SelecTKD: Selective Token-Weighted Knowledge Distillation for LLMs

Supplementary Material

A. Extended Related Work

Knowledge Distillation for LLMs. Knowledge distillation (KD) [18] compresses LLMs by transferring a strong teacher to a smaller student. Recent advances cluster into two threads: *objectives* and *teaching strategies*. On objectives, MiniLLM [15] leverages reverse KL, AKL [51] adaptively blends forward/reverse KL, DistiLLM [29] introduces skew KL, and DistiLLM-2 [30] extends it with a contrastive, asymmetric design. On strategies, ImitKD [34] mitigates exposure bias via on-policy imitation, PromptKD [28] elicits student-friendly teacher outputs through prompts, and ATKD [62] adjusts teaching modes for easy vs. hard tokens under uncertainty. These works primarily optimize how divergence is measured and how data is curated. Our SelecTKD decides *where* to apply supervision by selecting reliable tokens, and composes with the above losses and pipelines to further boost student performance.

Token Filtering and Selective Distillation. Recent work converges on the view that not all tokens are equally informative. Along the efficiency axis, token dropping in pretraining reduces computation by skipping low-importance tokens [19, 61]. For fine-grained supervision, RHO-1 [35] formalizes selective language modeling by reweighting losses across tokens, while ENT [33] improves robustness by truncating gradients of high-error tokens under noisy supervision. In the fine-tuning and alignment stages, Token Cleaning [43] and T-SHIRT [13] apply token-level data selection that outperforms coarse, sample-level filtering; system-level advances such as Collider [5] further make token filtering cost-effective on modern hardware. In the context of distillation, AdaSPEC [21] performs selective alignment between draft and target models, optimizing acceptance rates instead of full-sequence KL. Our approach is complementary: we adopt a simple verification mechanism and apply a token-weighted loss directly during KD, emphasizing high-confidence teacher signals while preserving compatibility with existing objectives and pipelines.

Orthogonal Directions. Several concurrent works address complementary aspects of LLM distillation. DSKD [58] resolves vocabulary mismatch by mapping teacher and student into a unified output space; EvoKD [37] synthesizes new training data via active learning to target student weaknesses; and Low-Rank Clone [16] achieves parameter-efficient distillation through low-rank weight cloning. These operate on the *space-alignment*, *data-generation*, and *structural* axes, respectively, while SelecTKD operates on the *loss-weighting* axis and can be combined with any of them.

B. Training Efficiency

SelecTKD introduces marginal overhead since verification reuses the teacher logits already computed for the distillation loss, requiring no additional forward passes or reference models. Measured on Qwen2-7B $\rightarrow$ 1.5B (4$\times$A100), relative to DistiLLM-2: Greedy Top-k adds $1.03\times$ wall-clock time and $1.02\times$ peak memory; Spec-k adds $1.08\times$ time and $1.02\times$ memory.

C. Limitations and Future Work

Firstly, our experiments evaluate SelecTKD with LLM teachers of up to 9B parameters and an 8B VLM teacher, scaling up to frontier-scale models remains an open challenge. Additionally, we currently use a fixed Top-k and a global rejected-token weight β; adaptively learning k/β, or conditioning them on token- or context-level uncertainty, may further enhance stability and efficiency. While our study covers single-turn, text-centric tasks and a vision-language scenario, extending selective token-weighted distillation to multi-image/video, speech, or preference-aligned pipelines is promising. Overall, SelecTKD reframes distillation from loss engineering to selective supervision, and we expect this principle to underpin more robust and efficient compact models.

D. Additional Results

As discussed in Section 1, our preliminary experiments suggest that different loss function geometries (e.g., forward/reverse/skewed KL) converge to similar fixed points despite differing optimization dynamics. To further validate this claim and demonstrate its robustness across different evaluation protocols, we conduct comprehensive experiments on multiple evaluation platforms, including DeepSeek, Qwen, OpenAI, and Kimi. These platforms employ distinct judging criteria and model preferences, providing a diverse testbed for assessing the generality of our hypothesis.

