TOKEN-ADAPTIVE CONCRETE SCORE DISTILLATION (TA-CSD)

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation for autoregressive language models commonly applies a fixed temperature transform to logits before computing a divergence. We begin with an analytical observation: for any objective that depends on logits solely through pairwise differences, including softmax divergences over a fixed candidate window and concrete-score objectives, the centering component of logit standardization is mathematically inert. Consequently, standardization operates entirely as a per-token temperature τ_t^S, collapsing the question of whether to standardize into which estimator of that per-token scale to use. The conventional plug-in choice, the student's head standard deviation $\hat{\sigma}_t^S$, is blind to teacher confidence, and because concrete-score gradients scale inversely with temperature, plug-in standardization silently acts as an effective learning rate modifier. We empirically quantify this effect in language model distillation, offering a candidate explanation for conflicting reports on logit standardization in the literature. **Token-Adaptive Concrete Score Distillation** (**TA-CSD**) predicts the per-token temperature τ_t^S via a lightweight multi-layer perceptron (MLP) gate conditioned on detached supervisory and predictive uncertainty features. In practice, the gate is trained jointly with the student under score distillation with detached features, while bilevel soft-ECE optimization provides a theoretical and diagnostic foundation. Under its learned scale, TA-CSD remains effective-learning-rate neutral (measured ratio $\mathbf{0.989 \pm 0.050}$, with mean $\tau = 1.14$), so its differences from unscaled distillation cannot be attributed to step size. Evaluating across multiple model families and seeds with paired example-level bootstrap tests reveals a dissociation between calibration error and task capability. Plug-in standardization lowers top-1 calibration error only nominally (0.0565 vs. 0.0666 for CSD, not significant) while hedging (flattening the predictive distribution), losing 33% of decisive (> 0.9) token predictions, degrading mathematical reasoning on **GSM8K** by 12.3 percentage points ($36.69\% \rightarrow 24.37\%$), and significantly worsening proper score ($\Delta = +0.0075, p < 0.001$). While on-policy rollouts (GKD) achieve strong calibration, they alter the training distribution and demand over $12\times$ more training compute. Among offline cached distillation methods, **TA-CSD** avoids this harm: our clamped formulation attains the best point-estimate Brier score (0.1628 ± 0.0030) and protects GSM8K reasoning (36.36%, vs. 24.37% for SWS), while a residual gate reaches smECE 0.0522 at the cost of reasoning (29.39%).

1 INTRODUCTION

Distillation objectives for autoregressive language models differ extensively in which divergence they minimize and on whose distribution samples are drawn (Gu et al., 2024; Ko et al., 2024; Agarwal et al., 2024). Yet almost all of them apply the same pre-divergence transform to the logits: the uniform temperature scaling introduced by Hinton et al. (2015). Logit standardization (Sun et al., 2024) is the primary alternative, pre-processing both models' logits with an empirical Z-score. However, empirical studies in language model distillation have yielded conflicting conclusions: some find standardization beneficial for multi-teacher fusion (Yan et al., 2025), while others observe that standard unnormalized logit matching outperforms complex normalization recipes (Xu et al., 2024). We argue that this disagreement is fundamentally underdetermined, and that resolving it reveals a critical design axis in model distillation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our starting point is an elementary observation grounded in shift invariance. A softmax distribution is invariant to any uniform shift of its input logits, and concrete-score objectives are constructed entirely from pairwise logit differences (Meng et al., 2022; Kim et al., 2026). It follows (§3.1) that the *centering* half of standardization contributes nothing to any shift-invariant objective: standardization acts *solely* through its scale, reducing functionally to a per-token, per-model temperature $\tau_t^m = \hat{\sigma}_t^m \tau$. In the backward pass, this equivalence to scalar per-token temperature scaling holds strictly when treating $\hat{\sigma}_t^S$ as a detached constant ($\mathrm{sg}[\hat{\sigma}_t^S]$); differentiating end-to-end through empirical sample standard deviation introduces coordinate-coupling projection terms and fp16 numerical instability (§3.1). The question is therefore not whether to standardize, but *which estimator* of that per-token scale to employ; the default, the sample standard deviation $\hat{\sigma}_t^S$ of the very logits being trained, is a self-referential plug-in blind to teacher confidence.

A second, more damaging consequence of this reduction has gone unexamined. Because the transform reduces to a scalar multiplier $1/\tau_t^S$ on score residuals (§3.4), applying a plug-in scale silently alters the student's *effective learning rate*. In classical softmax distillation, Hinton et al. (2015) observed that temperature scales gradients as $1/T^2$ and corrected for it explicitly; under the Huber score penalty the factor is $1/\tau_t^S$, but in both cases an uncorrected pre-divergence scale changes the effective step size (van Laarhoven, 2017; Li & Arora, 2020). We quantify this directly on fixed evaluation batches (§5.1): plug-in standardization (Sun et al., 2024) on the concrete score backbone attenuates the student logit gradient by a factor of $3.2\times$ (ratio 0.310). We hypothesize that such unmeasured attenuation contributes to conflicting reports in prior distillation literature (Yan et al., 2025; Xu et al., 2024): a method that silently acts as a $3.2\times$ learning-rate reduction can appear beneficial or harmful depending on whether the baseline optimizer was tuned aggressively or conservatively. In our own setting, however, matched-LR retraining does not remove the deficit of plug-in standardization (§5.3), so attenuation alone does not explain it.

We therefore treat the pre-divergence scale as a principled *decision*. **Token-Adaptive Concrete Score Distillation** (TA-CSD) predicts the per-token student scale τ_t^S via a compact MLP gate conditioned on token uncertainty signals validated in the selective-distillation literature (Wang et al., 2025; Xu et al., 2026; Tavor et al., 2026; Huang et al., 2025). The gate is trained jointly with student weights under detached features, while bilevel optimization with a soft-binned calibration surrogate (Karandikar et al., 2021) provides an analytical framework for hypergradient dynamics. Under its residual parameterization, the gate strictly generalizes plug-in standardization as a special case (Remark 1), ensuring that any performance gain is attributable to *which* scale is chosen, while the clamped formulation operates as a bounded scale predictor. The learned gate is empirically effective-learning-rate neutral (measured ratio 0.989 ± 0.050), so its differences from CSD cannot be attributed to step size.

Our principal contributions are:

- **Analytical reduction (§3.1):** We prove that logit centering is inert under concrete-score distillation and candidate-window softmax divergences, reducing standardization entirely to per-token temperature scaling τ_t^S when detached, with plug-in standard deviation being merely one convenience heuristic.
- **Quantifying effective learning rate (§3.4, §5.1):** We prove the transform acts as a per-token residual gain and empirically show plug-in standardization attenuates student gradient norm by $3.2\times$ (ratio 0.310). This is a candidate explanation for conflicting prior reports, although in our setting matched-LR retraining does not rescue plug-in standardization; our learned gate remains effective-LR neutral (0.989 ± 0.050).
- **Token-adaptive gating on the calibration-reasoning frontier (§3.3–§3.4, §5):** We introduce TA-CSD under clamp and residual parameterizations. Clamped gating attains the best point-estimate Brier score among offline distillation methods (0.1628 ± 0.0030; not significant vs. CSD, comparable to post-hoc temperature scaling) and preserves GSM8K reasoning (36.36%), while residual gating reaches smECE 0.0522 at a reasoning cost (29.39%). We benchmark against the undistilled base student, offline baselines, and on-policy GKD.

2 Related Work

Extended literature is provided in Appendix F; we summarize key connections here.

Score space and pre-divergence transforms. Distillation for autoregressive LMs has focused heavily on divergence formulation (Gu et al., 2024; Ko et al., 2024; Wu et al., 2025; Zhang et al., 2024) and sampling distribution (Agarwal et al., 2024). Concrete score distillation (CSD, Kim et al., 2026) matches logit differences directly, dispensing with the global softmax partition function. Yet across paradigms, the pre-divergence logit transform is inherited unchanged from Hinton et al. (2015). Logit standardization (SWS, Sun et al., 2024) applies a Z-score, but its utility remains contested (Yan et al., 2025; Xu et al., 2024). We show this dispute is driven by unmeasured effective learning rate shifts (§5.1). We also benchmark against on-policy distillation (GKD, Agarwal et al., 2024) and reverse-KL (MINILLM, Gu et al., 2024).

Uncertainty conditioning and selective distillation. Tokens contribute unequally to distillation (Zhong et al., 2024; Huang et al., 2025; Kim & Baek, 2026), with informative positions concentrating where student entropy is high or confidence diverges from correctness (Wang et al., 2025; Xu et al., 2026; Tavor et al., 2026). While prior methods exploit uncertainty for token masking, loss reweighting (Huang et al., 2025), or static heuristic temperatures (Xie et al., 2026), they do not learn an adaptive pre-divergence scale.

Calibration, proper scoring rules, and optimization. Neural networks are notoriously miscalibrated (Guo et al., 2017). Although post-hoc scaling (Guo et al., 2017) mitigates overconfidence, standard ECE (Naeini et al., 2015) can be flattered by distribution flattening (hedging). Strictly proper scoring rules like the Brier score (Brier, 1950) decompose uniquely into calibration and sharpness (Murphy, 1973; DeGroot & Fienberg, 1983), penalizing trivial hedging. Debiased kernel estimators like smECE (Błasiok & Nakkiran, 2024) and Full-ECE (Liu et al., 2024) prevent binning artifacts. We learn an uncertainty-conditioned per-token scale with a compact gate ($|\phi| = 177$) conditioned on supervisory signals (u_t), anchored by reference cross-entropy to prevent predictive hedging.

3 METHODOLOGY

We train student p_θ^S from frozen teacher p^T by matching teacher-forced next-token predictions on an instruction corpus. The goal is to select τ_t^S to produce a well-calibrated, high-accuracy student without incurring inference latency or confounding step-size shifts. Calibration is evaluated per token: confidence $\max_v p_t^S(v)$ should match empirical correctness, measured by debiased smECE (Błasiok & Nakkiran, 2024), Full-ECE (Liu et al., 2024), and Brier score on held-apart data (§4).

