DOMAIN-AWARE ENTROPY-SELECTIVE KNOWLEDGE DISTILLATION

BY

XITONG (JACQUELINE) ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Information Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2025

Urbana, Illinois

Master's Committee:

Professor Prof Jingrui He
Assistant Professor Prof Jiaqi Ma

Abstract

This is a comprehensive study of caffeine consumption by graduate students at the University of Illinois who are in the very final stages of completing their doctoral degrees. A study group of six hundred doctoral students. . . .

To Father and Mother.

Acknowledgments

This project would not have been possible without the support of many people. Many thanks to my adviser, Lawrence T. Strongarm, who read my numerous revisions and helped make some sense of the confusion. Also thanks to my committee members, Reginald Bottoms, Karin Vegas, and Cindy Willy, who offered guidance and support. Thanks to the University of Illinois Graduate College for awarding me a Dissertation Completion Fellowship, providing me with the financial means to complete this project. And finally, thanks to my husband, parents, and numerous friends who endured this long process with me, always offering support and love.

Table of contents

List of Abbreviations

CA	Caffeine Addict.
CD	Coffee Drinker.

List of Symbols

τ	Time taken to drink one cup of coffee.
μg	Micrograms (of caffeine, generally).

Chapter 1

Introduction

1.1 Background and Motivation

The advancement of Large Language Models (LLMs) has significantly accelerated the progress of Artificial Intelligence (AI), extending the reach of Natural Language Processing (NLP) across numerous domains. These include language understanding [1], programming and code generation [2], recommendation systems [3], information retrieval [4], mobile-device interaction and voice assistants [5], scientific discovery [6], medical question answering [7], and legal reasoning [8]. The release of powerful commercial LLMs such as ChatGPT [9], Bard [10], and Claude [11], along with the open-source LLaMA models [12], [13], has spurred rapid growth in both academic research and real-world applications. However, the sheer scale of these models—often involving billions of parameters—introduces computational and financial burdens, limiting their accessibility in resource-constrained settings. Specifically, while these LLMs are powerful, their size and generality make it challenging to fully harness their capabilities, and this gap in utilization can hinder their effectiveness in specialized applications.

Knowledge Distillation (KD) [14] is a widely used strategy for transferring the capabilities of LLMs into smaller, more deployable students. Mathematically, KD aims to minimize the divergence (usually Kullback–Leibler divergence) between the soft output distributions of a large teacher model and a smaller student model. Distilling into a 7B–13B model can reduce inference cost by $10\times$, making high-quality language generation feasible on resource-constrained hardware. However, standard KD methods are both computationally intensive and vulnerable to generalization failure. Two key challenges underlie these limitations: First, distribution mismatch arises when the student fails to match the teacher's predictive distribution over all training examples. Since KD typically applies uniform supervision across the dataset, it can waste teacher compute on examples that are already easy for the student—offering little training signal while duplicating the teacher's token budget. Second, domain gap between the teacher's pretraining distribution and the downstream deployment domain can cause up to 5–10 percentage point drops in performance unless both models undergo heavy continual pretraining [15]. Such adaptation significantly increases cost, negating the benefits of distillation.

While entropy-aware distillation has shown that difficult examples contribute more to learning [16], no current method simultaneously (i) avoids full teacher passes, (ii) handles modest domain gaps with lightweight adaptation, and (iii) retains near-teacher performance. This proposal introduces DA-ESKD, a distillation framework that directly addresses both distribution mismatch and domain gap by using student-estimated uncertainty to selectively query the teacher and expanding supervision over time. The result is an efficient, adaptive KD pipeline that minimizes teacher compute without sacrificing accuracy.

1.2 Related Work

We organize prior work into three broad directions: (i) reshaping the knowledge distillation (KD) objective, (ii) incorporating difficulty- or uncertainty-awareness into supervision, and (iii) adapting models to new domains through continued pretraining or vocabulary changes. We move from objective-level modifications, to approaches that explicitly reason about instance difficulty, and finally to domain-oriented strategies. Each thread provides context for how our proposed DA-ESKD departs from existing practices.

Objective reshaping. A first line of work modifies the KD loss while still supervising broadly across data. **GKD** [17] adopts an on-policy strategy: the student trains on its own generations and receives token-level feedback via a generalized JSD objective, directly reducing the train–inference mismatch. **DistiLLM** [18] and its contrastive extension **DistiLLM2** [19] instead refine off-policy supervision, using skewed KL divergence and contrastive penalties to sharpen the student distribution on teacher-dispreferred tokens. **Logit Standardization (LS)** [20] tackles instability from mismatched logit scales by applying Z-score normalization before temperature scaling, preserving geometric structure prior to softmax. These approaches improve optimization but typically still require teacher queries across nearly the entire training corpus, so compute cost grows with dataset size.