As illustrated in Figure 5, the performance differences between symmetric and asymmetric loss combinations remain relatively small across all four evaluation platforms, with variations typically within 1–2% in win rate. This consistency across diverse judging criteria strengthens our argument that the choice of loss function geometry, while affecting training dynamics, does not fundamentally alter the achievable performance ceiling. Instead, the key insight lies in selectively applying supervision to high-confidence teacher signals, which is the core principle of SelecTKD.

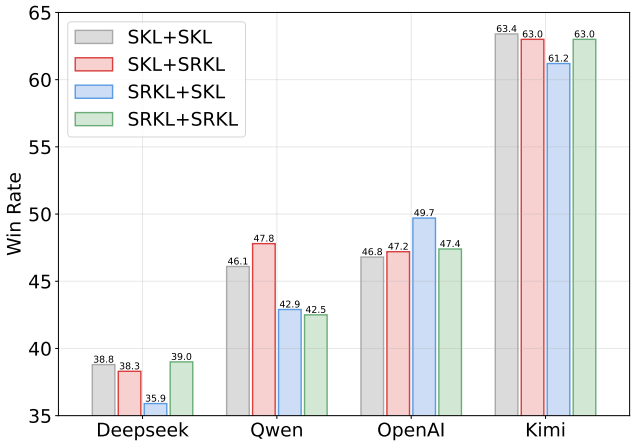

Figure 5. Performance comparison of symmetric and asymmetric loss function combinations on Qwen2-7B-Inst → Qwen2-1.5B-SFT across four different evaluation platforms. The chart shows the Win Rate of four loss function combinations: two symmetric forms (SKL+SKL, SRKL+SRKL) and two asymmetric forms (SKL+SRKL, SRKL+SKL). Results are evaluated using DeepSeek (deepseek-chat), Qwen (qwen-plus-2025-01-25), OpenAI (gpt-4o), and Kimi (moonshot-v1-8k) as judges. The consistent performance patterns across platforms indicate that loss function geometry is not the dominant factor in determining final student performance, supporting our reframing of the distillation problem from "how to measure divergence" to "where to apply supervision."

Table 6 presents the complete ablation study on verification mechanisms and hyperparameters, extending the analysis in Section 5. The results demonstrate that while different verification strategies (Hellinger-based, Greedy Top-k, Non-Greedy Spec-k) yield varying degrees of improvement over Vanilla KD, the Non-Greedy Spec-k variant consistently achieves the best performance. Furthermore, the robustness of SelecTKD to hyperparameter choices (particularly $k \in \{3, 5, 7, 10\}$ and $\beta \in \{0, 0.01, 0.05, 0.1\}$) indicates that the framework is practical and easy to deploy without extensive hyperparameter tuning.

E. Theoretical Analysis of SelecTKD

In this section, we provide a rigorous theoretical analysis of SelecTKD, focusing on the monotonic improvement property of the Token Acceptance Rate (TAR) during training. This analysis establishes a mathematical foundation for the implicit curriculum learning effect observed empirically in our experiments (Section 5). The following theorem provides a lower bound on the TAR increment per update step, which guarantees that the student-teacher alignment improves monotonically under appropriate conditions.

Theorem 2. *Assume a sufficiently small learning rate $\eta > 0$ and that the language model satisfies standard Lipschitz continuity conditions [8]. Then, for both SelecTKD variants (Greedy Top-k and Non-Greedy Spec-k), each gradient update step improves the Token Acceptance Rate (TAR), with the improvement lower-bounded as:*

$$\mathrm{TAR}_{t+1} - \mathrm{TAR}_t \geq \eta\, \kappa\, (1 - \mathrm{TAR}_t), \tag{14}$$

where $\kappa > 0$ is a positive constant that reflects the average ease of correcting a rejected proposal. The constant κ depends on: (i) the teacher's confidence margin at the Top-k decision boundary, (ii) the model's smoothness properties (Lipschitz constant), and (iii) the gradient magnitude for tokens near the acceptance threshold.

Table 6. Comprehensive ablation on verification mechanisms, selection types, and key hyperparameters k and β for SelecTKD. Results are reported as average win rate (%) on instruction-following benchmarks (Qwen2-7B-Inst → Qwen2-1.5B).