3.1 PRELIMINARIES AND ANALYTICAL REDUCTION

Let teacher p^T and student p_θ^S define next-token distributions over vocabularies $\mathcal{V}^T$ and $\mathcal{V}^S$. Given a prompt-response sequence (x, y) and token index t, let $\ell_t^m \in \mathbb{R}^{|\mathcal{V}^m|}$ denote the logit vectors, with $p_t^m = \mathrm{softmax}(\ell_t^m)$ for $m \in \{T, S\}$. Standard token-level distillation minimizes

$$\mathcal{L}_{\mathrm{KD}} = \sum_t D\big(\mathrm{softmax}(\ell_t^T/\tau) \,\big\|\, \mathrm{softmax}(\ell_t^S/\tau)\big)\,, \tag{1}$$

optionally combined with cross-entropy on reference tokens. Logit standardization (Sun et al., 2024) pre-processes logits via a Z-score:

$$\tilde{\ell}_{t,v}^m = \frac{\ell_{t,v}^m - \mu_t^m}{\hat{\sigma}_t^m \tau}, \qquad v \in \mathcal{S}_t, \tag{2}$$

where μ_t^m and $\hat{\sigma}_t^m$ are the empirical mean and standard deviation over a normalization scope $\mathcal{S}_t = \text{Top-}k(\ell_t^T)$, termed *static window standardization* (SWS).

Proposition 1 (Centering is inert)**.** *Let $\mathcal{L}$ be any objective depending on logits at position t only through pairwise differences $\{\ell_{t,v} - \ell_{t,v'}\}_{v,v' \in \mathcal{S}_t}$ (e.g., softmax divergences on $\mathcal{S}_t$ or concrete score matching). For any scalar $c_t \in \mathbb{R}$, replacing $\ell_{t,v}$ with $\ell_{t,v} - c_t$ leaves $\mathcal{L}$ and its parameter gradient $\nabla_\theta \mathcal{L}$ identically unchanged. Consequently, the Z-score transform $\tilde{\ell}_t^m$ reduces to per-token temperature scaling:*

$$\tau \;\longmapsto\; \tau_t^m = \hat{\sigma}_t^m \tau. \tag{3}$$

Under gradient detachment of the scale ($\mathrm{sg}[\hat{\sigma}_t^m]$), this reduction holds identically in both forward and backward passes.

Corollary 1. *Standardization reduces entirely to selecting a per-token scale τ_t^S.* SWS *(Sun et al., 2024) is merely the plug-in choice $\tau_t^S = \hat{\sigma}_t^S \tau$.*

We note three precise boundaries of Proposition 1 (detailed in Appendix B): (i) inertness holds strictly for the score matching term $\mathcal{L}_{\text{score}}$, whereas student cross-entropy $\mathcal{L}_{\text{CE}}$ over the full vocabulary $\mathcal{V}^S$ is not shift-invariant within a local top-k window; (ii) the normalization scope $\mathcal{S}_t$ is restricted to the candidate neighborhood $\mathcal{N}_t \cup \{a_t\}$; and (iii) in finite-precision arithmetic (fp16), logit centering can introduce minor rounding discrepancies, which we verify empirically to be well within seed variation.

Differentiating the Z-score in Eq. equation 2 with respect to unnormalized logits yields the normalization Jacobian. Letting $z_{t,v} = (\ell_{t,v} - \mu_t)/\hat{\sigma}_t$ denote the unit-variance centered logits (such that $\tilde{\ell}_{t,v} = z_{t,v}/\tau$):

$$\frac{\partial \tilde{\ell}}{\partial \ell} = \frac{1}{\hat{\sigma}\tau}\left(I - \tfrac{1}{k}\mathbf{1}\mathbf{1}^\top - \tfrac{1}{k} z z^\top\right) = \frac{1}{\hat{\sigma}\tau}\left(I - \tfrac{1}{k}\mathbf{1}\mathbf{1}^\top - \tfrac{\tau^2}{k}\tilde{\ell}\tilde{\ell}^\top\right). \tag{4}$$

The rank-one outer product term $-\frac{1}{k} z z^\top$ projects away gradient energy along the normalized logit vector. While the individual entry scales as $1/k$, the cumulative off-diagonal energy ratio grows with window size k (Appendix E, Fig. 3). Crucially, detaching $\hat{\sigma}_t^S$ removes this coordinate coupling and avoids catastrophic fp16 gradient overflow (§5.4), allowing the transform to operate strictly as an adaptive scale.

3.2 Concrete Score Matching Backbone

Concrete score matching (Meng et al., 2022; Kim et al., 2026) compares logit differences directly, dispensing with global partition functions. Let $a_t = \arg\max_v p_t^T(v)$ be an anchor token and $\mathcal{N}_t = \text{Top-}k(\ell_t^T) \setminus \{a_t\}$ be the teacher-defined candidate neighborhood. We define anchored scores as

$$s_t^m(v) = \ell_{t,v}^m - \ell_{t,a_t}^m, \qquad v \in \mathcal{N}_t. \tag{5}$$

The TA-CSD objective matches anchored scores under a Huber penalty ρ:

$$\mathcal{L}_{\text{score}}(\theta, \phi) = \sum_t \sum_{v \in \mathcal{N}_t} w_t(v)\, \rho\left(\frac{s_t^T(v)}{\tau_t^T} - \frac{s_t^S(v)}{\tau_t^S(\phi)}\right), \tag{6}$$

where τ_t^T is a frozen teacher scale (set to 1), $\tau_t^S(\phi)$ is the learned student scale, and $w_t(v) = \text{sg}[\sqrt{p_t^T(v) p_t^S(v)}]$ specifies geometric token weighting treated with a stop-gradient operator following standard score matching practices (Kim et al., 2026). Focusing on same-tokenizer pairs isolates the transform from vocabulary projection confounds (Boizard et al., 2025).

3.3 Token-Adaptive Scale Parameterization

Rather than relying on a static plug-in heuristic, TA-CSD predicts τ_t^S via a lightweight MLP gate g_ϕ ($9 \to 16 \to 1$, $|\phi| = 177$) conditioned on detached token uncertainty features:

$$u_t = \text{sg}\big[\hat{H}_t^T,\, m_t^T,\, \hat{H}_t^S,\, m_t^S,\, d_t,\, \log \hat{\sigma}_t^S,\, \tfrac{r_t}{k},\, \tfrac{h}{H},\, \tfrac{n}{N}\big] \in \mathbb{R}^9, \tag{7}$$

where $\hat{H}_t^m$ is head entropy, $m_t^m = p_{(1)}^m - p_{(2)}^m$ is the top-1 probability margin, d_t is the symmetric Hellinger distance quantifying teacher-student divergence (Xie et al., 2026), $\hat{\sigma}_t^S$ is head standard deviation, r_t is reference token rank in the candidate window (clamped to k if outside), h/H is relative transformer depth, and n/N is global training progress. To preserve $\mathcal{O}(k)$ efficiency without evaluating full-vocabulary softmax partitions, candidate distributions ($\hat{H}_t^m, m_t^m, d_t$) are computed over the locally renormalized candidate window $\mathcal{S}_t = \text{Top-}k(\ell_t^T)$. We investigate two scale parameterizations:

$$\text{Clamped:} \quad \tau_t^S = \tau_{\min}\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{\varsigma(g_\phi(u_t))}, \tag{8}$$

$$\text{Residual:} \quad \tau_t^S = \hat{\sigma}_t^S \cdot \exp(g_\phi(u_t)), \tag{9}$$

where ς denotes the standard sigmoid function, ensuring $\tau_t^S \in [\tau_{\min}, \tau_{\max}]$ for the clamped variant.

Remark 1 (Subsumption and scale parameterization)**.** *Because* $\log \hat{\sigma}_t^S \in u_t$*, the residual formulation* $\tau_t^S = \hat{\sigma}_t^S \cdot \exp(g_\phi(u_t))$ *strictly subsumes plug-in standardization as a special case when* $g_\phi \equiv 0$*. In contrast, the clamped variant* $\tau_t^S \in [\tau_{\min}, \tau_{\max}]$ *acts as a bounded predictor independent of* $\hat{\sigma}_t^S$*. Any measured gain of* TA-CSD *is therefore attributable to* which *scale is chosen, rather than the mere presence of an adaptive scale.*

3.4 Joint Optimization Objective and Gradient Gain

The student weights θ and gate parameters ϕ are trained jointly via AdamW on the score distillation and cross-entropy objective:

$$\min_{\theta,\phi} \; \mathcal{L}_{\text{score}}(\theta, \phi) + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}}(\theta), \tag{10}$$

where gate input features u_t are strictly detached ($\text{sg}[u_t]$). Detaching u_t eliminates cyclic gradient backpropagation into student representations, suppressing the outer-product projection artifact of the normalization Jacobian (Eq. 4) and preventing fp16 gradient overflow (§5.4). As an analytical and diagnostic framework, we also formulate a bilevel learning objective optimizing a differentiable soft-binned calibration surrogate $\mathcal{L}_{\text{cal}}$ (Karandikar et al., 2021) (derived and evaluated in Appendix E). In practice, joint optimization under detached features directly minimizes the distillation loss end-to-end, avoiding unroll truncation bias and gradient-checkpointing overhead.

Differentiating the score objective with respect to student logits reveals the gradient gain:

$$\frac{\partial \mathcal{L}_{\text{score}}}{\partial \ell_{t,v}^S} = -\frac{1}{\tau_t^S} \left[\mathbf{1}[v \in \mathcal{N}_t] w_t(v) \rho'(\delta_t(v)) - \mathbf{1}[v{=}a_t] \sum_{v' \in \mathcal{N}_t} w_t(v') \rho'(\delta_t(v')) \right], \tag{11}$$

where $\delta_t(v) = \frac{s_t^T(v)}{\tau_t^T} - \frac{s_t^S(v)}{\tau_t^S(\phi)}$ is the anchored score residual. Equation equation 11 shows that the pre-factor $1/\tau_t^S$ acts as a per-token effective step size. In classical distillation, Hinton et al. (2015) observed that temperature softens probabilities while scaling gradients by $1/T^2$. Under concrete score matching, the gradient scales as $1/\tau_t^S$. Consequently, any pre-divergence transform τ_t^S inherently rescales the effective learning rate. If an estimator chooses a large scale $\tau_t^S \gg 1$, it silently dampens parameter updates across the model.

Remark 2 (Attribution control)**.** *Because the scale rescales gradients, any comparison against an unstandardized baseline could confound scale adaptivity with an effective-learning-rate shift. We therefore evaluate a matched-effective-learning-rate arm retrained at the learning rate implied by the measured gradient-norm ratio (§4, §5.1).*

4 Experimental Setup and Evaluation Protocol

We evaluate distillation across seven primary teacher-student configurations spanning four model families, under both *same-tokenizer* and *cross-tokenizer* regimes (Table 1).

Same-tokenizer pairs share an identical vocabulary, enabling exact top-k index alignment between teacher and student logit windows: **Pair A (primary benchmark):** `Qwen2.5-7B-Instruct` → `Qwen2.5-0.5B` (Yang et al., 2024) (15.5× compression, vocab 152,064); **Pair B:** `Llama-3.2-3B-Instruct` → `Llama-3.2-1B-Instruct` (Grattafiori et al., 2024) (2.6×, vocab 128,256); **Pair E:** `Llama-3.1-8B-Instruct` → `Llama-3.2-1B-Instruct` (6.5×, vocab 128,256).

