Impacto da Direção da Divergência KL na Calibração e Entropia Preditiva em Destilação de Conhecimento para LLMs

Victor Pasquini Ribeiro Campos [1], **Isabela Neves Drummond** [1]

[1]Instituto de Matemática e Computação – Universidade Federal de Itajubá

d2023003590@unifei.edu.br, isadrummond@unifei.edu.br

Abstract. *Knowledge Distillation (KD) transfers the predictive distribution of a large teacher model to a smaller student. The choice between Forward KL (FKL) and Reverse KL (RKL) as the distillation loss governs how the student organises its output distribution, yet the impact on predictive entropy and calibration in open-ended generation remains uncharacterised. This work presents a controlled experiment comparing CE baseline, FKL and RKL for distilling Qwen2.5-7B into Qwen2.5-1.5B on Dolly-15k, measuring token-level entropy $H(t)$, Expected Calibration Error (ECE) and Teacher-Student KL divergence. RKL produced the lowest entropy and highest ECE, indicating overconfidence not observed in classification tasks. FKL with $T{=}4$ achieved the best ECE across all conditions, while FKL $T{=}2$ produced the worst ECE within the FKL family, revealing a non-monotonic effect of temperature on calibration. Temperature had no effect on RKL, confirming its theoretical invariance empirically.*

Resumo. *A destilação de conhecimento (KD) transfere a distribuição preditiva de um modelo professor maior para um estudante menor. A escolha entre Forward KL (FKL) e Reverse KL (RKL) como função de perda determina como o estudante organiza sua distribuição de saída, mas o impacto sobre entropia preditiva e calibração em geração aberta permanecia não caracterizado. Este trabalho apresenta um experimento controlado comparando CE baseline, FKL e RKL na destilação de Qwen2.5-7B para Qwen2.5-1.5B no Dolly-15k, medindo entropia token-level $H(t)$, Erro de Calibração Esperado (ECE) e divergência KL Teacher-Student. A RKL produziu menor entropia e maior ECE, indicando overconfidence não observado em tarefas de classificação. FKL com $T{=}4$ obteve o menor ECE entre todas as condições, enquanto FKL com $T{=}2$ produziu o maior ECE da família FKL, revelando efeito não-monotônico da temperatura sobre a calibração. A temperatura não exerceu efeito sobre a RKL, confirmando empiricamente sua invariância teórica.*

1. Introdução

A destilação de conhecimento (*Knowledge Distillation* – KD) é uma técnica que transfere o comportamento de um modelo maior (professor) para um modelo menor (estudante), treinando-o para imitar a distribuição de probabilidade do professor ao invés de apenas os rótulos corretos [Hinton et al. 2015]. Em LLMs (*Large Language Models*), onde o custo computacional de inferência limita o uso em ambientes com recursos restritos, como

computação *mobile* e embarcada, a compressão por destilação viabiliza modelos menores sem abrir mão do conhecimento acumulado pelo professor [Gu et al. 2024].

Em geração autoregressiva, a destilação opera *token* a *token* e a função de perda determina como o estudante organiza sua distribuição preditiva. A *Forward* KL (FKL) minimiza $\mathrm{KL}(p_T \| p_S)$ (*mean-seeking*), forçando cobertura de todos os modos do professor; a *Reverse* KL (RKL) minimiza $\mathrm{KL}(p_S \| p_T)$ (*mode-seeking*), concentrando massa nos modos dominantes [Wu et al. 2025, Jung et al. 2025], com impacto na dinâmica de gradiente em épocas limitadas [Rao et al. 2024]. Modelos treinados apenas com entropia cruzada tendem a ser excessivamente confiantes [Kadavath et al. 2022]; na literatura de classificação, FKL reduz o ECE em relação ao *baseline* [Hosseini and Caragea 2022], a temperatura modula esse efeito de forma não-linear [Stanton et al. 2021], e a calibração do professor impacta diretamente a do estudante [Kim et al. 2025]. Em geração aberta com LLMs, essa relação permanece sem caracterização empírica.

Este trabalho conduz um experimento controlado comparando CE *baseline*, FKL e RKL na destilação de Qwen2.5-7B para Qwen2.5-1.5B no Dolly-15k, com varredura $T \in \{1, 2, 4\}$. São medidas entropia *token-level* $H(t)$, ECE e divergência KL *Teacher–Student*, além de ROUGE-L como métrica auxiliar. O MMLU (*Massive Multitask Language Understanding*) é utilizado como avaliação secundária de acurácia *cross-domain* [Hebbalaguppe et al. 2024].

