Beyond One-Step Distillation: Bridging the Capacity Gap in Small Language Models via Multi-Step Knowledge Transfer*

Gaeun Yim[1,2] **Nayoung Ko**[2] **Manasa Bharadawaj**[3†]
[1]Ulsan National Institute of Science and Technology, Republic of Korea
[2]University of Toronto, Canada
[3]LG Electronics, Toronto AI Lab, Canada
gaeungraceyim@unist.ac.kr nayoung.ko@mail.utoronto.ca manasa.bharadwaj@lge.com

Abstract

Large Language Models (LLMs) excel across diverse NLP tasks but remain too large for efficient on-device deployment. Although knowledge distillation is a promising compression strategy, direct one-step distillation from a large teacher to a small student often leads to substantial performance loss due to the capacity gap. In this work, we revisit multi-step knowledge distillation (MSKD) as an effective remedy, exploring how staged, size-aware transfer paths can better preserve teacher knowledge across students of varying scales. Through extensive experiments with GPT-2 and OPT, we demonstrate that MSKD always improves perplexity and ROUGE-L score over single-step approaches without requiring specialized fine-tuning. Our results establish multi-step transfer as a simple yet powerful framework for progressively compressing LLMs into efficient, high-performing Small Language Models (SLMs).

1 Introduction

Large Language Models (LLMs) (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023) have achieved remarkable success across a wide range of natural language processing tasks, including text generation, question answering, and code completion. Despite recent progress, LLMs still contain billions of parameters, requiring substantial compute and multiple high-end GPUs for both training and inference. Models such as GPT-3 (175B) (Brown et al., 2020) and Llama 3.1 (405B) (Grattafiori et al., 2024) exemplify this scale, making deployment on edge devices or smartphones largely impractical. As a result, on-device adoption of LLMs remains limited, especially under privacy, latency, and connectivity constraints. As organizations seek to embed language capabilities in everyday devices, *compression without compromise* has become a critical research challenge.

Knowledge Distillation (Hinton et al., 2015) has emerged as a popular approach, where a large, high-performing model transfers its knowledge to a smaller, more efficient model. However, when the representational gap between the two models is too large, direct distillation typically leads to severe performance collapse, a phenomenon known as the *curse of capacity gap* (Yang et al., 2022; Zhang et al., 2024; Hsieh et al., 2023; Zhou and Ai, 2024; Lee et al., 2024). Prior studies have attempted to address this issue through data filtering, rationale extraction, and intermediate-task tuning, (Li et al., 2021; Lee et al., 2024; Hsieh et al., 2023; Zhou and Ai, 2024) but these methods introduce heavy computational overheads and task-specific dependencies.

To avoid additional overheads during distillation, we investigate how multi-step distillation can effectively bridge the capacity gap between teacher and student. Our research is inspired by Teacher Assistant Knowledge Distillation (TAKD) (Mirzadeh et al., 2020), which demonstrated the concept of multi stage transfer in Convolutional Neural Networks (CNNs) via intermediate models. While TAKD's insights were originally grounded in computer vision, their potential in Transformer and LLM domains remains largely unexplored, despite its growing relevance today. In parallel, our work builds on the loss function from MiniLLM (Gu et al., 2024), which advanced LLM compression but remained limited to one-step distillation.

In the following sections, we comprehensively analyze MSKD focusing on how the sizes and number of TAs influence overall effectiveness.

2 Related Work

Recent surveys (Yang et al., 2024; Xu et al., 2024) highlight the rapid expansion of research on knowl-

*Work performed during Gaeun Yim and Nayoung Ko were visiting students at University of Toronto and research interns at LG Electronics Toronto AI Lab.

†Corresponding author in the absence of Wei Zhong.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 182–187

edge distillation for Transformer-based models, ranging from early efforts such as DistilBERT (Sanh et al., 2020) to more recent approaches like KARD (Kang et al., 2023) and MiniLLM (Gu et al., 2024). Most of these methods follow a direct distillation paradigm, distilling knowledge directly from teacher to student in a single step without explicitly addressing intermediate models.

Follow-up studies such as Dynamic KD (Li et al., 2021) and Mentor-KD (Lee et al., 2024) revisited intermediate-model designs and incorporated CoT-based supervision (Hsieh et al., 2023; Zhou and Ai, 2024) to enhance student reasoning. Step-by-Step Distillation (Hsieh et al., 2023) generates rationales as auxiliary signals, while TA-in-the-Loop (Zhou and Ai, 2024) filters teacher outputs via an intermediate TA. However, these approaches require multi-stage training or filtering pipelines, resulting in substantial computational overhead. In contrast, we aim to simplify distillation by introducing hierarchical TA models to avoid complex procedures.