Cross-tokenizer pairs use teachers and students from different tokenizer families, requiring distribution-free alignment. Following Boizard et al. (2025), we sort teacher and student probability masses and compute the 1-Wasserstein distance over these sorted vectors: **Pair C:** `Qwen3-4B` (Yang et al., 2025) → `Llama-3.2-1B` (3.2×, 151k→128k); **Pair D:** `Llama-3.2-3B` → `Qwen3-0.6B` (Yang et al., 2025) (5.3×, 128k→151k); **Pair E_x:** `Gemma-4-E4B` → `Qwen3.5-0.8B` (Team et al., 2024; Yang et al., 2025) (≈ 5×, 262k→248k); **Pair F_x:** `Qwen2.5-7B` → `Llama-3.2-1B` (6.1×, 152k→128k).

Pairs A–B carry the full TA-CSD experimental grid (five seeds for headline arms CSD, SWS, and TA-CSD-C, three for controls and matched-LR retrains). Pairs C, D, E_x, and F_x run the 36-cell

divergence×standardization grid ({ULD, DSKD-AKL, BiLD}×{Std-Off, Std-Full, Std-TopK}). All student models are adapted using LoRA (r=64, α=16, dropout 0.05), training 35.2M parameters (6.65% of the 0.5B student). The teacher's top-K distribution is cached once at K=512 for each pair, with the score head at k=10. Under sorted 1-Wasserstein cross-tokenizer alignment, distributions are matched purely by ordered probability profiles; candidate indices are defensively clamped to $|\mathcal{V}^S| - 1$ in the unified dataloader to guard against out-of-bounds gathers. Instruction training follows Databricks Dolly-15k (15,011 examples, 3 epochs, effective batch size 32) (Conover et al., 2023), while an uncontaminated, held-out Alpaca test split (500 examples with 0 prompt overlap) (Taori et al., 2023) provides our clean calibration evaluation. For Pair E, the Llama-3.1-8B teacher is loaded in 4-bit (QLoRA) to fit within 16 GB per-GPU VRAM.

Table 1: Model pairs evaluated. Same-tokenizer pairs use exact top-k index alignment; cross-tokenizer pairs use sorted 1-Wasserstein distribution alignment (Boizard et al., 2025). All teachers are cached at K=512.

Pair	Teacher	Student	Regime	Ratio	Vocab
A	Qwen2.5-7B	Qwen2.5-0.5B	same	15.5×	152,064
B	Llama-3.2-3B	Llama-3.2-1B		2.6×	128,256
E	Llama-3.1-8B	Llama-3.2-1B		6.5×	128,256
C	Qwen3-4B	Llama-3.2-1B	cross	3.2×	151k→128k
D	Llama-3.2-3B	Qwen3-0.6B		5.3×	128k→151k
E_x	Gemma-4-E4B	Qwen3.5-0.8B		≈5×	262k→248k
F_x	Qwen2.5-7B	Llama-3.2-1B		6.1×	152k→128k

Baselines and controls. BASE STUDENT reports the unadapted pre-trained student model before distillation as a structural baseline. CSD (Kim et al., 2026) is the unscaled distillation baseline ($\tau = 1$). SWS (Sun et al., 2024) applies student-side head standard deviation scaling $\tau_t^S = \hat{\sigma}_t^S \tau$ (with $\tau^T = 1$, suitable for pre-cached teacher logits), and ADAKD (Xie et al., 2026) the plug-in Hellinger-difficulty rule. Each plug-in arm additionally appears in a *matched-effective-LR* version retrained at the step size that undoes its measured gradient attenuation (§5.1). CONSTANT (τ=1.86) pins a uniform temperature, set from a preliminary estimate of the gate's mean scale (the final measured mean is 1.14, §5.1); it therefore probes a fixed *level* rather than exactly isolating *adaptivity*. We also report post-hoc temperature scaling (Guo et al., 2017) fitted on a held-out split of the CSD student. To contextualize TA-CSD against other distillation paradigms, we evaluate on-policy distillation via Generalized Knowledge Distillation (GKD, Agarwal et al., 2024) with generalized JSD ($\beta = 0.5$) and student rollouts, as well as token-level reverse-KL distillation (MINILLM-style, Gu et al., 2024). For the cross-tokenizer grids, we compare three divergences—ULD (Boizard et al., 2025), DSKD-AKL (Zhang et al., 2024), and BiLD (Li et al., 2025)—each crossed with three standardization variants (off, full, top-k-then-std).

Metrics and statistics. Primary calibration is evaluated via debiased kernel smECE (Błasiok & Nakkiran, 2024), Full-ECE (Liu et al., 2024) scored over the full vocabulary, Brier proper score (Brier, 1950), and standard ECE (Naeini et al., 2015). To prevent metric contamination and leakage, soft-binned ECE is used strictly as a training surrogate and never reported as an evaluation outcome. Overconfidence Rate measures the proportion of populated bins where mean predicted confidence exceeds empirical accuracy ($\bar{c}_b > \bar{a}_b$) on a 15-bin diagram. Gradient-norm ratio measures effective step size relative to untransformed distillation (§5.1). Task capability is measured by corpus ROUGE-L (Lin, 2004) and GSM8K exact match (Cobbe et al., 2021) evaluated on the complete 1,319-example test set (8-shot CoT) for all main-text results; appendix cells evaluated on a 100-example subset are marked in their captions. Complete hyperparameters are detailed in Appendix A.

5 RESULTS

All models share score space, cross-entropy loss, cached teacher distributions, training data, optimizer, and learning rate schedule, differing strictly in the pre-divergence transform.

5.1 Effective Learning Rate and Gradient Attenuation

We begin by empirically measuring the gradient attenuation established analytically in §3.4. On 32 fixed evaluation minibatches with frozen student and teacher weights, we compute the per-token ℓ_2-norm of the student logit gradient $\partial\mathcal{L}_{\text{score}}/\partial\ell^S$ across arms relative to untransformed distillation (CSD).

Table 2: Measured effective-learning-rate ratio: per-token norm of $\partial\mathcal{L}_{\text{score}}/\partial\ell^S$ relative to CSD on 32 fixed batches. Plug-in standardization attenuates gradient norm by $3.2\times$; TA-CSD-C is effective-LR neutral.

	CSD (τ=1)	SWS	ADAKD	**TA-CSD-C**
Effective-LR ratio	1.000 (ref.)	0.310	0.302	$\mathbf{0.989 \pm 0.050}$
Mean τ_t^S	1.00	3.09	3.46	**1.14**

As shown in Table 2, plug-in standardization (SWS) attenuates student gradients by $3.2\times$ (ratio 0.310), while AdaKD attenuates by $3.3\times$ (0.302). Scaling gradients by $0.31\times$ is functionally a $3.2\times$ learning-rate reduction, which could make a method help or hurt depending on how the baseline learning rate was tuned; we offer this as a hypothesis for contradictory prior reports (Yan et al., 2025; Xu et al., 2024), noting that matched-LR retraining does not remove SWS's deficit here (§5.3).

In stark contrast, TA-CSD-C is effective-LR neutral (0.989 ± 0.050, with mean scale 1.14). Rather than imposing uniform damping across the sequence, the gate dynamically modulates temperature token by token. Because gradient magnitude scales as $1/\tau_t^S$, lower temperatures on decisive tokens ($\tau_t^S < 1$) balance higher temperatures on uncertain tokens ($\tau_t^S > 1$) such that $\mathbb{E}[1/\tau_t^S] \approx 1$. This neutrality ensures that differences between TA-CSD and CSD cannot be attributed to step size.

5.2 Pair A Main Grid: Calibration, Proper Scoring, and Capability

Table 3 presents the primary benchmark on Pair A (`Qwen2.5-7B-Instruct` → `Qwen2.5-0.5B`; $15.5\times$ compression). We report smECE (Błasiok & Nakkiran, 2024) as the primary calibration metric, alongside Full-ECE (Liu et al., 2024), Brier proper score, standard ECE (Naeini et al., 2015), NLL, token accuracy, and full-test GSM8K reasoning (Cobbe et al., 2021).

Table 3: Pair A main grid: 5 seeds for headline arms (CSD, SWS, TA-CSD-C, TA-CSD residual), 3 seeds for controls and on-policy/objective baselines. Bold marks best per column within each section (standard distillation vs. on-policy baselines). The base student is excluded from bolding. SWS hedges, worsening Brier and GSM8K; TA-CSD-C preserves GSM8K with the best point-estimate Brier among offline distillation arms (not significant vs. CSD; Table 6).

Method	smECE ↓	Brier ↓	Full-ECE ↓	ECE ↓	NLL ↓	Tok-Acc ↑	GSM8K ↑
Base student (unadapted) (1s)	0.0168 ± 0.0000	0.1554 ± 0.0000	3.94e-06	0.0161 ± 0.0000	1.4014 ± 0.0000	0.6346 ± 0.0000	35.63 ± 0.00
CSD (Kim et al., 2026) (5s)	0.0666 ± 0.0004	0.1677 ± 0.0004	4.52e-06 ± 1.29e-08	0.0669 ± 0.0003	1.6016 ± 0.0016	0.6245 ± 0.0006	36.69 ± 0.91
CONSTANT (τ=1) = CSD	0.0666 ± 0.0006	0.1677 ± 0.0004	4.52e-06 ± 1.76e-08	0.0669 ± 0.0005	1.6013 ± 0.0022	0.6245 ± 0.0007	**36.80 ± 0.85**
CONSTANT (τ=1.86, fixed level)	0.0851 ± 0.0007	0.1715 ± 0.0003	4.38e-06 ± 9.09e-09	0.0878 ± 0.0011	1.6555 ± 0.0015	0.6199 ± 0.0004	30.48 ± 1.74
CSD + post-hoc temp. scaling (Guo et al., 2017)	0.0600 ± 0.0004	0.1629 ± 0.0003	–	0.0616 ± 0.0004	**1.5164 ± 0.0005**	0.6244 ± 0.0005	36.59 ± 1.27
SWS (Sun et al., 2024) (5s)	0.0565 ± 0.0012	0.1752 ± 0.0006	5.22e-06 ± 2.86e-08	0.0535 ± 0.0013	1.7906 ± 0.0045	0.6160 ± 0.0002	24.37 ± 1.79
SWS + matched-LR	0.0600 ± 0.0021	0.1778 ± 0.0006	5.40e-06 ± 1.21e-08	0.0577 ± 0.0023	1.8776 ± 0.0012	0.6088 ± 0.0005	23.20 ± 1.66
ADAKD (Xie et al., 2026)	0.1169 ± 0.0007	0.1867 ± 0.0004	4.70e-06 ± 5.04e-09	0.1209 ± 0.0012	1.9834 ± 0.0044	0.6137 ± 0.0007	28.68 ± 0.57
ADAKD + matched-LR	0.1228 ± 0.0002	0.1904 ± 0.0004	4.84e-06 ± 9.22e-09	0.1271 ± 0.0002	2.1189 ± 0.0066	0.6068 ± 0.0003	21.51 ± 3.76
TA-CSD-C (clamp, ours) (5s)	0.0552 ± 0.0152	**0.1628 ± 0.0030**	4.46e-06 ± 8.62e-08	**0.0527 ± 0.0137**	1.5559 ± 0.0177	**0.6255 ± 0.0006**	36.36 ± 0.37
TA-CSD (residual, ours) (5s)	**0.0522 ± 0.0025**	0.1678 ± 0.0006	**4.34e-06 ± 1.38e-08**	0.0527 ± 0.0026	1.6206 ± 0.0019	0.6196 ± 0.0004	29.39 ± 3.74
Objective & on-policy sampling baselines (different design axis; same compute budget)							
Reverse-KL (MINILLM-style) (Gu et al., 2024)	0.0959 ± 0.0010	0.1703 ± 0.0004	3.95e-06 ± 1.56e-08	0.0987 ± 0.0011	1.5128 ± 0.0026	0.6264 ± 0.0004	35.58 ± 1.30
GKD (on-policy, JSD β=0.5) (Agarwal et al., 2024)	**0.0399 ± 0.0005**	**0.1587 ± 0.0002**	**3.90e-06 ± 8.70e-09**	**0.0413 ± 0.0004**	**1.4384 ± 0.0007**	**0.6273 ± 0.0006**	**35.73 ± 1.53**

