Sequential Distillation of LLMs

Devashish Juyal* juyal@umich.edu
Rohan Singh* rjsingh@umich.edu
Krish Kapoor† krishkap@umich.edu
Erik Nielsen† erikvn@umich.edu

Abstract

The enormous size of frontier models poses significant challenges for deployment in resource-constrained environments. Knowledge distillation (KD) introduced by Hinton et al. [2015] has emerged as a promising approach to compress such models by transferring their capabilities into smaller, more efficient students. In this project, we conduct a comparative empirical study of two distillation paradigms: parallel distillation, in which all student models are directly distilled from a large teacher (e.g., GPT-3), and sequential (chain) distillation, where each student is trained by its predecessor in a cascade.

1 Introduction

Recent advances in LLMs have exponentially expanded their capacity for complex reasoning, particularly through techniques like chain-of-thought (CoT) reasoning. Models like GPT-4 and Claude, now approaching the trillion-parameter scale [OpenAI, 2024], demonstrate impressive few-shot and zero-shot capabilities across muliple domains. However, the computational and memory requirements of these models render them impractical for deployment in resource-constrained settings. To address this scalability challenge, knowledge distillation has emerged as a promising method for compressing LLMs by transferring their knowledge and abilities to smaller, more efficient student models.

While the majority of prior work focuses on single-step distillation, wherein a large teacher model supervises the training of a smaller student, this setup often suffers when the capacity gap between teacher and student is too large. Such gaps can hinder the student's ability to fully absorb the teacher's behavior, especially in complex reasoning tasks. In this project, we compare the effectiveness of two distillation paradigms: (1) **parallel distillation**, where a large teacher model independently supervises multiple students of varying sizes, and (2) **sequential distillation**, where knowledge is passed down through a cascade of models, with each student serving as the next teacher. We evaluate these strategies on reasoning-centric tasks using a Bayesian loss formulation based on KL divergence over soft teacher logits Hinton et al. [2015], Fang et al. [2025], alongside standard cross-entropy objectives. Model performance is assessed across multiple dimensions including final answer accuracy, convergence speed, hallucination rate, and output fidelity (measured via MAUVE Pillutla and et al. [2021] and BERTScore Zhang and et al. [2020]). Through this project, we aim to understand how the structure of the distillation process influences the quality and reliability of distilled LLMs.

1.1 Motivation

At its core, distillation involves a high-capacity teacher model f_t guiding the training of a smaller student model f_s, typically by providing soft predictions that encode representations beyond ground-truth labels. This process mimics the pedagogical relationship between a teacher and student. In this work, we consider a generational propagation of knowledge where a once-trained student becomes a teacher for the next, forming a sequence of increasingly compressed models.

*Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor
†Department of Computer Science, University of Michigan, Ann Arbor

Preprint.

Our goal is to systematically explore how this sequential distillation strategy compares to the more conventional parallel setup. During our literature review, we read a variety of distillation frameworks, including multi-teacher distillation (where a student aggregates knowledge from several teachers), self-distillation (where a model learns from its own softened predictions), and parallel distillation (where a single teacher supervises students of different sizes). While each paradigm offers unique advantages, few studies have investigated the impact of distillation depth, i.e., how knowledge degrades or persists when passed through multiple student-teacher generations. By evaluating the trade-offs between sequential and parallel knowledge transfer, we aim to contribute new insights into how model size, supervision structure, and distillation dynamics interact in the context of reasoning-oriented LLM compression.

1.2 Paper Summary

Our project is inspired by two recent papers. “Cooperative Knowledge Distillation: A Learner-Agnostic Approach" Zhang et al. [2021] introduces a flexible framework where models act as both teachers and students, sharing knowledge based on their strengths and weaknesses. “Keypoint-based Progressive Chain-of-Thought Distillation for LLMs" Feng et al. [2024] presents a progressive strategy for transferring reasoning skills from a single teacher to students of varying sizes. These works motivated us to explore how the structure of knowledge transfer, specifically parallel vs. sequential distillation, affects reasoning performance in compressed LLMs.

2 Extension and Methods

To evaluate these strategies, we will fine-tune students on a range of reasoning tasks involving language understanding, inference, and math problem-solving. For example, questions such as: “What is the total calorie count if you add the calories from lettuce and cucumber, which are 30 and 80 respectively?" require the model to read, reason, and compute. We will use similar multi-step questions to assess reasoning depth and consistency.