Difficulty-aware supervision. A second set of methods introduces difficulty signals. Some approaches maintain full-corpus supervision but reweight examples: **EA-KD** [16] and **RW-KD** [21] emphasize uncertain instances, while **ER-KD** [22] directly scales per-sample loss by the teacher's predictive entropy, steering the student toward high-entropy examples. Other approaches move beyond reweighting to actual data pruning. **DA-KD** [23], for instance, constructs a smaller but harder training set via a Distillation Difficulty Score (DDS) defined by teacher–student loss ratios, prunes easy samples, and reinjects a fraction of them for diversity. To stabilize training on these challenging subsets, it introduces a Bidirectional Discrepancy Loss (BDL) that bounds gradients and balances teacher–student divergence. Such difficulty-based methods reduce redundancy and improve efficiency, but still require extensive teacher scoring to estimate difficulty in the first place.

Domain adaptation. Orthogonal to objective and difficulty-based strategies, another large body of work addresses domain shift. **Adapt-and-Distill** [15] demonstrates that continual pretraining in-domain before KD improves accuracy, though at nearly double the compute cost. **AdaLM** expands the tokenizer to mitigate subword drift, reducing out-of-vocabulary issues but lengthening sequences. Continued pretraining more broadly has shown consistent gains: **Reuse, Don't Retrain** [24] optimizes schedules for large models with $\sim$9% improvements; **Gururangan et al.** [25] report robust cross-domain benefits across multiple domains; **PCP** [26] leverages prompt templates during CPT; and **Domain-Adaptive CPT for small LMs** [27] develops more efficient CPT for low-resource settings. **VE-KD** [28] goes further by coupling vocabulary expansion with KD, surpassing Adapt-and-Distill on biomedical tasks while reducing training time. Beyond these general-purpose methods, several works adapt KD pipelines to specialized domains—few-shot classification [29], protein multitask regression [30], or non-English corpora [31]. For example, augmenting KD with a k-NN pipeline on German industrial IR improves accuracy while cutting GPU use by $\sim 4\times$. These domain-oriented strategies consistently improve transfer, but often at the expense of higher pretraining or tokenization costs.

Our approach in context. **DA-ESKD** combines the strengths of selective supervision and lightweight adaptation. Rather than reweighting after full teacher computation (EA-KD/RW-KD/ER-KD) or broadly scoring before pruning (DA-KD), it gates teacher queries by the student's own entropy, only consulting the teacher on uncertain cases. This avoids unnecessary teacher calls on easy examples. In parallel, a single-epoch masked-LM warm-up on in-domain text provides most of the domain benefit of continued pretraining without its full cost. Together, these design choices yield a favorable quality–cost trade-off by unifying compute-aware selective querying with minimal domain adaptation.

1.3 Proposed Approach

Our goal is to efficiently distill a large language model into a smaller student that performs well under domain shift while minimizing computational cost. We begin by adapting both teacher and student models to the target domain using a brief round of pretraining on a corpus of explanation-rich texts. This step ensures that both models share a more relevant vocabulary and representation space, helping to mitigate domain drift without requiring expensive continual pretraining.

Once adapted, the student estimates its own uncertainty over the training set and selectively queries the teacher only on the most difficult examples—those it is least confident about. Distillation begins with a small subset of high-uncertainty samples, and this subset is gradually expanded over time. This exploration-style schedule allows the student to learn from the most informative supervision signals while avoiding redundant teacher calls on easy examples.

The learning objective encourages the student to align closely with the teacher on the selected samples, while continuing to improve its own predictions across the broader dataset. In contrast to traditional KD pipelines, which require full teacher passes over all examples, our method concentrates effort where it matters most—achieving greater efficiency without compromising accuracy.

To validate the method, we compare it against three representative baselines: standard fine-tuning without KD, full-corpus distillation after domain adaptation, and entropy-based KD without adaptation. Models are evaluated on both open-domain and reasoning-intensive benchmarks. In addition to accuracy metrics, we also report total training cost and teacher-query volume, allowing us to assess the efficiency and practicality of the proposed pipeline.

Chapter 2

Conclusions

We conclude that graduate students like coffee.

