens in BTS): the extra regularization from pseudo-labels and adaptive temperature introduces harmful noise on longer, more diverse summaries.

Config	R-1	R-2	R-L	BLEU	BS F1	Sem
A1 Baseline	.2957	.1389	.2314	11.67	.7441	.7538
A2 +LogitKD	**.3527**	**.2034**	**.2877**	**12.05**	**.7486**	**.7603**
A3 +Pseudo	.3014	.1444	.2405	11.84	.7465	.7584
A4 +Adapt. τ	.2866	.1352	.2253	10.88	.7393	.7443
A5 +InterMatch	.2881	.1361	.2268	10.87	.7391	.7466
Teacher	*.3663*	*.2135*	*.2998*	*.1254*	*.7506*	*.7689*

Table 3: BanSum ablation results (BanglaT5-small student, 109.9M). R = ROUGE, BS = BERTScore, Sem = Semantic Similarity. Best student scores in **bold**.

5.3 Dual-Teacher EWAD+CPDP Ablation (Qwen-2.5)

Table 4 reports the eight-experiment ablation on the quality-filtered BanSum subset (20K; 2,000 test samples) using the Qwen-2.5 family.The baseline (direct fine-tuning of Qwen2.5-3B without any distillation) achieves the highest ROUGE and BERTScore. Single-teacher configurations (32B, 14B) and fixed-weight dual-teacher perform comparably to the baseline but do not surpass it. Confidence-only weighting performs worst (0.099 ROUGE-L), and even full EWAD/EWAD+CPDP remain below the baseline.

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 4: EWAD+CPDP ablation on filtered BanSum (20K subset), Qwen2.5-3B + LoRA student. B-4 = BLEU-4. R = ROUGE, BS = BERTScore, Sem = Semantic Similarity. Best student scores in **bold**.

5.4 Full-Scale Validation

Based on the selective evaluation above, the baseline configuration (direct fine-tuning) was identified as the strongest and retrained on the full 141K BanSum corpus for 5 epochs. Table 5 compares the 20K and 141K results. Scaling from 20K to 141K yields consistent improvements across all metrics: +0.015 ROUGE-1, +0.015 ROUGE-2, +0.017 ROUGE-L, and +0.019 semantic similarity. The full-scale model also exceeds the BanglaT5-small A1 baseline (Table 3) on ROUGE-L (0.2335 vs. 0.2314), demonstrating that the larger Qwen2.5-3B architecture benefits substantially from additional training data.

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	*+.0154*	*+.0148*	*+.0175*	*+.0038*	*+.0048*	*+.0189*

Table 5: Qwen2.5-3B baseline fine-tuning: 20K subset vs. full 141K BanSum. 14,120 test samples for 141K; 2,000 for 20K.R = ROUGE, BS = BERTScore, Sem = Semantic Similarity.

5.5 Cross-Lingual Distillation

Table 6 presents cross-lingual pseudo-label distillation results across ten languages. The student (mT5-small, 300M) achieves 71–93% of teacher ROUGE-L on nine languages and *surpasses* the teacher on Pashto (122.4% retention), at 3.2$\times$ compression. The consistency across typologically diverse languages spanning Devanagari (Hindi, Nepali), Nastaliq (Urdu), Perso-Arabic (Persian, Pashto), Ge'ez (Amharic), and Latin (Hausa, Indonesian, Turkish, Swahili) scripts confirms that the offline pseudo-label KD pipeline generalizes beyond bangla.

5.6 Multi-Judge LLM Evaluation

We evaluate 1,000 BanSum A2 student summaries with two LLM judges (GPT-5.2, Claude Sonnet 4.6) scoring faithfulness, coverage, coherence, and conciseness (1–10) plus hallucination detection. Table 7 reports the results: high faithfulness (8.73) and coherence (7.61), with coverage (6.36) weakest. The hallucination divergence (0% GPT-5.2 vs. 25.5% Claude) underscores the need for multi-judge evaluation.