Method	Verification	k	β	Avg. Win Rate (%)
Vanilla KD	None	–	1	41.19
SelecTKD (H)	Hellinger-based	–	–	41.70
SelecTKD (G)	Greedy Top-k	3	0.01	42.98
SelecTKD (G)	Greedy Top-k	5	0.01	43.09
SelecTKD (G)	Greedy Top-k	7	0.01	43.04
SelecTKD (G)	Greedy Top-k	10	0.01	42.95
SelecTKD (G)	Greedy Top-k	5	0.1	42.82
SelecTKD (G)	Greedy Top-k	5	0.05	42.86
SelecTKD (G)	Greedy Top-k	5	0	42.27
SelecTKD (NG)	Non-Greedy Spec-k	3	0.01	43.22
SelecTKD (NG)	Non-Greedy Spec-k	5	0.01	**43.33**
SelecTKD (NG)	Non-Greedy Spec-k	7	0.01	43.26
SelecTKD (NG)	Non-Greedy Spec-k	10	0.01	43.13
SelecTKD (NG)	Non-Greedy Spec-k	5	0.1	43.20
SelecTKD (NG)	Non-Greedy Spec-k	5	0.05	43.18
SelecTKD (NG)	Non-Greedy Spec-k	5	0	42.83

Proof. Let c denote a context tuple $(x, y_{<t})$ representing the input and previous tokens. Define the acceptance indicator function at training step t as $A_t(c) = \mathbb{I}\big(V_t(c) = 1\big)$, where V_t is the SelecTKD verification weight. For the Greedy Top-k variant, $V_t(c) = 1$ if and only if $\arg\max_y q_\theta(y|c) \in \mathrm{Top}_k(p(\cdot|c))$; for the Non-Greedy Spec-k variant, $V_t(c) = 1$ if and only if at least one of the k sampled candidates from $q_\theta(\cdot|c)$ passes the speculative acceptance test. The Token Acceptance Rate at step t is defined as $\mathrm{TAR}_t = \mathbb{E}_c[A_t(c)]$, where the expectation is taken over the data distribution.

The change in TAR after one gradient update step is:

$$\mathrm{TAR}_{t+1} - \mathrm{TAR}_t = \mathbb{E}_c[A_{t+1}(c) - A_t(c)]. \tag{15}$$

Under the assumption of a sufficiently small learning rate η and smooth model dynamics, the probability of an accepted token becoming rejected is negligible compared to the reverse transition. This follows from the fact that gradient updates are local and smooth, and accepted tokens are already within the teacher's high-confidence region. Therefore, the TAR increment is dominated by tokens that transition from

rejected to accepted:

$$\begin{aligned}\mathrm{TAR}_{t+1} - \mathrm{TAR}_t &\approx \Pr(A_{t+1}=1,\, A_t=0) \\ &\underbrace{-\Pr(A_{t+1}=0,\, A_t=1)}_{\approx 0} \geq \Pr(A_{t+1}=1,\, A_t=0). \quad (16)\end{aligned}$$

Let $B_t = \{c : A_t(c) = 0\}$ denote the set of rejected contexts at step t. The probability of a context being rejected is $\Pr(c \in B_t) = 1 - \mathrm{TAR}_t$. The improvement in TAR comes from tokens in B_t that become accepted after the update:

$$\begin{aligned}\Pr(A_{t+1}=1,\, A_t=0) &= \Pr(c \in B_t) \cdot \Pr(A_{t+1}=1 \mid c \in B_t) \\ &= (1 - \mathrm{TAR}_t) \cdot \Pr(A_{t+1}=1 \mid c \in B_t). \quad (17)\end{aligned}$$

The conditional probability $\Pr(A_{t+1} = 1 \mid c \in B_t)$ represents the likelihood that a previously rejected context becomes accepted after one update. This probability can be lower-bounded by $\eta\kappa$, where κ depends on: (i) the gradient magnitude for tokens near the Top-k boundary, (ii) the teacher's probability margin between the k-th and $(k+1)$-th ranked tokens, and (iii) the model's Lipschitz smoothness, which ensures that small parameter updates lead to predictable distribution changes. Formally, κ can be expressed as:

$$\kappa = \min_{c \in B_t, y \in \mathrm{Boundary}_k(c)} \left\{ \frac{|\nabla_\theta \log q_\theta(y|c)| \cdot \mathrm{margin}_k(p(\cdot|c))}{L} \right\}, \quad (18)$$