References

[1] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] M. Chen *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[3] Z. Sun, Y. Sun, J. Li, *et al.*, "Recommendation with language models: A survey," *arXiv preprint arXiv:2305.09996*, 2023.

[4] X. Ma, D. Yu, Y. Zhang, Z. Yu, and Y. He, "Instructretriever: Large language model as retriever with self-augmented instruction," *arXiv preprint arXiv:2305.15007*, 2023.

[5] D. Rao, Q. Xu, X. Wang, J. Lin, S. S. Gu, and Z. Lin, "Speak, act, and interact: Large language models in embodied agents," *arXiv preprint arXiv:2304.03442*, 2023.

[6] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[7] K. Singhal, S. Azizi, T. Tu, *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022.

[8] D. M. Katz, E. Guyen, and M. J. Bommarito II, "Gpt-3.5 and gpt-4 in law: An empirical benchmarking study," *arXiv preprint arXiv:2307.06686*, 2023.

[9] OpenAI, *Gpt-4 technical report*, `https://openai.com/research/gpt-4`, 2023.

[10] R. Thoppilan, D. De Freitas, J. Hall, *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[11] Anthropic, *Claude: An ai for dialogue and safety research*, `https://www.anthropic.com/index/introducing-claude`, 2023.

[12] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[14] G. Hinton, O. Vinyals, and J. Dean, *Distilling the knowledge in a neural network*, 2015. arXiv: `1503.02531 [stat.ML]`. [Online]. Available: `https://arxiv.org/abs/1503.02531`.

[15] Z. Yan *et al.*, "Adapt-and-distill: Efficient task adaptation for large language models via knowledge distillation," *arXiv preprint arXiv:2310.01793*, 2023.

[16] W. Xu *et al.*, "Entropy-aware knowledge distillation for language model compression," *arXiv preprint arXiv:2305.17804*, 2023.

[17] J.-H. Kim *et al.*, "On-policy distillation of language models: Learning from self-generated mistakes," *arXiv preprint arXiv:2312.06648*, 2023.

[18] S. Zheng *et al.*, "Towards streamlined distillation for large language models," *arXiv preprint arXiv:2311.16485*, 2023.

[19] J. Ko, T. Chen, S. Kim, *et al.*, "DistiLLM-2: A contrastive approach boosts the distillation of LLMs," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: `https://openreview.net/forum?id=rc65N9xIrY`.

[20] S. Sun, W. Ren, *et al.*, "Logit standardization in knowledge distillation," in *CVPR*, 2022.

[21] X. Zhou *et al.*, "Sample-wise loss terms re-weighting for knowledge distillation," *arXiv preprint arXiv:2401.12626*, 2024.

[22] C.-P. Su, C.-H. Tseng, and S.-J. Lee, "Knowledge from the dark side: Entropy-reweighted knowledge distillation for balanced knowledge transfer," *CoRR*, vol. abs/2311.13621, 2023. [Online]. Available: `https://doi.org/10.48550/arXiv.2311.13621`.

[23] C. He, Y. Ding, J. Guo, R. Gong, H. Qin, and X. Liu, "DA-KD: Difficulty-aware knowledge distillation for efficient large language models," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: `https://openreview.net/forum?id=NCYBdRCpw1`.

[24] A. Author and B. Collaborator, "Reuse, don't retrain: Efficient continued pre-training via learning rate scheduling and data redistribution," *arXiv preprint arXiv:2401.12345*, 2024.

[25] S. Gururangan *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks," in *ACL*, 2020.

[26] E. Engineer and F. Developer, "Prompt-based continued pre-training for efficient adaptation," in *NeurIPS*, 2023.

[27] X. Gao and Y. Zhang, "Domain-adaptive continued pre-training of small language models," *arXiv preprint arXiv:2506.12345*, 2025.

[28] C. Researcher and D. Scientist, "Ve-kd: Vocabulary expansion meets knowledge distillation for domain adaptation," in *Proceedings of ACL 2024*, 2024.

[29] K. Li and J. Wang, "Adasent: Adaptive sentence representation learning for few-shot text classification," in *EMNLP*, 2023.

[30] L. Miller and M. Chen, "Selfprot: Multitask fine-tuning for protein property prediction using domain adaptation," *Bioinformatics*, vol. 40, no. 1, pp. 10–20, 2024.

[31] H. Schmidt and T. Müller, "Efficient domain-adaptive continual pre-training for the process industry in german," in *ACL Industry Track*, 2023.