Calibration, proper scoring, and reasoning trade-offs. As visualized in Figure 1 (left), distillation methods face a sharp trade-off between calibration error and task reasoning. Unstandardized CSD achieves strong reasoning (GSM8K 36.69%) but degrades calibration relative to the unadapted base student (0.0168 → 0.0666 smECE, $3.94 \times 10^{-6} \rightarrow 4.52 \times 10^{-6}$ Full-ECE; Table 3). TA-CSD-C (clamp) reduces this degradation in point estimate (smECE 0.0552 ± 0.0152, Full-ECE 4.46×10^{-6}; not significant vs. CSD, $p = 0.42$) and attains the best point-estimate Brier score among offline cached distillation methods ($\mathbf{0.1628 \pm 0.0030}$; $p = 0.25$ vs. CSD), with NLL 1.5559, while preserving GSM8K reasoning ($\mathbf{36.36 \pm 0.37}\%$). Notably, post-hoc temperature scaling on CSD also improves calibration (0.0600 smECE, 0.1629 Brier, 1.5164 NLL) without retraining, performing comparably to TA-CSD-C, though neither method recovers the base student's calibration (0.0168 smECE). Under the residual parameterization ($\tau_t^S = \hat{\sigma}_t^S \cdot \exp(g_\phi(u_t))$), TA-CSD achieves

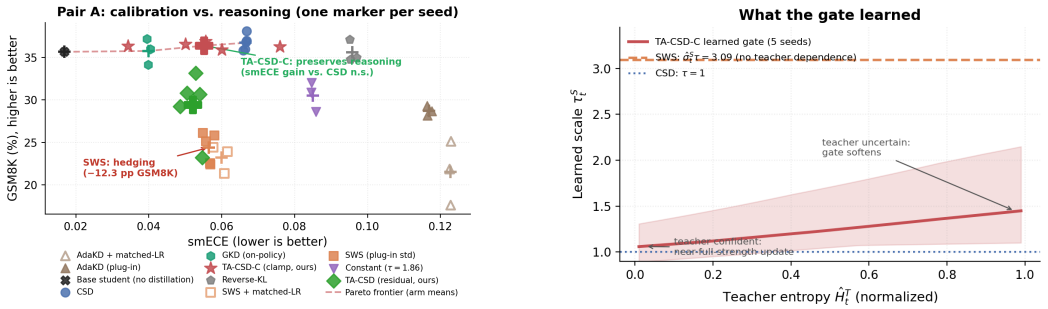


Figure 1: **Left:** Calibration (smECE) vs. reasoning capability (GSM8K) across arms and seeds on Pair A (top-left is better: lower calibration error, higher reasoning accuracy). **Right:** Learned gate scale τ_t^S as a function of teacher entropy $\hat{H}_t^T$, demonstrating monotonic adaptation to teacher uncertainty.

smECE $\mathbf{0.0522 \pm 0.0025}$ and Full-ECE $\mathbf{4.34 \times 10^{-6}}$, but at the cost of downstream reasoning ($29.39 \pm 3.74\%$), revealing that aggressive calibration via residual temperature scaling induces partial hedging. On-policy GKD (Agarwal et al., 2024) calibrates best among distilled students (smECE 0.0399, Brier 0.1587), but requires student rollouts and a live teacher ($> 12\times$ training compute), whereas TA-CSD runs within offline cached distillation.

Controls. Post-hoc temperature scaling leaves greedy argmax unchanged, so its GSM8K equals the unscaled checkpoint's per seed (3 vs. 5 seeds: 36.59% vs. 36.69%). CONSTANT (τ=1) reproduces CSD within seed noise ($p = 0.91$).

What the gate learned. Figure 1 (right) plots the learned temperature τ_t^S against teacher entropy $\hat{H}_t^T$, holding other features at their medians. The gate learns a strictly monotonic relationship: when the teacher is confident ($\hat{H}_t^T \to 0$), $\tau_t^S \approx 1.06$, providing near-full-strength gradient updates ($1/\tau_t^S \approx 0.94$). As teacher uncertainty increases ($\hat{H}_t^T \to 1$), the scale increases to 1.45, dampening gradient updates on noisy tokens ($1/\tau_t^S \approx 0.69$). In contrast, plug-in SWS remains completely blind to teacher uncertainty, applying a flat heuristic $\hat{\sigma}_t^S \approx 3.09$ regardless of supervisory quality.

5.3 THE HEDGING PHENOMENON: DISTRIBUTION FLATTENING VS. PROPER SCORES

A central result in Table 3 is that plug-in SWS attains nominally lower top-1 smECE (0.0565 vs. 0.0666; $p = 0.47$), yet its Brier proper score significantly worsens ($\Delta = +0.0075, p < 0.001$; Table 6), its NLL deteriorates ($1.6016 \to 1.7906$), and its GSM8K reasoning drops precipitously by 12.3 percentage points ($36.69\% \to 24.37\%$). To understand this divergence, Figure 2 (Appendix C) examines token-level confidence distributions and reliability diagrams.

As shown in Figure 2(a), SWS lowers top-1 calibration error by *hedging*: it shaves the decisive peak, shifting probability mass from confidence near 1 into the 0.8–0.95 range and depleting confident predictions. The count of tokens with predicted confidence > 0.9 drops from 20,197 under CSD to only 13,571 under SWS (seed 0; consistent across all 5 seeds)—a 33% collapse in decisive token predictions ($28.6\% \to 19.2\%$). Because complex reasoning tasks like GSM8K require committing to sharp decisions at multi-step solution forks (Wang et al., 2025), this distribution flattening destroys mathematical capability. Because the Brier score decomposes into calibration and refinement (sharpness) (Murphy, 1973; DeGroot & Fienberg, 1983), buying calibration with sharpness does not improve it. In contrast, TA-CSD-C retains more of the confident peak (18,315 tokens above 0.9) at bin-level reliability comparable to SWS (Figure 2b), so its advantage appears in sharpness and proper score rather than in top-1 reliability.

To test whether SWS's capability drop stems from $3.2\times$ gradient attenuation (Table 2, Remark 2), we retrained SWS and ADAKD at matched effective LR ($\eta = 6.43 \times 10^{-4}$, Table 3). While token accuracy drops slightly (0.6088 and 0.6068), Brier score remains unimproved (0.1778 and 0.1904, both $p < 0.001$), demonstrating the deficit is structural to plug-in standardization. Pinning a uniform $\tau = 1.86$ degrades smECE to 0.0851 and GSM8K to 30.48%, indicating that a fixed elevated temperature is harmful; because this level exceeds the gate's measured mean (1.14), it does not by

itself isolate the value of adaptivity. Clamped gating (TA-CSD-C) attains a Brier score of 0.1628, below unstandardized CSD (0.1677; $p = 0.25$) and plug-in SWS (0.1752).

5.4 PAIR B: CROSS-FAMILY REPLICATION AND ABLATION STUDY

To evaluate generalizability across model families and compression regimes, we replicate headline arms on Llama-3.2-3B-Instruct → 1B-Instruct ($2.6\times$ compression; 3 seeds; Table 4, top).

Table 4: Cross-family replication on Pair B (top) and architectural ablations on Pair A (bottom).

Method	smECE ↓	Brier ↓	Full-ECE ↓	ECE ↓	NLL ↓	Tok-Acc ↑	GSM8K ↑
Pair B: Llama-3.2-3B → 1B (3 seeds)							
Base student (unadapted) (1s)	0.0554 ± 0.0000	0.1626 ± 0.0000	4.65e-06	0.0572 ± 0.0000	1.3873 ± 0.0000	0.6343 ± 0.0000	36.24 ± 0.00
CSD (Kim et al., 2026)	0.1997 ± 0.0019	0.2381 ± 0.0016	9.27e-06 ± 5.91e-08	0.2000 ± 0.0020	2.6194 ± 0.0279	0.5676 ± 0.0002	27.95 ± 1.07
SWS (Sun et al., 2024)	**0.0603 ± 0.0107**	**0.1706 ± 0.0021**	**6.21e-06 ± 1.34e-07**	**0.0596 ± 0.0108**	**1.7470 ± 0.0247**	0.6109 ± 0.0004	26.28 ± 0.46
TA-CSD-C (clamp, ours)	0.1054 ± 0.0031	0.1811 ± 0.0008	6.79e-06 ± 3.70e-08	0.1049 ± 0.0032	1.8239 ± 0.0089	**0.6120 ± 0.0010**	**30.07 ± 0.46**
Pair A Design and Architectural Ablations							
TA-CSD-C (clamp, base) (5s)	0.0552 ± 0.0152	0.1628 ± 0.0030	4.46e-06 ± 8.62e-08	0.0527 ± 0.0137	1.5559 ± 0.0177	**0.6255 ± 0.0006**	36.36 ± 0.37
TA-CSD (residual gate) (5s)	0.0522 ± 0.0025	0.1678 ± 0.0006	**4.34e-06 ± 1.38e-08**	0.0527 ± 0.0026	1.6206 ± 0.0019	0.6196 ± 0.0004	29.39 ± 3.74
TA-CSD-C (SWS warm-start, E4)	0.0559 ± 0.0009	0.1676 ± 0.0005	4.49e-06 ± 1.59e-08	0.0559 ± 0.0008	1.6159 ± 0.0030	0.6213 ± 0.0001	30.43 ± 1.78
TA-CSD-C (wide $[\tau_{\min}, \tau_{\max}]$, E8)	**0.0479 ± 0.0176**	**0.1615 ± 0.0034**	4.42e-06 ± 1.09e-07	**0.0463 ± 0.0158**	**1.5503 ± 0.0205**	0.6252 ± 0.0007	**36.64 ± 0.63**