Each student will be trained using a combined distillation loss [Fang et al., 2025] that blends KL divergence between the student and teacher logits with standard cross-entropy on the ground-truth or reference outputs. This formulation encourages the student to learn both accurate predictions and faithful approximations of the teacher’s uncertainty.:

$$L = \alpha\, L_{\mathrm{CE}}(y, p_S) + (1-\alpha)\, T^2\, \mathrm{KL}(p_T \,||\, p_S)$$

2.1 Frameworks

We will use PyTorch for the bulk of our training and experimentation pipeline. The dynamic compute graph will allow us to calculate gradients for the cross entropy and KL divergence. In order to load pretrained models for distillation, we will use HuggingFace libraries to load models like GPT-2 and GPT-3 and other open source variants. We can use HuggingFace’s tokenizers and datasets as well. To keep track of loss curves, we will use Weights and Biases, which allows us to analyze the curves in great detail and with a good UI.

2.2 Work Allocation

We intend on breaking our work up into the following responsibilities:

1. Question Generation and Evaluation: It is important to correctly set up experiments that evaluate the reasoning difficulty of each question and generate questions that show off a varying level of reasoning difficulty to test our models on.

2. Training and Fine Tuning: We must make sure hyperparameters like alpha in our loss function are optimally set, which will require trials and hypotheses. We must also run the training of our students on the Great Lakes cluster, which will take some oversight.

3. Result Interpretation and Synthesis: Interpreting our results correctly and drawing conclusions is vital to our study. We must accurately compare the two methods and use data to show the advantages and disadvantages of each approach.

3 Figures and Diagrams

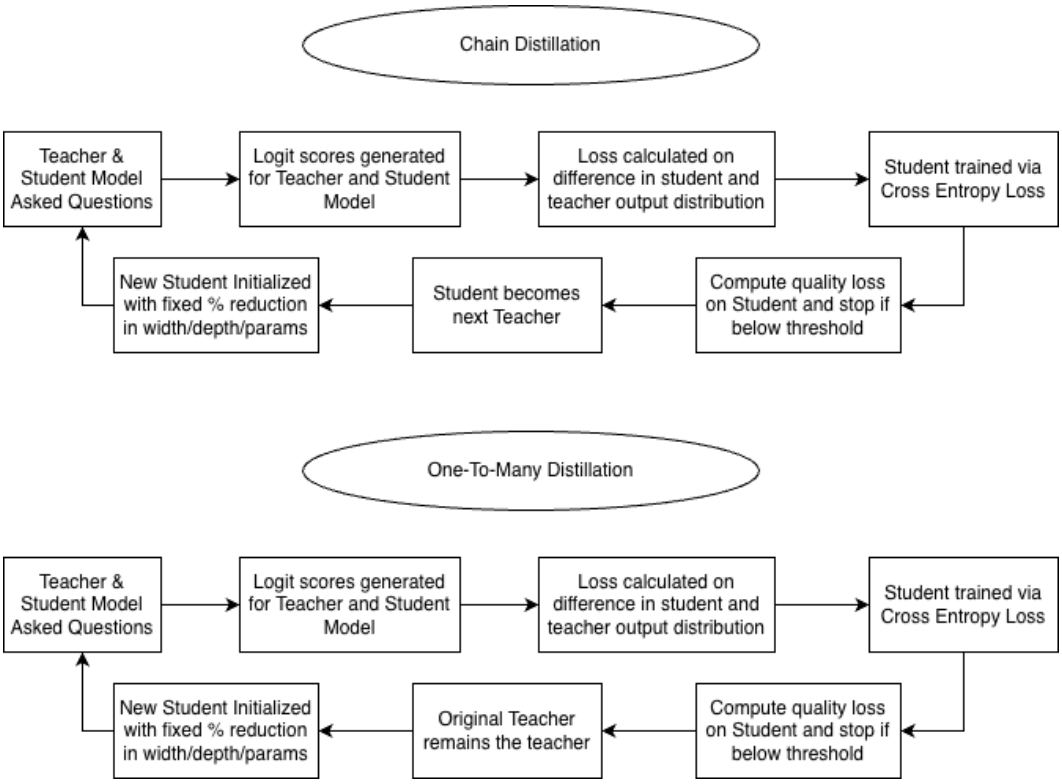

Figure 1: Chain Distillation and One-To-Many Distillation Flow Diagram

References

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL `https://arxiv.org/abs/1503.02531`.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2403.05530*, 2024. URL `https://arxiv.org/abs/2403.05530`.

Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, Terry Ma, Wei Ruan, Ali Abbasi, Jing Zhang, Tao Wang, Ehsan Latif, Wei Liu, Wei Zhang, Soheil Kolouri, Xiaoming Zhai, Dajiang Zhu, Wenxuan Zhong, Tianming Liu, and Ping Ma. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions, 2025. URL `https://arxiv.org/abs/2504.14772`.

Krishna Pillutla and et al. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*, 2021.

Tianyi Zhang and et al. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020.

Yichi Zhang, Yatao Ren, Jing Ma, and et al. Cooperative knowledge distillation: A learner-agnostic approach. In *ICML*, 2021.

Siyang Feng et al. Keypoint-based progressive chain-of-thought distillation for llms. *arXiv preprint arXiv:2405.16064*, 2024.