where $\mathrm{Boundary}_k(c)$ denotes tokens near the Top-k decision boundary, $\mathrm{margin}_k(p(\cdot|c))$ is the probability gap between the k-th and $(k+1)$-th ranked tokens under the teacher distribution, and L is the Lipschitz constant. Combining equations (16) and (17), we obtain:

$$\mathrm{TAR}_{t+1} - \mathrm{TAR}_t \geq (1 - \mathrm{TAR}_t) \cdot \eta\kappa, \quad (19)$$

which completes the proof. □

Theorem 2 provides a unified mathematical foundation for the implicit and adaptive curriculum induced by SelecTKD. The lower bound in equation (14) reveals several important properties:

1. **Monotonic improvement:** The theorem guarantees that TAR increases (or at least does not decrease) at each step, ensuring stable training dynamics.
2. **Adaptive curriculum:** The term $(1 - \mathrm{TAR}_t)$ quantifies the proportion of "unlearned" or misaligned tokens. The bound shows that the largest improvements occur early in training when misalignment is high ($\mathrm{TAR}_t \approx 0$), while the improvement rate naturally decreases as the student approaches the teacher ($\mathrm{TAR}_t \to 1$). This creates an automatic, self-paced curriculum that starts with easier tokens and gradually incorporates more challenging ones.
3. **Robustness to hyperparameters:** The constant κ depends on the teacher's confidence margin, which is typically well-separated for high-quality teachers. This ensures that the improvement bound remains meaningful across different model sizes and tasks.

This theoretical result is empirically supported by the TAR curves observed in our experiments (see Figure 4 (a) in the main text), which exhibit the predicted quasi-monotonic increase and saturating behavior. The theorem confirms that SelecTKD prioritizes correcting the student's most significant deviations first, leading to stable optimization and improved generalization, as evidenced by the flatter loss landscapes discussed in Section 5.

F. Gradient Derivations

This section provides detailed derivations of the gradients for various divergence measures used in knowledge distillation. These derivations support the convergence analysis presented in the main text and clarify the relationship between different loss functions. We begin with the forward and reverse KL divergences, then extend to the skewed variants (SKL and SRKL) used in modern distillation frameworks.

F.1. Derivation of FKL Gradient

The forward KL divergence measures the expected log-likelihood ratio under the teacher distribution, encouraging the student to cover all modes of the teacher. Consider the forward KL divergence term at time step t and token v_i:

$$D_{\mathrm{FKL}}^{(t,i)}(p, q_\theta) = p_i \log \frac{p_i}{q_i}, \quad (20)$$

where we use the shorthand notation:

$$p_i := p(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x}), \quad q_i := q_\theta(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x}). \quad (21)$$

Since the teacher distribution p_i is fixed (treated as a constant with respect to the student parameters θ), the derivative with respect to q_i is:

$$\frac{\partial}{\partial q_i} D_{\mathrm{FKL}}^{(t,i)}(p, q_\theta) = -\frac{p_i}{q_i}. \quad (22)$$

The negative sign indicates that increasing q_i reduces the divergence when $p_i > 0$, which aligns with the mass-covering behavior of forward KL. Using the chain rule through the softmax function, the gradient with respect to the student distribution is:

$$\frac{\partial}{\partial q_\theta(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x})} D_{\mathrm{FKL}}^{(t,i)}(p, q_\theta) = -\frac{p(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x})}{q_\theta(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x})}. \quad (23)$$

This gradient is large when p_i is high and q_i is low, emphasizing the importance of learning high-probability teacher tokens.

F.2. Derivation of RKL Gradient

The reverse KL divergence measures the expected log-likelihood ratio under the student distribution, encouraging mode-seeking behavior. For the reverse KL divergence at position (t, i):

$$D_{\text{RKL}}^{(t,i)}(p, q_\theta) = q_i \log \frac{q_i}{p_i}, \tag{24}$$

with the same probability definitions:

$$p_i := p(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x}), \quad q_i := q_\theta(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x}). \tag{25}$$