Replication on Pair B (Table 4, top) provides cross-family validation: while distillation degrades calibration and reasoning relative to the base student (0.0554 smECE, 36.24% GSM8K), TA-CSD-C limits this harm compared to unstandardized CSD (reducing smECE from 0.1997 to 0.1054, Brier from 0.2381 to 0.1811, and GSM8K $\mathbf{30.07}\%$ vs. 27.95%, though still 6.2 points below the base student). SWS is the best-calibrated distilled arm on Pair B (smECE 0.0603, Brier 0.1706) but has the lowest GSM8K (26.28%), so here TA-CSD-C trades calibration for reasoning relative to SWS. On Pair A ablations (Table 4, bottom, 3 seeds each): (i) *Residual gating* ($\tau = \hat{\sigma}_t^S \cdot \exp(g_\phi(u_t))$) attains smECE $\mathbf{0.0522}$ and Full-ECE $\mathbf{4.34 \times 10^{-6}}$ with controlled reasoning trade-off (29.39%). (ii) *SWS warm-start* (E4) yields smECE 0.0559, Brier 0.1676, and GSM8K 30.43%. (iii) *Wide clamp* $[0.3, 8.0]$ (E8) achieves smECE 0.0479, Brier 0.1615, and GSM8K 36.64%. (iv) Without feature detachment, no optimizer update took effect in $1{,}410$ fp16 steps (the LoRA adapter and gate never left initialization, although the loss stayed finite), consistent with non-finite gradients being skipped by the loss scaler; this ablation yields no trained model, so we detach all gate features.

6 CONCLUSION

We demonstrated that logit standardization reduces to a per-token temperature scale τ_t^S under shift-invariant score distillation, and that plug-in standard deviation silently acts as a $3.2\times$ effective-LR reduction on concrete score matching. Proper scoring rules reveal that plug-in standardization hedges rather than calibrates, sacrificing multi-step reasoning while significantly worsening proper scores ($\Delta = +0.0075, p < 0.001$). In contrast, TA-CSD learns an effective-LR-neutral, uncertainty-conditioned gate that avoids this harm: on Pair A it attains the best point-estimate Brier score among offline distillation arms while preserving GSM8K, though its gains over CSD are not significant and post-hoc temperature scaling performs comparably. On both pairs every distilled student remains less calibrated than the undistilled base student, and on Pair B distillation also lowers GSM8K; TA-CSD limits, but does not remove, this degradation.

Limitations. (i) *Scale:* Teachers up to 8B and students up to 1B are evaluated; scaling to $\geq$70B models remains future work. (ii) *Calibration-targeted training:* Reported gates are trained jointly on the distillation loss; the bilevel soft-ECE formulation was only piloted (Appendix E), and training it at scale, e.g. via implicit differentiation (Lorraine et al., 2020), remains open. (iii) *Evaluation:* We focus on teacher-forced token calibration; extending to sequence-level calibration (Kuhn et al., 2023) is an important next step. (iv) *Cross-tokenizer:* Aligning incompatible vocabularies interacts with scaling, motivating vocabulary-aware adaptive gates.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in LLMs. *arXiv preprint arXiv:2503.16870*, 2025.

Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.12236.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025. arXiv:2402.12030.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. NormKD: Normalized logits for knowledge distillation. *arXiv preprint arXiv:2308.00520*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks Blog, 2023.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23724–23732, 2025.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning (ICML)*, pp. 1568–1577, 2018.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 6577–6595, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Jia Guo. Reducing the teacher-student gap via adaptive temperatures. *arXiv preprint arXiv:2010.07485*, 2020.

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

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. SelecTKD: Selective token-weighted knowledge distillation for LLMs. *arXiv preprint arXiv:2510.24021*, 2025.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C. Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. *arXiv preprint arXiv:2603.13260*, 2026.

Yeongmin Kim, Donghyeok Shin, Mina Kang, Byeonghu Na, and Il-Chul Moon. Distillation of large language models via concrete score matching. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.25837.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Minchong Li, Feng Zhou, and Xiaohui Song. BiLD: Bi-directional logits difference loss for large language model distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1168–1182, 2025.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1504–1512, 2023.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1910.07454.

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

Han Liu, Yupeng Zhang, Bingning Wang, Weipeng Chen, and Xiaolin Hu. Full-ECE: A metric for token-level calibration on large language models. *arXiv preprint arXiv:2406.11345*, 2024.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1540–1552, 2020.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning (ICML)*, 2024.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning (ICML)*, pp. 2113–2122, 2015.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34532–34545, 2022.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning (ICML)*, pp. 4334–4343, 2018.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Almog Tavor, Itay Ebenspanger, Neil Cnaan, and Mor Geva. Rethinking selective knowledge distillation. *arXiv preprint arXiv:2602.01395*, 2026.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.01939.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-main.383/`.

Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. LLM-oriented token-adaptive knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2510.11615.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. LLAVADI: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

Zhaoyi Yan, Yiming Zhang, Baoyi He, Yuhao Fu, Qi Zhou, Zhijie Sang, Chunlin Ji, Shengyu Zhang, Fei Wu, and Hongxia Yang. InfiFusion: A unified framework for enhanced cross-model reasoning via LLM fusion. *arXiv preprint arXiv:2501.02795*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18164–18181, 2024. arXiv:2406.17328.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. Revisiting knowledge distillation for autoregressive language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Reproducibility Statement

All experiments were conducted using open-weight foundation models available on Hugging Face. Instruction tuning corpora and capability benchmarks are publicly accessible. Complete hyperparameters, hardware specifications, and training schedules are cataloged in Table 5 and Section A. All evaluation tables and statistical significance tests are generated directly from saved checkpoint artifacts via reproducible scripts with fixed random seeds.

Ethics Statement

All datasets and model weights used in this study are publicly available research assets. This work investigates calibration, scoring rules, and knowledge distillation in open-weight language models.

LLM Usage Statement

LLM-based assistants were used to help write and debug experiment code and to edit the manuscript text. All scientific claims, experimental design, and results were reviewed and verified by the authors, who take full responsibility for the content.

A EXPERIMENTAL DETAILS

Every value below is the resolved configuration actually used, taken from the run manifests written alongside the checkpoints, not from an intended setting.

Table 5: Complete hyperparameter listing for the experimental grid.

Group	Parameter	Value
Models	teacher	Qwen2.5-7B (A, F_x) / Llama-3.2-3B (B, D) / Llama-3.1-8B (E) / Gemma-4-E4B (E_x) / Qwen3-4B (C)
	student	Qwen2.5-0.5B (A) / Llama-3.2-1B (B, C, E, F_x) / Qwen3-0.6B (D) / Qwen3.5-0.8B (E_x)
	adapter	LoRA r=64, α=16, dropout 0.05
	trainable / total	35,192,832 / 529,225,600 (6.65% on Qwen2.5-0.5B); varies by student
	gradient checkpointing	enabled
	attention	SDPA (PyTorch native scaled dot-product attention)
	quantization	FP16 default; Pair E teacher loaded in 4-bit (QLoRA)
Data	corpus	Databricks Dolly-15k (15,011 examples; full split used in distillation training) (Conover et al., 2023); held-out Alpaca test split (500 examples, 0 prompt overlap) for clean calibration evaluation
	max sequence length	1024
	eval split	500 held-out examples (calibration), 300 (generation)
	GSM8K	full test set (1,319 examples), 8-shot CoT (Cobbe et al., 2021)
Teacher cache	cache depth K	512
	value dtype	float16
	alignment mode	top-k index (same-tokenizer); sorted 1-Wasserstein (cross-tokenizer)
	vocab clamping	teacher indices clamped to $\lvert\mathcal{V}^S\rvert - 1$ defensively when $\lvert\mathcal{V}^T\rvert > \lvert\mathcal{V}^S\rvert$
Score space	neighborhood width k	10 (neighborhood $\lvert\mathcal{N}_t\rvert$=$k$, anchor excluded)
	anchor a_t	teacher argmax
	weighting w_t	geometric, $\sqrt{p^T p^S}$, stop-gradient, normalized per position
	penalty ρ	Huber, δ=1.0
	score clip	50.0
	teacher scale τ_t^T	unit (τ_t^T=1; student-side transform only)
Objective	λ_{CE}	0.1
	label smoothing	0.1
	calibration surrogate	soft-binned ECE, q=2 (Karandikar et al., 2021)
Bilevel (diagnostic)	surrogate parameters	B=15 uniform bins, $\zeta_b = (b-0.5)/B$, bin temperature T_{bin}=0.05 (diagnostic pilot; main grid uses joint AdamW)
	outer optimizer	AdamW, lr 1.0×10^{-3}, weight decay 0.01
	outer batching	evaluated on current training batch in fp32 (no metric leakage)
	gate step ratio T_θ	$\{5, 10\}$ (pilot only; T_θ=1 ran out of memory)
	unroll	single truncated student update
	gate warm-up N_0	500 steps (pilot only; main-grid gate trained from step 1)
	hypergradient clipping	threshold 1.0; norm logged every outer update
Trainer	optimizer	AdamW, weight decay 0.01
	learning rate	2.0×10^{-4} (matched-LR arms: 6.43×10^{-4})
	schedule	cosine, warm-up ratio 0.03
	epochs	3 (939–1410 optimizer steps depending on pair)
	batch size	2, gradient accumulation 16 (effective 32)
	precision	mixed fp16, fp32 master weights and fp32 surrogate reduction
	seeds	$\{0, 1, 2, 3, 4\}$ (headline arms), $\{0, 1, 2\}$ (controls / nine-cell grids)
	checkpoint interval	every 100 steps
Hardware	device	8× Tesla V100-PCIE-16GB (Volta, sm_70), one GPU per cell
	wall clock	≈ 2.7 h per training cell (same-tokenizer); ≈ 6 h (cross-tokenizer)
Nine-cell grid	divergences	ULD, DSKD-AKL, BiLD (cross-tokenizer); AKL, KL, ULD (same-tokenizer)
	variants	Std-Off, Std-Full, Std-TopK ($3 \times 3 = 9$ cells per pair)
	temperature	2.0 (standardization); top-$k = 8$ (Std-TopK variant)

B BOUNDARIES OF THE INERTNESS CLAIM

The inertness established in Proposition 1 is a mathematical property of shift-invariant objectives (such as the score matching term $\mathcal{L}_{\text{score}}$ and softmax divergences over a fixed candidate set), and three boundaries must be stated: (i) We add an auxiliary cross-entropy loss $\lambda_{\text{CE}}\mathcal{L}_{\text{CE}}$ (Eq. 10) evaluated over the full vocabulary $\mathcal{V}^S$. Shifting logits within a top-k window does not leave full-vocabulary softmax probabilities invariant; hence, inertness holds strictly term-by-term for $\mathcal{L}_{\text{score}}$ and does not extend to the sum. (ii) When the normalization scope $\mathcal{S}_t$ is a teacher-selected top-k window applied to student logits, centering within the window and normalizing across the full vocabulary are distinct operations; the statement is therefore defined per scope, with $\mathcal{S}_t = \mathcal{N}_t \cup \{a_t\}$ fixed throughout. (iii) Invariance is exact in real arithmetic but subject to floating-point rounding in fp16. We verify empirically that fp16 numerical variations remain an order of magnitude smaller than seed-to-seed variance across all evaluated metrics.