Human Validation. We sampled 100 concordant and 100 discordant judge pairs for blind evaluation

Lang	Teacher (966M) R-1	R-2	R-L	Student (300M) R-1	R-2	R-L	Ret. (%)
Hindi	.419	.218	.372	.344	.165	.308	83.0
Urdu	.418	.213	.372	.370	.195	.330	88.8
Persian	.352	.167	.324	.292	.130	.254	78.2
Amharic	.286	.148	.262	.253	.140	.245	93.5
Hausa	.462	.269	.408	.426	.258	.379	92.9
Nepali	.367	.193	.343	.308	.163	.303	88.4
Pashto	.450	.229	.401	.525	.327	.491	122.4
Indonesian	.367	.180	.328	.263	.126	.238	72.5
Turkish	.291	.145	.270	.200	.104	.192	71.1
Swahili	.392	.211	.343	.289	.139	.253	73.8
Average	*.380*	*.197*	*.342*	*.327*	*.175*	*.299*	*86.5*

Table 6: Cross-lingual pseudo-label KD (mT5-XLSum → mT5-small, 3.2× compression). Ret. = ROUGE-L retention.R = ROUGE, BS = BERTScore, Sem = Semantic Similarity.

Dimension	GPT-5.2	Claude	Avg.
Faithfulness	9.64	7.82	8.73
Coverage	6.46	6.26	6.36
Coherence	8.00	7.21	7.61
Conciseness	8.00	7.08	7.54
Overall	8.05	7.01	**7.53**
Halluc. rate	0.0%	25.5%	12.8%

Table 7: LLM judge evaluation of 1,000 A2 student summaries (BanSum).

by five annotators. On concordant samples, annotators confirmed both judges in 93% of cases. On discordant samples, annotators sided with Claude in 84% of cases (vs. 38% for GPT-5.2), revealing GPT-5.2's systematic positive bias it assigns high faithfulness even to samples with minor factual inconsistencies. Table 8 summarizes the results.

Metric	GPT-5.2	Claude
Human agree. (concordant)	91%	95%
Human agree. (discordant)	38%	84%
Halluc. label match	41%	82%
Mean score $\lvert\Delta\rvert$	1.73	0.58

Table 8: Human validation of LLM judges on 200 samples (100 concordant + 100 discordant), evaluated by five annotators. "$\lvert\Delta\rvert$" = mean absolute score difference from human consensus.

6 Analysis and Discussion

We analyze the results not only to measure improvements but also to identify regimes where multi-teacher distillation becomes unstable.

Logit-level KD dominates. Across both datasets, the largest improvement occurs from A1 to A2, showing that most gains come from transferring softened token distributions.

Output length modulates KD effectiveness. Additional KD components improve semantic similarity on short-output BTS but consistently degrade performance on long-output BanSum, suggesting that pseudo-label noise compounds across longer autoregressive sequences.

EWAD and CPDP reveal practical limits of multi-teacher KD. In the Qwen-2.5 experiments, neither EWAD nor CPDP surpasses direct fine-tuning of the 3B student. This suggests the presence of a *teacher-quality ceiling*: when the student already approaches teacher performance, reliability-aware multi-teacher routing may introduce additional variance without providing cleaner supervision.

Cross-lingual distillation generalizes broadly. Offline pseudo-label KD retains strong performance across ten languages and five writing systems, achieving 86.5% average teacher ROUGE-L retention and even surpassing the teacher in Pashto. This suggests that the distillation pipeline itself is robust across typologically diverse settings.

Data scaling outweighs loss engineering. Increasing training data from 20K to 141K examples improves ROUGE-L more than any EWAD or CPDP modification. This indicates that, at current model scales, improving data coverage may be more effective than designing increasingly complex KD objectives.