2. Referencial Teórico

A KD comprime um professor em um estudante menor, explorando a *dark knowledge* nas probabilidades sobre classes incorretas [Hinton et al. 2015]. Na formulação original, o professor produz distribuições suavizadas pelo parâmetro T (Equação 1):

$$p_T^{(T)}(k \mid x) = \frac{\exp(z_k/T)}{\sum_j \exp(z_j/T)}, \tag{1}$$

e a perda combina entropia cruzada e destilação com fator T^2 para compensar o escalonamento do gradiente (Equação 2):

$$\mathcal{L} = \alpha\, \mathcal{L}_{\mathrm{CE}} + (1-\alpha)\, T^2\, \mathcal{L}_{\mathrm{KD}}, \tag{2}$$

com degradação sistemática do ECE para $T \geq 8$ [Stanton et al. 2021]. Em LLMs, a destilação opera *token* a *token* sobre $p_T(\cdot \mid y_{<t}, x)$ [Gu et al. 2024]; a divergência KL (Equação 3) apresenta comportamentos geométricos distintos: FKL é *mean-seeking*, RKL é *mode-seeking* [Wu et al. 2025, Jung et al. 2025]:

$$\mathrm{KL}(P \| Q) = \sum_k P(k) \log \frac{P(k)}{Q(k)}. \tag{3}$$

Calibração é a correspondência entre confiança e frequência de acerto [Guo et al. 2017], quantificada pelo ECE (Equação 4):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{4}$$

onde B_m são os *bins* de confiança. Redes profundas tendem a ser *over-confident* [Kadavath et al. 2022]: FKL reduz o ECE versus CE em classificação [Hosseini and Caragea 2022], professores bem calibrados melhoram a calibração do estudante [Kim et al. 2025], e $T \geq 5$ destrói sistematicamente o ECE [Hebbalaguppe et al. 2024]. O MiniLLM reporta ECE(RKL) $<$ ECE(FKL) em classificação de sentimento, em contraste com o regime de geração aberta examinado neste trabalho [Gu et al. 2024].

3. Metodologia

O experimento compara CE *baseline*, FKL e RKL em destilação de texto aberto, totalizando 7 execuções independentes: 1 CE (sem temperatura variável), 3 FKL e 3 RKL com *sweep* $T \in \{1, 2, 4\}$. Todos os hiperparâmetros permanecem fixos; varia-se exclusivamente a função de perda e o valor de T.

3.1. Modelos e Dados

O professor é o `Qwen/Qwen2.5-7B-Instruct` (quantização 4-bit); o estudante é o `Qwen/Qwen2.5-1.5B-Instruct` (precisão `bfloat16`). Ambos compartilham vocabulário e *template* de *chat*, requisito mandatório para destilação *token-wise* com FKL e RKL [Jung et al. 2025], com razão de parâmetros de $\approx$4,7$\times$ [Gu et al. 2024].

O Databricks Dolly-15k [Gu et al. 2024, Rao et al. 2024], com 15.011 exemplos em 7 categorias, foi escolhido por permitir comparabilidade direta com trabalhos recentes de KD para LLMs. Partição determinística (`seed=42`): 14.000 treino, 1.011 teste; máximo de 512 *tokens*, truncamento à direita (taxa: 6,8%), com os *tokens* do *prompt* mascarados (`label = -100`). O MMLU [Kadavath et al. 2022] é *benchmark* secundário de acurácia de múltipla escolha (14.042 exemplos, 57 domínios).

3.2. Funções de Perda

As três condições são normalizadas por $|y|$. A **CE (baseline)** é dada pela Equação 5:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log p_S(y_t \mid y_{<t}, x). \tag{5}$$

A **FKL** (Equação 6) inclui T^2 para compensar o escalonamento do gradiente [Hinton et al. 2015, Stanton et al. 2021], com $\alpha = 0{,}5$:

$$\mathcal{L}_{\mathrm{FKL}} = \alpha\, \mathcal{L}_{\mathrm{CE}} + (1{-}\alpha)\, T^2\, \frac{1}{|y|} \sum_{t=1}^{|y|} \mathrm{KL}\Big(p_T^{(T)}(\cdot \mid y_{<t}, x) \Big\| p_S^{(T)}(\cdot \mid y_{<t}, x)\Big)\,. \tag{6}$$

A **RKL** (Equação 7) segue [Gu et al. 2024], sem T^2:

$$\mathcal{L}_{\mathrm{RKL}} = \alpha\, \mathcal{L}_{\mathrm{CE}} + (1{-}\alpha)\, \frac{1}{|y|} \sum_{t=1}^{|y|} \mathrm{KL}\Big(p_S(\cdot \mid y_{<t}, x) \Big\| p_T(\cdot \mid y_{<t}, x)\Big)\,. \tag{7}$$

Na RKL, quando a mesma temperatura é aplicada a p_S e p_T, a suavização se cancela na razão da KL; *checkpoints* para $T \in \{1, 2, 4\}$ são numericamente idênticos (Seção 4.1).