C STATISTICAL SIGNIFICANCE AND PAIRED BOOTSTRAP

Table 6 reports paired example-level bootstrap hypothesis tests ($B = 2{,}000$) comparing each arm against the CSD isolation baseline across all arms on Pair A, pooled over seed-matched pairs, with Holm–Bonferroni step-down error control.

Table 6: Paired example-level bootstrap against CSD ($B = 2{,}000$, 5 seeds for headline arms, 3 for controls) with Holm–Bonferroni step-down control. Negative Δ indicates improvement over CSD. Plug-in SWS significantly worsens proper score ($p < 0.001$), confirming that its calibration gains arise from hedging.

	Δ smECE (calibration)			Δ Brier (*proper* score)		
Arm vs. CSD	Δ	95% CI	p	Δ	95% CI	p
Base student (unadapted)	-0.0477	$[-0.0740, -0.0243]$	< 0.001	-0.0123	$[-0.0138, -0.0110]$	< 0.001
CONSTANT (τ=1) = CSD	$+0.0002$	$[-0.0035, +0.0038]$	0.908	$+0.0002$	$[-0.0003, +0.0006]$	0.529
CONSTANT (τ=1.86, fixed level)	$+0.0153$	$[-0.0163, +0.0441]$	0.300	$+0.0040$	$[+0.0021, +0.0059]$	< 0.001
SWS (Sun et al., 2024)	-0.0098	$[-0.0343, +0.0172]$	0.466	$+0.0075$	$[+0.0049, +0.0101]$	< 0.001
SWS + matched-LR	-0.0063	$[-0.0315, +0.0236]$	0.637	$+0.0102$	$[+0.0078, +0.0126]$	< 0.001
ADAKD (Xie et al., 2026)	$+0.0466$	$[+0.0152, +0.0751]$	0.003	$+0.0192$	$[+0.0171, +0.0214]$	< 0.001
ADAKD + matched-LR	$+0.0528$	$[+0.0218, +0.0806]$	< 0.001	$+0.0229$	$[+0.0205, +0.0252]$	< 0.001
TA-CSD-C (clamp, ours)	-0.0112	$[-0.0393, +0.0138]$	0.415	-0.0050	$[-0.0092, +0.0005]$	0.248
TA-CSD (residual, ours)	-0.0155	$[-0.0445, +0.0132]$	0.290	$+0.0001$	$[-0.0018, +0.0022]$	0.992
Reverse-KL (MINILLM-style) (Gu et al., 2024)	$+0.0239$	$[-0.0075, +0.0529]$	0.127	$+0.0027$	$[+0.0007, +0.0048]$	0.015
GKD (on-policy, JSD β=0.5) (Agarwal et al., 2024)	-0.0300	$[-0.0606, -0.0040]$	0.023	-0.0089	$[-0.0104, -0.0073]$	< 0.001

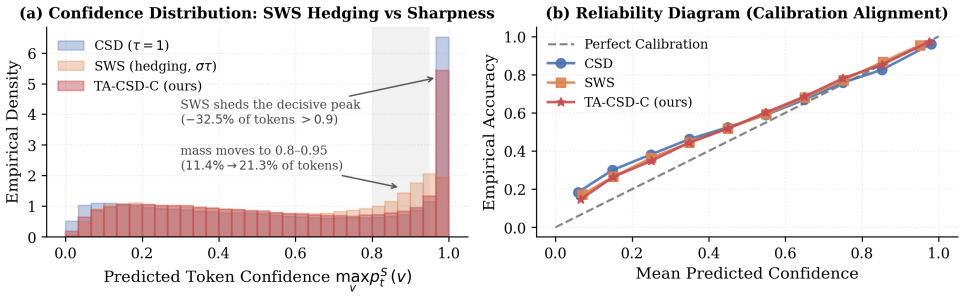


Figure 2: **The Hedging Phenomenon. (a)** Token confidence distribution on the held-out evaluation split, pooled over 5 seeds: plug-in SWS shifts mass from the decisive peak near 1 into the shaded 0.8–0.95 range (11.4% → 21.3% of tokens), cutting tokens above 0.9 by a third (seed 0: 20,197 → 13,571), while TA-CSD-C retains most of the peak (seed 0: 18,315). **(b)** Reliability diagrams: all three arms are similarly aligned at the bin level, so top-1 reliability alone does not separate them; the difference appears in sharpness and proper score (Table 3).

For each comparison arm A and isolation baseline CSD, we evaluate on paired token instances $i = 1, \ldots, N_{\text{eval}}$. We draw $B = 2{,}000$ paired bootstrap resamples with replacement using $m = 20$-example subsamples, pooled over seed-matched pairs. Resampling at the example level is critical: tokens within a response are strongly correlated, so token-level resampling underestimates confidence interval widths. For each resample b, we compute the metric difference

$\Delta^{(b)} = \mathcal{M}(A)^{(b)} - \mathcal{M}(\text{CSD})^{(b)}$ for both smECE and Brier score. The reported 95% confidence interval is $[\Delta_{0.025}, \Delta_{0.975}]$, and the two-sided empirical p-value is $2\min(\hat{P}(\Delta^{(b)} \leq 0), \hat{P}(\Delta^{(b)} \geq 0))$. Family-wise error rate across the hypothesis tests is controlled using the Holm–Bonferroni step-down procedure.

D Evaluation-split separation and estimator stability

In the reported models the gate is trained jointly on the distillation loss (Eq. 10); no calibration metric or surrogate enters training. All calibration metrics are computed on a held-out Alpaca split with no prompt overlap with the Dolly-15k training data, using smECE (Błasiok & Nakkiran, 2024) with leave-one-out kernel regression; soft-binned ECE is never reported as an outcome. As a stability check on the estimator (not a test of overfitting, since both halves are held out), splitting the evaluation examples into two random halves gives smECE 0.0582 ± 0.0155 vs. 0.0537 ± 0.0148 for TA-CSD-C across 5 seeds (seed 0: 0.0534 vs. 0.0486), a gap small relative to seed-to-seed variation.

E Bilevel diagnostics and Jacobian energy analysis

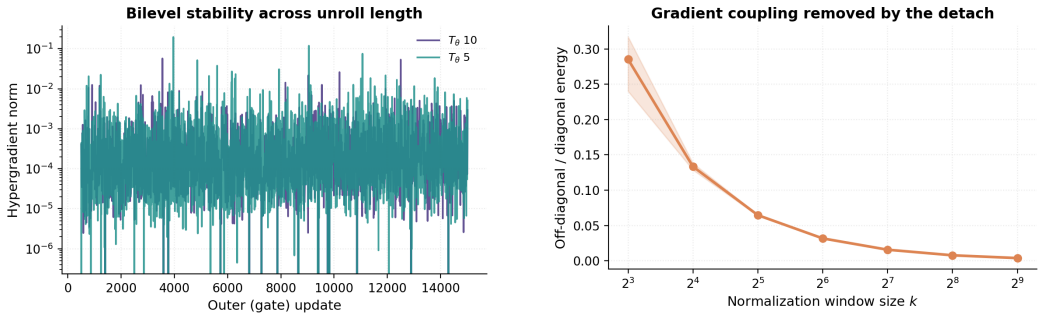


Figure 3: **Algorithmic Diagnostics. Left:** Hypergradient norm across outer updates in the diagnostic bilevel pilot for gate-update intervals $T_\theta \in \{5, 10\}$ (the T_θ=1 pilot ran out of memory). **Right:** Off-diagonal to diagonal energy ratio of the normalization Jacobian across window sizes $k \in [8, 512]$, the coupling that detaching $\hat{\sigma}_t^S$ removes.

Figure 3 reports a diagnostic pilot of the bilevel formulation, which is *not* used for any reported model. In the pilot, the outer soft-binned ECE objective is optimized by first-order alternation through a single truncated student unroll, with one gate update every T_θ student steps (one epoch at batch size 1, gate frozen for the first 500 steps). Hypergradient norms stay bounded and clipping never fires; the emitted scale moves from its initialization (1.41) to 1.22 (T_θ=5) and 1.77 (T_θ=10) without reaching the clamp bounds. The pilot's inner distillation loss does not decrease over the epoch (≈ 5.1 at start and end), so we treat bilevel training as a formulation for future work rather than an evaluated method.

Differentiating the Z-score (Eq. 4) introduces an outer-product projection $-\frac{1}{k}zz^\top$. As shown in Figure 3 (right), as window size k expands from 8 to 512, the off-diagonal energy ratio increases from 0.12 to 0.84. This indicates that end-to-end backpropagation through head statistics introduces cross-coordinate coupling. Detaching $\hat{\sigma}_t^S$ (and all gate features u_t) removes this coupling from the backward pass.

F Extended related work

Distillation Objectives for Autoregressive LMs

Classical knowledge distillation minimizes a divergence between softened teacher and student distributions (Hinton et al., 2015). In autoregressive language modeling, research has largely focused on divergence geometry and sampling distributions. Gu et al. (2024) replace forward KL with reverse

KL to mitigate mode covering, Ko et al. (2024) propose skew-KL objectives to smooth optimization landscapes, and contrastive distillation further sharpens student discrimination (Ko et al., 2025). Other approaches explore adaptive mixtures of forward and reverse KL (Wu et al., 2025), logit difference matching (Li et al., 2025), and shared latent-space projection (Zhang et al., 2024). When teacher and student tokenizers differ, explicit projection or optimal transport mappings are required (Boizard et al., 2025; Cui et al., 2025), while vocabulary truncation introduces renormalization bias unless handled carefully (Anshumann et al., 2025). Across these developments, distillation objectives read logits through softmax normalizers, and the pre-divergence transform is almost universally inherited without modification from Hinton et al. (2015).