Differentiating with respect to q_i using the product rule:

$$\frac{\partial}{\partial q_i} D_{\text{RKL}}^{(t,i)}(p, q_\theta) = \frac{\partial}{\partial q_i} [q_i \log q_i - q_i \log p_i] \tag{26}$$

$$= (\log q_i + 1) - \log p_i \tag{27}$$

$$= \log \frac{q_i}{p_i} + 1. \tag{28}$$

The gradient is positive when $q_i > \frac{p_i}{e}$ and negative when $q_i < \frac{p_i}{e}$, pushing the student distribution toward the teacher. The full gradient with respect to the student parameters is:

$$\frac{\partial}{\partial q_\theta(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x})} D_{\text{RKL}}^{(t,i)}(p, q_\theta) = \log \frac{q_\theta(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x})}{p(v_i \mid \boldsymbol{y}_{<t}, \boldsymbol{x})} + 1. \tag{29}$$

Note that the gradient depends on the ratio q_i/p_i, making it sensitive to tokens where the student over- or underestimates the teacher's probability.

F.3. Convergence Analysis of FKL and RKL

We now analyze the convergence properties of forward and reverse KL divergences, showing that both objectives converge to the same fixed point $q_\theta = p$ under appropriate conditions. This analysis supports the claim in the main text that different loss geometries share the same optimal solution, despite differing optimization dynamics.

We follow the notation established in the main text: p denotes the teacher distribution, q_θ the student distribution parameterized by θ, and z_j^q the student logit for token j in the vocabulary $\mathcal{V} = \{1, \ldots, |\mathcal{V}|\}$.

Convergence of Forward KL. The convergence condition for forward KL requires that all gradients with respect to student logits vanish:

$$\frac{\partial D_{\text{FKL}}(p, q_\theta)}{\partial z_j^q} = 0, \quad \forall j \in \{1, \ldots, |\mathcal{V}|\}. \tag{30}$$

Using the standard result for cross-entropy with soft targets, the gradient of forward KL with respect to student logits is:

$$\frac{\partial D_{\text{FKL}}(p, q_\theta)}{\partial z_j^q} = q_\theta(j|\mathbf{y}_{<t}, \boldsymbol{x}) - p(j|\mathbf{y}_{<t}, \boldsymbol{x}). \tag{31}$$

This follows from the fact that forward KL is equivalent to cross-entropy with the teacher distribution as targets. The gradient in equation (31) vanishes if and only if:

$$q_\theta(j|\mathbf{y}_{<t}, \boldsymbol{x}) = p(j|\mathbf{y}_{<t}, \boldsymbol{x}), \quad \forall j \in \{1, \ldots, |\mathcal{V}|\}. \tag{32}$$

Since forward KL is convex in q_θ and the constraint $\sum_j q_\theta(j) = 1$ defines a convex set, the stationary point in equation (32) is the unique global minimum. Therefore, forward KL converges to $q_\theta = p$. □

Convergence of Reverse KL. The convergence condition for reverse KL is:

$$\frac{\partial D_{\text{RKL}}(p, q_\theta)}{\partial z_j^q} = 0, \quad \forall j \in \{1, \ldots, |\mathcal{V}|\}. \tag{33}$$

Let $q_\theta = \text{softmax}(z^q)$ denote the student distribution obtained by applying the softmax function to logits z^q. The gradient of reverse KL with respect to logits can be computed using the chain rule. A compact matrix form is:

$$\frac{\partial D_{\text{RKL}}(p, q_\theta)}{\partial z^q} = J_{\text{softmax}}(z^q) \left(\log q_\theta - \log p + \mathbf{1} \right), \tag{34}$$

where $J_{\text{softmax}}(z^q) = \text{diag}(q_\theta) - q_\theta q_\theta^\top$ is the softmax Jacobian matrix, $[\log q_\theta]_i = \log q_\theta(i)$, $[\log p]_i = \log p(i)$, and $\mathbf{1}$ is the all-ones vector of dimension $|\mathcal{V}|$.

When $q_\theta = p$ (elementwise equality), the term $(\log q_\theta - \log p + \mathbf{1})$ reduces to $\mathbf{1}$. Since the softmax Jacobian satisfies $J_{\text{softmax}}(z^q)\mathbf{1} = \mathbf{0}$ (due to the constraint $\sum_j q_\theta(j) = 1$), the gradient in equation (34) is zero. Therefore, $q_\theta = p$ is a stationary point.