Building on the insight from Dynamic KD that mid-sized teachers can transfer knowledge more effectively than very large ones, we extend this idea to enable dynamic knowledge transfer across multiple teacher scales simultaneously. Unlike Dynamic KD's single-teacher setup, our multi-step framework leverages successive teachers for more stable and size-aware distillation.

3 Motivation and Method

The results of multi step distillation from TAKD (Mirzadeh et al., 2020) clearly demonstrate that both the number and scale of Teacher Assistants (TAs) are crucial. When model capacities decrease gradually (e.g., 5x $\rightarrow$ 4x $\rightarrow$ 3x $\rightarrow$ 2x $\rightarrow$ x), the knowledge transfer becomes smoother and yields the best performance. In contrast, skipping steps (e.g., 5x $\rightarrow$ 3x $\rightarrow$ x) results in smaller gains, while direct distillation (e.g., 5x $\rightarrow$ x) performs worst due to large capacity gaps.

Building on this insight, we extend multi-step distillation beyond TAKD's CNN setting to modern Transformer-based LLMs. Unlike convolutional architectures, Transformers show different scaling dynamics, and even our smallest students exceed the largest TAKD teachers in size. This disparity, along with the cost of large-model training, motivates us to explore up to use two TA stages.

Our method is a multi-step distillation framework that progressively transfers knowledge through intermediate Teacher Assistant (TA) models. We hypothesize that the optimal TA lies near the geometric midpoint between teacher and student capacity, balancing gradient stability and knowledge retention. However, too many steps may accumulate information loss, especially in LLMs with long-tailed output distributions. To mitigate this, we adopt the reverse KL objective from MiniLLM (Gu et al., 2024), preventing overfitting to low-probability teacher outputs. This reduces error propagation and ensures that each TA refines, rather than distorts, inherited knowledge.

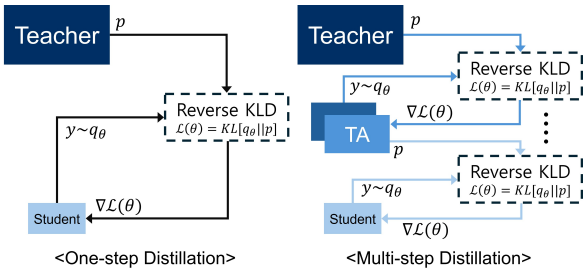

Figure 1: The visualization of our multi-step distillation.

As illustrated in Figure 1, the left panel depicts MiniLLM's (Gu et al., 2024) one-step distillation setup. In contrast, the right panel presents our multi-step distillation framework, which introduces a sequence of TAs between teacher and student. In this framework, $p(y \mid x)$ denotes the teacher model's output distribution, while $y \sim q_\theta(y \mid x)$ represents a response sampled from the student model. By minimizing the reverse KL divergence $\mathrm{KL}(q_\theta \,\|\, p)$, the student updates its parameters so that its generated responses attain higher likelihood under the teacher distribution. This formulation allows the student to benefit from the teacher's distributional feedback without explicitly imitating every sample. It enables incremental, loss-resilient compression that preserves the teacher's expressive fidelity while producing lightweight yet robust students. These distillation paths mirror TAKD's stepwise compression idea, adapted to multi-scale depth of Transformer checkpoints.

4 Experiments and Results

Since MiniLLM offers a strong Transformer-based distillation framework with higher ROUGE-L (Lin, 2004) scores and scalable student sizes for GPT-2 (Radford et al., 2019) and OPT (Zhang et al., 2022), we adopt it as the basis for our multi-step experiments. We use GPT-2 and OPT as both teachers and students, varying only in

	Teacher	TA model		Student	ROUGE-L ↑
MiniLLM	1.5B	-	-	-	27.60
	1.5B	-	-	760M	**26.40**
	1.5B	-	-	340M	25.40
Ours	1.5B	760M	-	340M	26.30
MiniLLM	1.5B	-	-	120M	24.60
Ours	1.5B	760M	-	120M	**27.20**
Ours	1.5B	-	340M	120M	27.00
Ours	1.5B	760M	340M	120M	24.10

Table 1: ROUGE-L score of GPT-2 based multi-step distillation.

	Teacher	TA model		Student	ROUGE-L ↑
MiniLLM	13B	-	-	-	29.20
	13B	-	-	6.7B	29.00
	13B	-	-	2.7B	27.40
Ours	13B	6.7B	-	2.7B	**31.29**
MiniLLM	13B	-	-	1.3B	26.70
Ours	13B	6.7B	-	1.3B	**30.57**
Ours	13B	-	2.7B	1.3B	30.13
Ours	13B	6.7B	2.7B	1.3B	30.46

Table 2: ROUGE-L score of OPT based multi-step distillation.

parameter size while keeping architectures and tokenizers identical. For fairness, all evaluations are performed strictly within the MiniLLM setup. TA and student sizes are chosen from public MiniLLM checkpoints to ensure consistent comparison across experiments[1].