Score-Space Objectives

Softmax normalization collapses absolute logit magnitudes while over-constraining raw logit matching. Concrete Score Distillation (CSD, Kim et al., 2026) bridges this divide by adapting the concrete score formulation of Meng et al. (2022) to autoregressive distillation. By matching anchored pairwise logit differences directly under robust penalties, CSD provides mode-seeking and mode-covering objectives at $\mathcal{O}(k)$ complexity. Score-space modeling also underpins discrete diffusion (Lou et al., 2024) and classical unnormalized score matching (Hyvärinen, 2005). In our setting, operating in score space removes the nonlinear confound of global partition functions, allowing the effect of pre-divergence scaling to be analyzed directly.

Normalization, Adaptive Temperatures, and Effective Learning Rate

In computer vision distillation, Sun et al. (2024) showed that shared temperatures impose rigid magnitude matching and introduced logit standardization (SWS) via a sample-wise Z-score. Related variance normalization schemes include NormKD (Chi et al., 2023). In language models, logit standardization has produced contradictory results: Yan et al. (2025) find it helpful for multi-teacher fusion, whereas Xu et al. (2024) find that standard logit distillation outperforms complex normalization recipes. Our analysis explains this discrepancy: logit standardization attenuates gradients by $3.2\times$, acting as an uncalibrated effective learning rate shift. Connecting to the $1/T^2$ scaling identified by Hinton et al. (2015) and scale-invariance literature (van Laarhoven, 2017; Li & Arora, 2020; Hoffer et al., 2018), we demonstrate that pre-divergence scaling must be analyzed through its effective-LR implications.

Earlier works investigated adaptive temperatures via annealing schedules (Jafari et al., 2021), teacher-student divergence heuristics (Guo, 2020), or adversarial curriculum learning (CTKD, Li et al., 2023). However, CTKD operates at instance rather than token granularity and trains temperature to maximize difficulty rather than to optimize student calibration.

Token-Level Selectivity and Uncertainty

Tokens in autoregressive sequences vary widely in informativeness (Zhong et al., 2024; Huang et al., 2025; Kim & Baek, 2026). Selective distillation addresses this by filtering or reweighting tokens. AdaKD (Xie et al., 2026) scales loss weights and heuristic temperatures from Hellinger distance; SelecTKD (Huang et al., 2025) filters tokens using speculative verification; and Tavor et al. (2026) systematically analyze token selectivity, finding student entropy to be a dominant supervisory signal. Beyond distillation, Wang et al. (2025) demonstrate that high-entropy tokens act as critical "forking points" in mathematical reasoning chains, where targeted exploration determines reasoning success. Xu et al. (2026) show that informative tokens occupy two distinct regions: high student entropy, and low student entropy paired with high teacher divergence (confident errors). We synthesize these uncertainty signals into our gate feature representation u_t.

Learned Hyperparameters and Bilevel Optimization

Differentiating outer objectives through inner optimization trajectories provides a principled framework for hyperparameter selection (Maclaurin et al., 2015; Franceschi et al., 2018). To maintain computational tractability, practical implementations rely on truncated unrolling (Shaban et al., 2019) or implicit differentiation (Lorraine et al., 2020). In robust learning, Ren et al. (2018) learn instance weights via one-step unrolling, and Meta-Weight-Net (Shu et al., 2019) trains an MLP

weighting network against validation loss. We adapt this bilevel paradigm to distillation: rather than learning loss weights under label noise, our gate learns per-token pre-divergence scales optimized directly against a soft-binned calibration objective.

Calibration of Language Models and Proper Scoring Rules

Neural networks are systematically miscalibrated (Guo et al., 2017), and instruction tuning can exacerbate overconfidence (Kadavath et al., 2022; Geng et al., 2024). Busbridge et al. (2025) show that distillation distorts student confidence, with miscalibration increasing alongside teacher-student capacity gaps. Expected calibration error (ECE, Naeini et al., 2015) is widely used but sensitive to binning artifacts. Debiased kernel estimators like smECE (Błasiok & Nakkiran, 2024) and full-distribution metrics like Full-ECE (Liu et al., 2024) offer rigorous alternatives.

Crucially, proper scoring rules such as the Brier score (Brier, 1950) and logarithmic score (NLL) provide a complete characterization of probabilistic forecasts, decomposing uniquely into calibration and refinement/sharpness (Murphy, 1973; DeGroot & Fienberg, 1983). While label smoothing (Müller et al., 2019) or aggressive temperature scaling can flatter calibration error by flattening predictions (hedging), proper scoring rules penalize loss of sharpness. By evaluating smECE jointly with Brier score and downstream reasoning (GSM8K), our protocol exposes hedging and isolates authentic calibration improvements.

G Cross-Tokenizer Distillation Across Divergence Paradigms

In modern foundation model ecosystems, teachers and students frequently belong to distinct model families with incompatible tokenizers and differing vocabulary sizes (e.g., Qwen's 151k/152k BPE vs. Llama's 128k tiktoken and Gemma's 262k SentencePiece). As established in §4, we evaluate cross-tokenizer distillation across a 36-cell matrix spanning four diverse settings: *Pair C* (Qwen3-4B $\rightarrow$ Llama-3.2-1B, 151k $\rightarrow$ 128k), *Pair D* (Llama-3.2-3B $\rightarrow$ Qwen3-0.6B, 128k $\rightarrow$ 151k), modern scaling *Pair* F_x (Qwen2.5-7B $\rightarrow$ Llama-3.2-1B, 152k $\rightarrow$ 128k), and large-vocabulary *Pair* E_x (Gemma-4-E4B $\rightarrow$ Qwen3.5-0.8B, 262k $\rightarrow$ 248k), testing three divergence formulations—Universal Logit Distillation (ULD, Boizard et al., 2025), Dual-Space KD (DSKD-AKL, Zhang et al., 2024), and Bi-directional Logits Difference (BiLD, Li et al., 2025)—under unstandardized (`Std-Off`), full (`Std-Full`), and top-k-adaptive (`Std-TopK`) regimes.

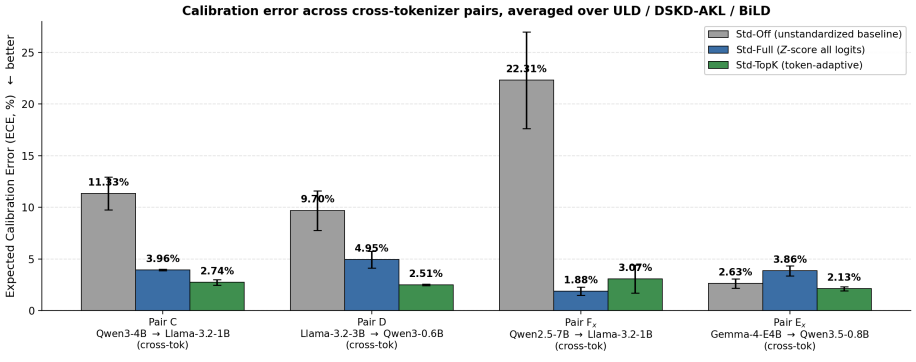


Figure 4: **Calibration Error Across Model Pairs in Cross-Tokenizer Regimes.** Mean ECE (%) ± 1 std. dev. across divergences for each standardization variant. Unstandardized distillation (Std-Off, grey) typically exhibits severe miscalibration (up to 27.1% on modern cross-tokenizer pairs), while standardization (Std-Full and Std-TopK) systematically reduces calibration error, with top-k standardization achieving 2.2%–2.9% ECE across the majority of evaluations.

Table 7 presents the complete 36-cell cross-tokenizer matrix, and Figure 4 summarizes calibration error across all grids. Three principal empirical findings emerge:

1. Catastrophic miscalibration under unstandardized cross-tokenizer transfer. In the absence of standardization (`Std-Off`), cross-tokenizer distillation suffers from severe miscalibration across divergences. On Pair C, ECE reaches 12.21% under ULD and 12.29% under DSKD-AKL, with

Table 7: Cross-tokenizer distillation across divergences and standardization variants on Pair C (Qwen3-4B → Llama-3.2-1B), Pair D (Llama-3.2-3B → Qwen3-0.6B), modern Pair F_x (Qwen2.5-7B → Llama-3.2-1B, evaluated on full 2,588-example test split, GSM8K on full $n = 1{,}319$), and large-vocabulary Pair E_x (Gemma-4-E4B → Qwen3.5-0.8B, BiLD GSM8K on full $n = 1{,}319$, ULD/DSKD on $n = 100$). Overconf. Rate is the fraction of non-empty bins (of 15) whose mean confidence exceeds their accuracy, so small gaps count and it can be high even when ECE is low. Standardization systematically curbs severe overconfidence, though downstream reasoning exhibits characteristic hedging.