7 Conclusion

We studied multi-teacher knowledge distillation for low-resource abstractive summarization through a reliability-aware perspective. Our experiments show that while reliability-gated distillation is conceptually promising, the most consistent gains still come from simple logit-level KD, and the effectiveness of more complex objectives depends strongly on output length and teacher–student capacity gaps.

These results suggest that future progress in generative KD may depend less on increasingly complex loss functions and more on understanding when teacher supervision is genuinely reliable. Cross-lingual experiments and human-validated evaluation further highlight the importance of robust data pipelines and trustworthy evaluation practices for summarization research.

8 Limitations

While our study provides a comprehensive analysis of reliability-aware multi-teacher distillation, several limitations remain.

First, our human validation study is limited in scale, covering 200 samples and five annotators on the BanSum dataset. Although this provides useful insight into LLM evaluation reliability, broader studies across datasets and annotator pools would strengthen the conclusions.

Second, although we evaluate cross-lingual generalization across ten languages and multiple scripts, the coverage does not include certain linguistic typologies such as highly agglutinative or tonal languages. The generalization of the proposed framework to such languages remains an open question.

Third, our findings indicate that reliability-aware distillation methods (EWAD and CPDP) do not consistently outperform direct fine-tuning when the student model approaches teacher capacity. However, we do not systematically explore wider teacher–student capacity gaps, where such methods may be more beneficial.

Fourth, we observe that output length significantly modulates the effectiveness of multi-component distillation, with improvements on short summaries and degradation on longer outputs. While we hypothesize that noise accumulation drives this behavior, we do not explicitly control for output length (e.g., via length-normalized objectives or decoding constraints).

Finally, our experiments rely on a fixed set of architectures (BanglaT5 and Qwen-2.5 families) and training setups. While the consistency of trends across these settings is encouraging, further validation across additional model families and training regimes would improve the robustness of our conclusions.

9 Ethics Statement

Data and Privacy. All datasets used in this work (BTS, BanSum, and XL-Sum subsets) are publicly available and widely used in prior research. We did not collect any new data involving human subjects. To the best of our knowledge, these datasets do not contain personally identifiable information (PII). No additional steps for anonymization were required.

Human Evaluation. We conducted a limited human validation study involving five annotators to assess agreement with LLM-based evaluations. Annotators were co-authors of this work and participated voluntarily without financial compensation. No sensitive or personal data was collected during this process. The evaluation task involved assessing generated summaries for quality dimensions such as faithfulness and coherence.

Ethical Considerations. As this study does not involve external participants or sensitive personal data, formal ethics board approval was not required. The work complies with standard ethical guidelines for NLP research using publicly available datasets.

Bias and Limitations. The datasets used are primarily news-based and may reflect inherent societal or reporting biases present in the source material. While our study focuses on model behavior rather than content generation in deployment settings, such biases may still influence model outputs. Additionally, LLM-based evaluation was found to exhibit calibration differences across models, which we explicitly analyze.

Use of AI Assistants. AI-assisted tools were used for minor writing support and language editing. All technical content, experimental design, and results were developed, verified, and validated by the authors.

Environmental Impact. All experiments were conducted on consumer-grade GPUs (e.g., RTX 5090/5080/4070 Ti Super) with an estimated total compute usage of approximately 320 GPU-hours. While model training incurs energy consumption, the use of knowledge distillation aims to reduce deployment costs and improve efficiency in downstream applications.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin Mubasshir, Md. Saiful Islam, Anindya Iqbal, M. Sohel Rahman, 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 1948–1968. Association for Computational Linguistics.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Zheng, Xian-Ling Huang, Xin-Ying Huang, and Furu Wei Huang. 2021. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683. 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 *Proceedings of Interspeech*, 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.

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. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

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. 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. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. 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*, pages 2980–2988.

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.

Hasan Moni. 2019. Bengali text summarization. https://www.kaggle.com/datasets/hasanmoni/bengali-text-summarization.

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. 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*, pages 3967–3976.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–503. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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*, pages 10817–10834. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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*, pages 1285–1294. ACM.

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