Dataset. We use the databricks-dolly-15K dataset (Conover et al., 2023), comprising 15K human-written instruction–response pairs, and adopt the same split as MiniLLM, with 500 samples for testing and the remainder for training.

Compute resources. All experiments were conducted on NVIDIA V100 16GB GPUs. Distilling GPT-2 120M from the 1.5B teacher finished in under 10 hours on four GPUs, while OPT 2.7B distilled from the 6.7B TA required about 40 hours on the same setup.

Hyper-parameters. We set the sampling temperature to 1.0, the learning rate to 5e-6, and the weight of the distillation loss to 0.5. The model is then trained for 5,000 steps.

Metrics. We use ROUGE-L to assess output similarity and Perplexity to measure language modeling quality, where lower values indicate better performance.

4.1 ROUGE-L score

Table 1 presents ROUGE-L scores of our multi-step distillation experiments, alongside the MiniLLM baseline, which evaluates only single-step distillation using GPT-2 as the base model. The results in Table 1 clearly demonstrate that adding an intermediate TA consistently improves GPT-2 performance.

Our approach of two-step distillation (1.5B → 760M → 340M) improves ROUGE-L from 25.40 to 26.30. Applying two-step distillation to smaller students (1.5B → 760M → 120M) also improves ROUGE-L from 24.60 to 27.20. We can also observe the difference of sizes of TA models causes slightly different improvements, and 760M was slightly better than the smaller TA. However, extending to three-step distillation results in degradation (24.10), suggesting diminishing returns when too many intermediate models accumulate errors. These findings are the same observation with CNNs (Mirzadeh et al., 2020): gradual but not excessive bridging is optimal.

We expand our experiments to language models roughly ten times larger, from GPT-2 (1.5B) to OPT (13B), to examine whether the teacher's size influences the resulting trends. In Table 2, two-step distillation (13B → 6.7B → 2.7B) yields a +3.89-point improvement over MiniLLM's one-step approach, showing consistent trends of gains for both 2.7B and 1.3B students. Notably, the best result occurs when the Teacher Assistant (TA) is much larger than the final student for the two-step distillations, showing that TA quality rather than the number of distillation stages is the primary determinant of final model fidelity, especially at larger scales.

Across both GPT-2 and OPT, performance peaks when the TA is large enough to preserve high-level representations before transfer, underscoring the need for a capacity-aware intermediate model that minimizes information loss. Overall, Table 2 shows the same improvement–degradation trend observed in GPT-2 except, demonstrating a consistent link between model scale and distillation effectiveness and confirming that multi-step distillation remains effective for much larger OPT models as well. In contrast, we find that three-step setup behaves differently, producing a clear performance boost, an effect that did not appear in the comparable multi-step distillation results for GPT-2.

[1]https://huggingface.co/MiniLLM

	Teacher	TA model		Student	Perplexity ↓
MiniLLM	1.5B	-	-	-	21.94
	1.5B	-	-	760M	**15.40**
	1.5B	-	-	340M	24.77
Ours	1.5B	760M	-	340M	17.55
MiniLLM	1.5B	-	-	120M	23.93
Ours	1.5B	760M	-	120M	**21.56**
Ours	1.5B	-	340M	120M	21.76
Ours	1.5B	760M	340M	120M	22.31

Table 3: Perplexity of GPT-2 multi-step distillation.

	Teacher	TA model		Student	Perplexity ↓
MiniLLM	13B	-	-	-	17.83
	13B	-	-	6.7B	11.35
	13B	-	-	2.7B	13.44
Ours	13B	6.7B	-	2.7B	**11.27**
MiniLLM	13B	-	-	1.3B	13.79
Ours	13B	6.7B	-	1.3B	12.52
Ours	13B	-	2.7B	1.3B	12.99
Ours	13B	6.7B	2.7B	1.3B	**11.88**

Table 4: Perplexity of OPT based multi-step distillation.

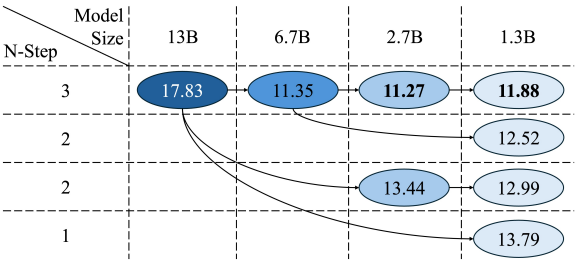

Figure 2: Perplexity of OPT across all possible paths.

4.2 Perplexity

Table 3 and 4 present the results of perplexity, which further reinforce the effectiveness of multi-step knowledge transfer.