3.3. Hiperparâmetros e Métricas

O otimizador é AdamW com agendador cosseno, reduzindo a taxa de aprendizado de $4{,}0 \times 10^{-5}$ a zero ao longo de 3 épocas, com aquecimento linear nos primeiros 10% dos passos; *batch* efetivo de 64, *weight decay* $0{,}05$, *gradient clipping* $1{,}0$, máximo de 512 *tokens* e $\alpha = 0{,}5$.

Três métricas principais são calculadas sobre os tokens da região de resposta. **Entropia token-level** (Equação 8):

$$H(t) = -\sum_{v \in V} p_S(v \mid y_{<t}, x) \log p_S(v \mid y_{<t}, x). \tag{8}$$

Nenhum estudo anterior mediu $H(t)$ temporal em geração autoregressiva com comparação direta FKL vs. RKL [Wu et al. 2025, Gu et al. 2024]. **ECE:** $M{=}10$ *bins equal-width* (Equação 4), antes de *temperature scaling*, calculado exclusivamente sobre o Dolly-15k; o *token* de maior probabilidade é tratado como predição [Hosseini and Caragea 2022]. **KL Teacher–Student** (Equação 9):

$$\overline{\mathrm{KL}}(T \| S) = \frac{1}{|y|} \sum_{t=1}^{|y|} \mathrm{KL}\Big(p_T(\cdot \mid y_{<t}, x) \Big\| p_S(\cdot \mid y_{<t}, x)\Big). \tag{9}$$

Perplexidade (PPL) e ROUGE-L são métricas auxiliares de qualidade.

4. Resultados

4.1. Métricas Agregadas

A Tabela 1 apresenta as métricas sobre o conjunto de teste do Dolly-15k. A RKL produziu resultados idênticos para $T \in \{1, 2, 4\}$, confirmando a invariância teórica; os três *checkpoints* são equivalentes e reportados como uma única condição.

Tabela 1. Métricas sobre o conjunto de teste do Dolly-15k. ECE: 10 *bins equal-width*, pré-*temperature scaling*. MMLU: acurácia em 14.042 questões. Melhor valor em negrito por coluna.

Condição	H(t)	ECE	$\overline{\mathrm{KL}}$	PPL	MMLU	ROUGE-L
CE (baseline)	1,5245	0,0447	0,7878	**6,27**	**0,5691**	**0,306**
FKL $T{=}1$	1,5592	0,0428	**0,4811**	6,45	0,5638	0,195
FKL $T{=}2$	1,3145	0,0909	0,4907	7,57	0,5465	0,198
FKL $T{=}4$	**1,7647**	**0,0382**	0,6943	6,94	0,5429	0,288
RKL	1,2752	0,1025	0,4815	7,13	0,5530	0,183

4.2. Calibração e Entropia Preditiva

O menor ECE do experimento (0,0382) foi obtido pela FKL $T{=}4$, ficando abaixo até do CE *baseline* (0,0447). A RKL atingiu o maior valor (0,1025), aumento de 129% sobre o *baseline*. Dentro da família FKL, $T{=}2$ produziu ECE de 0,0909, mais que o dobro dos valores em $T{=}1$ e $T{=}4$, em comportamento não-monotônico que reproduz o padrão documentado por [Stanton et al. 2021] e [Hebbalaguppe et al. 2024] em classificação. A entropia preditiva separou as condições em dois grupos: FKL $T{=}4$ e CE *baseline* (1,7647 e 1,5245) com entropia mais alta; FKL $T{=}2$ e RKL (1,3145 e 1,2752) com entropia mais baixa. A Figura 1 evidencia que ECE elevado coincide com $H(t)$ reduzido (*overconfidence*) e que a ordenação das condições é estável ao longo da sequência.

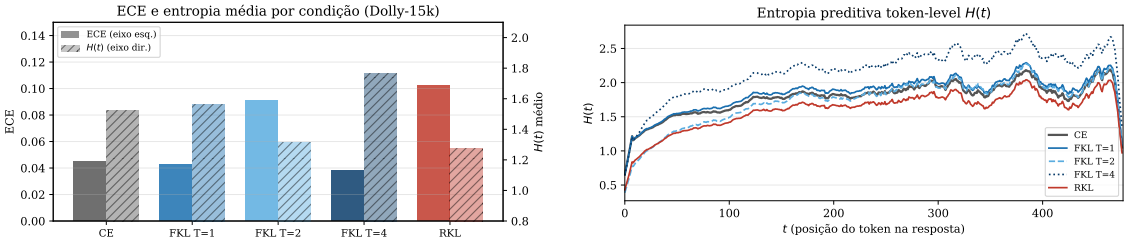

Figura 1. ECE e $H(t)$ por condição (esq.) e $H(t)$ por posição t, janela de 15 tokens (dir.). FKL $T{=}2$ e RKL: ECE elevado com $H(t)$ reduzido (*overconfidence*); separação entre condições estável ao longo da sequência.