To establish uniqueness, note that reverse KL is strictly convex in q_θ on the interior of the probability simplex. This follows from the fact that the function $f(q) = q \log(q/p)$ is strictly convex for $q > 0$ and $p > 0$. Therefore, the stationary point $q_\theta = p$ is the unique global minimum, and reverse KL converges to this fixed point. □

The convergence analysis above demonstrates that both forward and reverse KL divergences converge to the same fixed point $q_\theta = p$, despite their different optimization dynamics (mass-covering vs. mode-seeking). This result supports the empirical observation in the main text that different loss geometries yield similar final performance when training is sufficiently long.

F.4. Convergence Analysis of SKL and SRKL

We now extend the convergence analysis to the skewed variants (SKL and SRKL) introduced in DistiLLM [29] and used in DistiLLM-2 [30]. These variants interpolate between teacher and student distributions, providing a flexible mechanism to balance mass-covering and mode-seeking behaviors.

Convergence of Skew KL (SKL). The convergence condition for SKL requires:

$$\frac{\partial \mathcal{L}_{\text{SKL}}}{\partial z_j} = 0, \quad \forall j \in \{1, \ldots, |\mathcal{V}|\}. \tag{35}$$

Recall from Section 3 that SKL is defined as:

$$\mathcal{L}_{\text{SKL}} = D_{\text{KL}}(p \| m), \quad \text{where } m = \alpha p + (1 - \alpha) q_\theta, \tag{36}$$

and $\alpha \in [0, 1)$ is the skew coefficient. Since $D_{\mathrm{KL}}(p\|\cdot)$ is strictly convex in its second argument (the mixed distribution m), and m is an affine function of q_θ, the unique minimum of SKL is achieved when $m = p$. Substituting the definition of m:

$$\alpha p + (1-\alpha) q_\theta = p \quad \Longleftrightarrow \quad (1-\alpha)(q_\theta - p) = 0. \tag{37}$$

For $\alpha < 1$, this implies $q_\theta = p$. When $\alpha = 1$, SKL degenerates to a constant (zero) and the condition is trivially satisfied. Therefore, $q_\theta = p$ is the unique stationary point (and global minimum) for SKL, establishing convergence to the teacher distribution. □

Convergence of Skew Reverse KL (SRKL). The convergence condition for SRKL is:

$$\frac{\partial \mathcal{L}_{\mathrm{SRKL}}}{\partial z_j} = 0, \quad \forall j \in \{1, \ldots, |\mathcal{V}|\}. \tag{38}$$

Recall that SRKL is defined as:

$$\mathcal{L}_{\mathrm{SRKL}} = D_{\mathrm{KL}}(q_\theta \| m'), \quad \text{where } m' = (1-\alpha)p + \alpha q_\theta. \tag{39}$$

Since $D_{\mathrm{KL}}(\cdot\|\cdot)$ is minimized when its two arguments are equal, and $D_{\mathrm{KL}}(q_\theta\|m')$ is strictly convex in q_θ on the interior of the simplex, the unique minimum is achieved when $q_\theta = m'$. Substituting the definition of m':

$$q_\theta = (1-\alpha)p + \alpha q_\theta \quad \Longleftrightarrow \quad (1-\alpha)(q_\theta - p) = 0. \tag{40}$$

For $\alpha < 1$, this again implies $q_\theta = p$. When $\alpha = 1$, SRKL reduces to a constant and the condition is trivially satisfied. Therefore, $q_\theta = p$ is the unique stationary point for SRKL, confirming convergence to the teacher distribution. □

The convergence analysis of SKL and SRKL demonstrates that both skewed variants share the same fixed point $q_\theta = p$ as forward and reverse KL, regardless of the skew coefficient $\alpha \in [0, 1)$. This theoretical result supports the empirical findings in the main text that different loss geometries, including symmetric and asymmetric combinations, converge to similar performance levels when training is sufficiently long. The key insight is that while the optimization *path* may differ (e.g., different convergence rates, different intermediate behaviors), the final *destination* remains the same: matching the teacher distribution p.