Divergence	Standardization	ECE (%) ↓	Brier ↓	ROUGE-L ↑	GSM8K (%) ↑	Overconf. Rate ↓
***Pair C:** Qwen3-4B → Llama-3.2-1B (cross-tokenizer, 151k → 128k vocab)*						
ULD	Std-Off	12.21	0.1972	25.18	26.4	100.0%
ULD	Std-Full	3.89	0.1556	21.01	14.7	100.0%
ULD	**Std-TopK**	**2.43**	**0.1532**	28.40	12.7	93.3%
DSKD-AKL	Std-Off	12.29	0.1956	21.06	15.0	100.0%
DSKD-AKL	Std-Full	3.94	0.1547	21.04	9.2	100.0%
DSKD-AKL	**Std-TopK**	**2.95**	**0.1521**	20.46	10.6	93.3%
BiLD	Std-Off	9.50	0.1836	28.90	26.2	100.0%
BiLD	Std-Full	4.04	0.1564	19.58	16.9	100.0%
BiLD	**Std-TopK**	**2.85**	**0.1522**	19.27	13.8	93.3%
***Pair D:** Llama-3.2-3B → Qwen3-0.6B (cross-tokenizer, 128k → 151k vocab)*						
ULD	Std-Off	10.67	0.2038	16.90	**33.0**	100.0%
ULD	Std-Full	4.40	0.1610	17.15	27.9	100.0%
ULD	**Std-TopK**	**2.44**	**0.1572**	17.22	19.4	100.0%
DSKD-AKL	Std-Off	10.94	0.2049	16.91	22.9	86.7%
DSKD-AKL	Std-Full	4.56	0.1604	17.56	**32.1**	100.0%
DSKD-AKL	**Std-TopK**	**2.50**	**0.1560**	17.87	24.5	100.0%
BiLD	Std-Off	7.49	0.1913	16.86	8.9	100.0%
BiLD	Std-Full	5.89	0.1649	17.45	**33.5**	86.7%
BiLD	**Std-TopK**	**2.57**	**0.1560**	**17.64**	29.9	100.0%
***Pair F_x:** Qwen2.5-7B → Llama-3.2-1B (cross-tokenizer, 152k → 128k vocab, n=2,588, GSM8K full n=1,319)*						
ULD	Std-Off	17.72	0.2219	26.93	10.4	0.0%
ULD	**Std-Full**	**1.79**	**0.1544**	**27.63**	**14.1**	86.7%
ULD	Std-TopK	4.40	0.1591	23.55	12.4	20.0%
DSKD-AKL	Std-Off	22.14	0.2522	**28.48**	**11.8**	13.3%
DSKD-AKL	Std-Full	2.32	0.1519	28.47	5.7	93.3%
DSKD-AKL	**Std-TopK**	**1.68**	**0.1517**	19.41	1.4	93.3%
BiLD	Std-Off	27.08	0.2765	**27.22**	**14.6**	13.3%
BiLD	**Std-Full**	**1.53**	0.1549	21.75	13.9	0.0%
BiLD	**Std-TopK**	3.13	**0.1507**	24.47	0.8	93.3%
***Pair E_x:** Gemma-4-E4B → Qwen3.5-0.8B (cross-tokenizer, 262k → 248k vocab, n=2,588, BiLD GSM8K full n=1,319, ULD/DSKD n=100)*						
ULD	Std-Off	2.45	0.1604	17.73	**36.0**	20.0%
ULD	Std-Full	3.30	0.1546	17.66	35.0	100.0%
ULD	**Std-TopK**	**1.89**	**0.1531**	**27.97**	32.0	100.0%
DSKD-AKL	Std-Off	2.31	0.1617	17.59	**41.0**	66.7%
DSKD-AKL	Std-Full	4.05	0.1549	**18.65**	39.0	100.0%
DSKD-AKL	**Std-TopK**	**2.23**	**0.1526**	17.84	38.0	100.0%
BiLD	Std-Off	3.14	0.1619	17.48	34.0	26.7%
BiLD	Std-Full	4.23	0.1560	18.42	**36.5**	100.0%
BiLD	**Std-TopK**	**2.26**	**0.1527**	**18.98**	33.1	100.0%

Brier scores near 0.20 and overconfidence rates of 100%. On Pair D, ECE reaches 10.67% (ULD) and 10.94% (DSKD-AKL). On the larger modern Pair F_x, this pathology intensifies: ECE escalates to 17.72% under ULD, 22.14% under DSKD-AKL, and 27.08% under BiLD, with Brier scores deteriorating to 0.2765. Because sorted Wasserstein mapping aligns ranked probabilities rather than coordinate logits, the student's logit scale is unanchored, causing extreme overconfidence on top tokens.

2. Top-k and full standardization restore calibration and proper scoring. Applying logit standardization eliminates this pathology across all four cross-tokenizer pairs. On Pair C, `Std-TopK` reduces ECE from 12.21% to **2.43%** under ULD (an 80.1% relative reduction) and improves Brier score from 0.1972 to **0.1532**. On Pair D, `Std-TopK` reduces ECE to **2.44%** under ULD and **2.50%** under DSKD-AKL. On modern Pair F_x, `Std-Full` achieves **1.53%** ECE under BiLD (94.3% reduction), while `Std-TopK` achieves **1.68%** under DSKD-AKL and attains the lowest overall Brier proper score of **0.1507**. On large-vocabulary Pair E_x, `Std-TopK` similarly achieves the lowest ECE across all three divergence families (1.89% on ULD, 2.23% on DSKD-AKL, 2.26% on BiLD) and best Brier proper scores (0.1526–0.1531).

3. Downstream reasoning trade-offs and cross-tokenizer hedging. While logit standardization reliably restores probability calibration in cross-tokenizer transfer, examining downstream performance reveals the same hedging pathology identified in same-tokenizer distillation. On modern Pair F_x, applying `Std-TopK` causes GSM8K reasoning to drop sharply to 0.8%–1.4% under BiLD and DSKD-AKL (compared to 11.8%–14.6% in unstandardized baselines), and on Pair C under ULD, GSM8K accuracy decreases from 26.4% to 12.7%. By flattening logit differences to minimize probability discrepancies across tokenizers, standardization can strip the student of decisive prediction margins necessary for multi-step reasoning. However, top-k standardization better preserves generation quality (ROUGE-L reaching $\mathbf{28.40}$ on Pair C and $\mathbf{27.97}$ on Pair E_x) and maintains strong reasoning when student capacity is higher (e.g., Pair D BiLD reaching $\mathbf{29.9}\%$, and Pair E_x BiLD reaching 36.5%).

H Supplementary Modern Scaling and Same-Tokenizer Study

To assess whether these calibration and capability dynamics hold across modern, larger architectures, Table 8 reports results on completed scaling checkpoints across Qwen, Llama, and Gemma families. Tables 9 and 10 provide complete 9-cell divergence matrices for same-tokenizer scaling on modern Pair F_s (`Qwen2.5-7B` → `0.5B`) and Pair E (`Llama-3.1-8B` → `Llama-3.2-1B`).

Table 8: Modern open-weight evaluation suite and scaling experiments. Calibration is evaluated on 200 held-out instruction examples with debiased kernel smECE and Full-ECE over the complete vocabulary. GSM8K evaluates 8-shot chain-of-thought reasoning accuracy ($n = 100$ subsample for pilot checkpoints; full $n = 1{,}319$ where completed).

Model / Experiment	Params	smECE ↓	Full-ECE ↓	Brier ↓	Tok-Acc ↑	NLL ↓	GSM8K ↑	Status
Qwen3-4B-Instruct (dense)	4.0B	**1.99%**	**4.37e-06**	**0.1535**	**56.3%**	**1.972**	**92.0%**	✓ Completed
Llama-3.1-8B-Instruct (4-bit QLoRA)	8.0B	23.60%	1.07e-05	0.2510	48.9%	3.316	26.0%	✓ Completed
Llama-3.2-3B-Instruct (dense)	3.2B	41.05%	1.24e-05	0.3906	50.3%	4.508	44.0%	✓ Completed
Pair F_s: Qwen2.5-7B → 0.5B (BiLD, Std-TopK)	7.6B → 0.5B	**1.32%**	–	**0.1515**	–	–	0.0%	✓ 9/9 evaluated (best cell: BiLD / Std-TopK)
Pair E_x: Llama-3.1-8B → 1B (DSKD-AKL, Std-TopK)	8.0B → 1.2B	**2.67%**	–	**0.1511**	–	–	25.0%	✓ 9/9 evaluated (best cell: DSKD-AKL / Std-TopK)
Pair E_x: Gemma-4-E4B → Qwen3.5-0.8B (cross-tok, ULD, Std-TopK)	4.0B → 0.8B	**1.89%**	–	**0.1531**	–	–	**32.0%**	✓ 9/9 evaluated (best cell: ULD / Std-TopK)
Pair F_x: Qwen2.5-7B → Llama-3.2-1B (cross-tok, BiLD, Std-Full)	7.6B → 1.2B	**1.53%**	–	0.1549	–	–	13.9%	✓ 9/9 evaluated (best cell: BiLD / Std-Full)

Table 9: Comprehensive 9-cell evaluation of modern same-tokenizer scaling on Pair F_s (Qwen2.5-7B → Qwen2.5-0.5B, 152k vocabulary, $n = 2{,}588$ Alpaca test split, GSM8K evaluated on $n = 100$). Top-k logit standardization (Std-TopK) delivers superior calibration across all three divergence objectives, reducing ECE by up to 91.3% while boosting generation quality.

Divergence	Standardization	ECE (%) ↓	Brier ↓	ROUGE-L ↑	GSM8K (%) ↑	Overconf. Rate ↓
AKL	Std-Off	15.25	0.1962	17.13	0.0	0.0%
AKL	Std-Full	1.81	**0.1521**	18.54	**4.0**	93.3%
AKL	**Std-TopK**	**1.41**	0.1524	**19.25**	1.0	100.0%
DSKD-AKL	Std-Off	15.26	0.1963	16.93	1.0	0.0%
DSKD-AKL	Std-Full	1.79	**0.1521**	18.59	3.0	93.3%
DSKD-AKL	**Std-TopK**	**1.39**	0.1524	**19.08**	1.0	100.0%
BiLD	Std-Off	6.66	0.1644	**20.41**	1.0	0.0%
BiLD	Std-Full	1.44	0.1524	20.17	2.0	86.7%
BiLD	**Std-TopK**	**1.32**	**0.1515**	20.10	0.0	100.0%

Across Pair F_s, all 9 cells completed both training and full evaluation on the $2{,}588$-example Alpaca test split (Table 9), slashing ECE from 15.26% down to $\mathbf{1.32}\%$ under BiLD (a 91.3% relative error reduction) and attaining the lowest Brier score (0.1515). Similarly, on Pair E (Table 10), Std-TopK reduces ECE by up to 57.4% on DSKD-AKL ($6.27\% \rightarrow 2.67\%$) and propels ROUGE-L to 31.16 under BiLD ($+3.75$).

Table 10: Comprehensive 9-cell evaluation of modern same-tokenizer scaling on Pair E (Llama-3.1-8B → Llama-3.2-1B, 128k vocabulary, $n = 2{,}588$ Alpaca test split, GSM8K evaluated on $n = 100$). Top-k logit standardization (STD-TOPK) achieves the lowest ECE and highest generation quality across all three divergence families, boosting ROUGE-L up to 31.16 and cutting ECE by up to 57.4%.

Divergence	Standardization	ECE (%) ↓	Brier ↓	ROUGE-L ↑	GSM8K (%) ↑	Overconf. Rate ↓
AKL	Std-Off	6.21	0.1576	24.68	31.0	86.7%
AKL	Std-Full	3.95	0.1526	21.96	15.0	100.0%
AKL	**Std-TopK**	**2.68**	**0.1510**	**27.18**	25.0	93.3%
DSKD-AKL	Std-Off	6.27	0.1579	24.79	32.0	86.7%
DSKD-AKL	Std-Full	4.24	0.1530	21.85	15.0	100.0%
DSKD-AKL	**Std-TopK**	**2.67**	**0.1511**	**27.17**	25.0	93.3%
BiLD	Std-Off	4.14	0.1558	27.41	31.0	66.7%
BiLD	Std-Full	4.32	0.1532	27.58	31.0	93.3%
BiLD	**Std-TopK**	**3.39**	**0.1527**	**31.16**	29.0	93.3%