4.3. KL Teacher–Student e Discussão

Todas as condições com KD reduziram $\overline{\mathrm{KL}}$ (Equação 9) em relação ao CE *baseline* (0,7878): FKL $T{=}1$, $T{=}2$ e RKL convergiram para valores próximos (0,4811; 0,4907 e 0,4815); FKL $T{=}4$ ficou mais distante (0,6943), com perfis quase sobrepostos após $t{=}60$. Os resultados revelam uma dissociação entre calibração e fidelidade distribucional: as condições que mais reduzem $\overline{\mathrm{KL}}$ não produzem melhor ECE. A RKL combina baixa entropia com alto ECE (*overconfidence*): o modelo concentra massa em poucos *tokens* e erra a magnitude da confiança. O ROUGE-L (Tabela 1) corrobora esse padrão: condições de menor entropia apresentam *recall* alto mas precisão baixa, gerando textos mais longos (≈499–506 vs. 335 *tokens* no CE), consistente com o comportamento *mode-seeking* da RKL.

5. Conclusão

Foram comparadas empiricamente três condições de treinamento para KD em LLMs: CE, FKL e RKL, com *sweep* $T \in \{1, 2, 4\}$ no Dolly-15k. Os resultados revelam uma dissociação entre fidelidade distribucional e calibração: a RKL produziu a menor entropia (1,2752) e o maior ECE (0,1025), indicando que o comportamento *mode-seeking* gera *overconfidence* em geração aberta, em contraste com evidências obtidas em classificação [Gu et al. 2024]. Na família FKL, a calibração mostrou-se não-monotônica em relação à temperatura (FKL $T{=}4$ com ECE de 0,0382; $T{=}2$ com o pior ECE da família, 0,0909), enquanto a invariância teórica da RKL à temperatura foi confirmada empiricamente.

Este trabalho apresenta as seguintes limitações: *(i)* Seeds e intervalos de confiança: cada condição foi executada com seed única, impedindo inferências estatísticas formais. Diferenças pequenas de ECE (e.g., 0,0382 vs. 0,0428) devem ser interpretadas com cautela; *bootstrap* no conjunto de teste poderia estimar variância distribucional, mas não foi implementado; *(ii)* Professor quantizado e limite de transferência: a quantização 4-bit pode tornar as predições do professor mais concentradas (*peaked*), amplificando o *overconfidence* da RKL. O ECE e $H(t)$ do professor não foram coletados, impedindo a determinação do teto teórico de transferência e a separação entre efeito da quantização e efeito da direção da KL; *(iii)* Restrição à família Qwen2.5: a destilação *token-wise* com FKL/RKL exige vocabulário idêntico entre professor e estudante, requisito técnico que motivou a escolha da mesma família e introduz viés de pré-treinamento compartilhado. Testar com LLaMA ou DeepSeek requereria mapeamento de vocabulário ou técnicas alternativas. *(iv)* ECE adaptado, MMLU e corpus: o ECE sobre o *top-1* token não captura incerteza sequencial; os *logits* de inferência no MMLU não foram salvos, de modo que o

ECE nesse domínio não foi calculado; o MMLU serve aqui apenas como *benchmark* de acurácia. O Dolly-15k é gerado por LLMs em 2023, sem revisão sistemática por humanos, limitando a generalização para domínios de raciocínio estruturado. A aplicação de temperatura exclusivamente ao professor na RKL (variante assimétrica) permanece como extensão não testada para trabalhos futuros.

Agradecimentos

Os autores agradecem à Universidade Federal de Itajubá (UNIFEI) pelo apoio à apresentação e publicação deste trabalho, e ao CNPq pelo suporte financeiro por meio de bolsa de pesquisa.

Referências

Gu, Y., Dong, L., Wei, F., and Huang, M. (2024). MiniLLM: Knowledge distillation of large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330.

Hebbalaguppe, R. et al. (2024). Calibration transfer via knowledge distillation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hosseini, M. and Caragea, C. (2022). Calibrating student models for emotion-related tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9266–9278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jung, S., Yoon, S., Kim, D., and Lee, H. (2025). ToDi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8078–8091, Suzhou, China. Association for Computational Linguistics.

Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kim, S., Park, S., Lee, J., and Kwak, N. (2025). The role of teacher calibration in knowledge distillation. *IEEE Access*, 13.

Rao, J. et al. (2024). Exploring and enhancing the transfer of distribution in knowledge distillation for autoregressive language models. *arXiv preprint arXiv:2409.12512*.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. (2021). Does knowledge distillation really work? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 6906–6919.

Wu, T., Tao, C., Wang, J., Yang, R., Zhao, Z., and Wong, N. (2025). Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*.