In Table 3, two-step distillation, perplexity decreases in proportion to both the TA's size and the student's capacity, mirroring the trends observed in ROUGE-L performance. Notably, the best configuration (1.5B → 760M → 340M) reduces perplexity from 24.77 to 17.55, a substantial improvement approaching the original teacher's predictive confidence 21.94. Unlike the ROUGE-L results (Table 1), the three-step distillation still improves direct distillation from 23.93 to 22.31, but the gains remain smaller than those achieved with two-step distillation. This suggests that overly deep distillation chains may dilute useful signal and introduce noise, limiting the benefits of additional intermediate stages.

Unlike GPT-2, the result of a three-step configuration (13B → 6.7B → 2.7B → 1.3B) in Table 4 attains the lowest perplexity (11.88) among all variants. This suggests the huge findings that the optimal number of steps scales with the teacher–student size gap: larger models benefit from deeper transitions, while mid-scale models converge with two steps. Figure 2 visually represents all possible distillation paths to show the increasing gains proportional to depth as well as TA size. These results support our hypothesis that the ideal distillation depth is dependent on both the architecture and the size ratio between teacher and student models.

5 Key Insights

Our analysis reveals three central findings. Firstly, multi-step distillation outperforms direct distillation without incurring architectural modifications or complex auxiliary supervision. These results suggest that structured progression in capacity space is a more fundamental determinant of effective distillation.

Secondly, the performance of the TA directly predicts the final student's quality. In the GPT-2 experiments, the TA with lower perplexity (760M, 15.40) leads to a superior final student than a weaker TA (340M, 17.55). Similarly, in the OPT family, where the 13B → 6.7B → 2.7B chain achieves the lowest intermediate perplexity (11.27), the subsequent 1.3B student (three step) inherits that advantage. Thus, the TA acts as a bottleneck of knowledge fidelity. The success of MSKD is critically dependent on maintaining a high-quality, low-perplexity intermediate TA at each stage.

Finally, although larger transformer-based models show smaller relative gains compared to their smaller counterparts, their increased parameter count provides greater capacity for TAs, enabling deeper reasoning chains and resulting in better students at the end of these extended chains.

6 Conclusion

We empirically demonstrate that MSKD effectively transfers knowledge from large language models to smaller models, preserving LLM-level performance even on resource-constrained devices. By leveraging intermediate Teacher Assistants, our work shows which distillation paths are effective and how much they outperform standard one-step distillation, improving both ROUGE-L and perplexity without additional data, or complex overheads. This lightweight framework offers a practical pathway for the rapid development of deployment-ready, high-performing small models.

Limitations

While multi-step distillation generally enhances performance, the effectiveness of deeper distillation chains depends on careful selection of intermediate Teacher Assistant sizes and distillation depths. Overly long chains can occasionally lead to diminished gains or increased computational cost, highlighting the need for judicious tuning. Our strong empirical results in favor of Multi-Step Knowledge Distillation motivate future work to establish a deeper theoretical foundation that explains why different multi-step configurations yield varying degrees of improvement, ultimately enabling more principled design of TA sequences. Future work will focus on adaptive TA selection and dynamic step optimization to maintain a favorable balance between mathematically grounded performance improvements and efficient multi-step compression pipelines for next-generation on-device intelligence.

Ethics Statement

In multi-step distillation, identifying an appropriate teacher–assistant (TA) size typically requires multiple rounds of training across different model configurations. This process can incur substantial and infinite computational costs and energy consumption, raising concerns about the efficient use of computing resources. Therefore, further research on principled or mathematically grounded methods for estimating optimal TA sizes and the number of intermediate steps is necessary to reduce unnecessary training overhead. At the same time, once suitable TA configurations are identified, our approach can substantially improve the performance of small language models, which are essential for edge deployment and resource-constrained environments. Such improvements may enable more accessible, energy-efficient, and sustainable AI systems.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program). We sincerely thank Wei Zhong for his supervision and mentorship during the CARTE program, in which this work was conducted, as well as for his consistent guidance and support throughout the project.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36:48573–48602.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-KD: Making small language models better multi-step reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17643–17658, Miami, Florida, USA. Association for Computational Linguistics.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *In Proceedings of Text Summarization Branches Out (ACL 2004)*.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.

Chuanpeng Yang, Wang Lu, Yao Zhu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2024. Survey on knowledge distillation for large language models: Methods, evaluation, and application. *Preprint*, arXiv:2407.01885.

Yi Yang, Chen Zhang, and Dawei Song. 2022. Sparse teachers can be dense with knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3904–3915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Zhang, Yang Yang, Qifran Wang, Jiahao Liu, Jingang Wang, Wei Wu, and Dawei Song. 2024. Minimal distillation schedule for extreme language model compression. *Findings of the ASsociation for Computational Linguistics: EACL 2024*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 265–282, Bangkok, Thailand. Association for Computational Linguistics.