F.5. Detailed Derivation: Gradient of RKL with Respect to Student Logits

This subsection provides a step-by-step derivation of the gradient of reverse KL divergence with respect to student logits. This derivation complements the convergence analysis above and clarifies the mathematical structure underlying the RKL objective. We assume $q_\theta = \mathrm{softmax}(z^q)$ and treat the teacher distribution p as fixed.

Step 1: Component form of RKL. The reverse KL divergence can be written in component form as:

$$D_{\mathrm{RKL}}(p, q_\theta) = \sum_{i=1}^{|\mathcal{V}|} q_\theta(i)\big(\log q_\theta(i) - \log p(i)\big), \tag{41}$$

where $|\mathcal{V}|$ is the vocabulary size, and the summation is over all tokens in the vocabulary.

Step 2: Gradient with respect to q_θ. To compute the gradient with respect to the student distribution q_θ, we differentiate each term in the summation. Using the identity $\partial(q_i \log q_i)/\partial q_j = \delta_{ij}(1 + \log q_i)$, where δ_{ij} is the Kronecker delta, we obtain:

$$\begin{aligned}\frac{\partial\, D_{\mathrm{RKL}}}{\partial\, q_\theta(j)} &= \frac{\partial}{\partial q_\theta(j)} \sum_{i=1}^{|\mathcal{V}|} \big[q_\theta(i) \log q_\theta(i) - q_\theta(i) \log p(i)\big] \\ &= \log q_\theta(j) - \log p(j) + 1. \end{aligned} \tag{42}$$

In vector form, this can be written as:

$$\frac{\partial\, D_{\mathrm{RKL}}}{\partial\, q_\theta} = \log q_\theta - \log p + \mathbf{1}, \tag{43}$$

where $[\log q_\theta]_i = \log q_\theta(i)$, $[\log p]_i = \log p(i)$, and $\mathbf{1}$ is the all-ones vector of dimension $|\mathcal{V}|$.

Step 3: Chain rule through softmax. To obtain the gradient with respect to logits z^q, we apply the chain rule through the softmax function. The softmax Jacobian matrix is:

$$J_{\mathrm{softmax}}(z^q) = \mathrm{diag}(q_\theta) - q_\theta\, q_\theta^\top, \tag{44}$$

where $\mathrm{diag}(q_\theta)$ is a diagonal matrix with q_θ on the diagonal. The (i, j)-th element of the Jacobian is:

$$[J_{\mathrm{softmax}}(z^q)]_{ij} = \frac{\partial q_\theta(i)}{\partial z^q(j)} = q_\theta(i)\left(\delta_{ij} - q_\theta(j)\right). \tag{45}$$

Applying the chain rule:

$$\frac{\partial\, D_{\mathrm{RKL}}}{\partial\, z^q} = J_{\mathrm{softmax}}(z^q)\, \frac{\partial\, D_{\mathrm{RKL}}}{\partial\, q_\theta} = J_{\mathrm{softmax}}(z^q)\, \Big(\log q_\theta - \log p + \mathbf{1}\Big). \tag{46}$$

Step 4: Stationary point analysis. At the stationary point $q_\theta = p$ (elementwise equality), the term $(\log q_\theta - \log p + \mathbf{1})$ reduces to $\mathbf{1}$, since $\log q_\theta(i) - \log p(i) = 0$ for all i when $q_\theta = p$. The softmax Jacobian satisfies the property $J_{\mathrm{softmax}}(z^q)\,\mathbf{1} = \mathbf{0}$, which follows from the constraint $\sum_{i=1}^{|\mathcal{V}|} q_\theta(i) = 1$ and the fact that:

$$\begin{aligned}\sum_{i=1}^{|\mathcal{V}|} [J_{\mathrm{softmax}}(z^q)]_{ij} &= \sum_{i=1}^{|\mathcal{V}|} q_\theta(i)\big(\delta_{ij} - q_\theta(j)\big) \\ &= q_\theta(j) - q_\theta(j) \sum_{i=1}^{|\mathcal{V}|} q_\theta(i) = 0. \end{aligned} \tag{47}$$

Therefore, when $q_\theta = p$, the gradient in equation (46) is zero, confirming that $q_\theta = p$ is a stationary point. By the strict convexity of $D_{\mathrm{RKL}}(p, q)$ in q on the interior of the probability simplex, this stationary point is unique, establishing $q_\theta = p$ as the unique global minimum.