DistillKit by Arcee AI

Arcee AI has released DistillKit, an open-source toolkit designed to advance research and implementation of model distillation methods for Large Language Models (LLMs).

What is DistillKit?

DistillKit facilitates the creation of cost-effective, secure, and domain-specific Small Language Models (SLMs) – focusing on two primary distillation methods:

- **Logit-based distillation** transfers knowledge from a larger teacher model to a smaller student model, using both hard targets (actual labels) and soft targets (teacher logits). This method has shown the highest performance gains, but requires models to share the same tokenizer.
- **Hidden states-based distillation,** on the other hand, aligns the intermediate layer representations of the student model with those of the teacher model. This approach offers more flexibility, allowing for cross-architecture distillation, such as distilling a Llama-70B based model into a StableLM-2-1.6B.

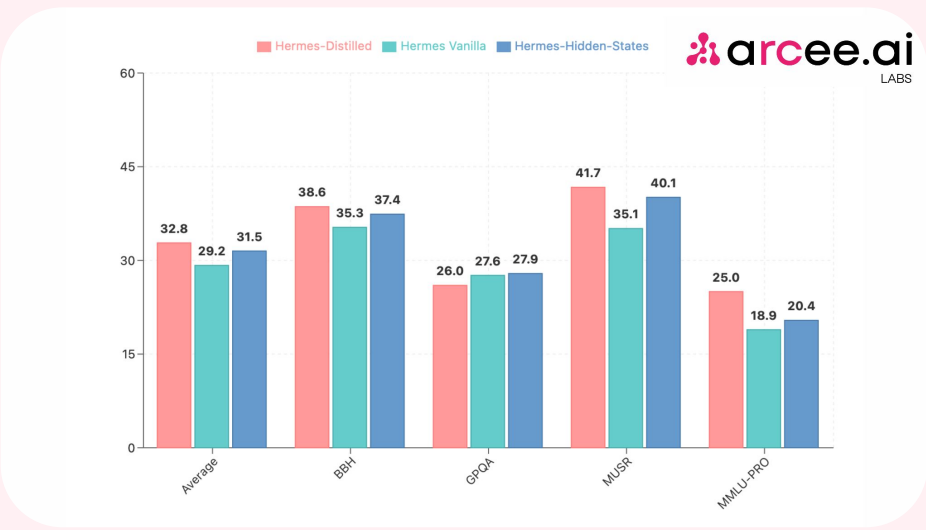
DistillKit's Stellar Evaluations

Evaluations of the distilled models were based on benchmarks from Hugging Face's OpenLLM leaderboard, using the lm-evaluation-harness tool. The results consistently showed improvements over standard Supervised Fine-Tuning (SFT) across most benchmarks.

- DistillKit demonstrates promising results on both general-purpose datasets and domain-specific tasks.
- When evaluated on subsets of datasets like openhermes, WebInstruct-Sub, and FineTome, the distilled models showed encouraging performance improvements – gains in MMLU and MMLU-Pro benchmarks indicated enhanced knowledge absorption capabilities for these SLMs compared to standard SFT alone.

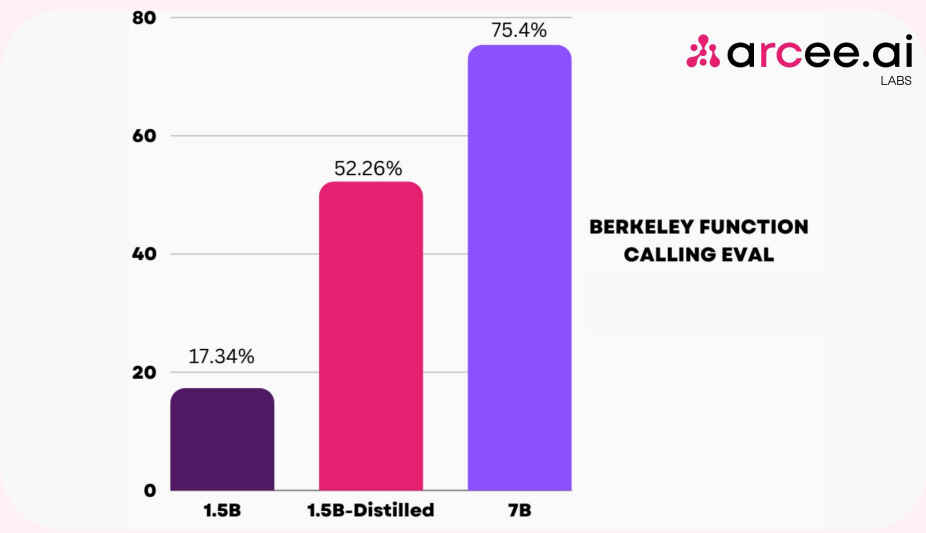
Qwen2-1.5B-Base was trained on a 200,000-sample subset of Teknium's OpenHermes-2.5 dataset. Three variants were produced:

1. Hermes-Distilled: Used logit-based distillation
2. Hermes-Hidden-States: Employed hidden state-based distillation
3. Hermes Vanilla: Utilized the same SFT routine and hyper parameters without any distillation



Qwen2-1.5B-Base was trained on our 450k-sample Agent-Data dataset:

1.1.5B: Standard SFT with no distillation
2.1.5B: Distilled: Employed logit-based distillation with **Arcee-Agent (7B):** as the teacher model
3.7B - Arcee-Agent: A 7B finetune on the same Agent-Data dataset using standard SFT. This was the teacher model



Qwen2-1.5B-Instruct (without additional training) is compared to a distilled version trained for 3 epochs on a 250,000-sample subset of TigerLabs' WebInstruct-Sub dataset.

Qwen2-1.5B-Instruct	52.34%
1.5B-Instruct-Distilled	55.52%

MMLU

arcee.ai LABS