Model Merging

Leveraging OS models to train the most performant & efficient domain-specific language models

arcee.ai

Table of Contents

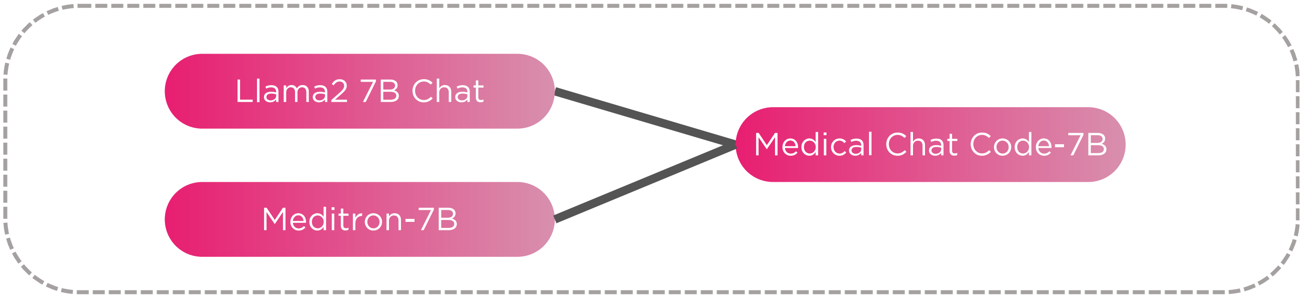
I. WHAT IS MODEL MERGING?

Model Merging is an approach that involves combining two or more neural network models into a superior, unified model – retaining the strengths or qualities of each.

To achieve this, you can use our open source toolkit, **MergeKit**. By loading only the tensors necessary for each individual operation into working memory, MergeKit can scale from a high-end research cluster all the way down to a personal laptop with **no GPU and limited RAM.**

This process involves **no training (so no GPUs)**. It also enhances the performance and accuracy of LLMs, while significantly reducing the time and resources required for training from scratch.

Unlike traditional model ensembling methods, which require the use of multiple models during inference time, Model Merging is a more efficient approach. It yields a single model that **maintains the same size** as each of the individual input models, as illustrated in the figure below.



THE EVOLUTION OF MODEL MERGING & THE RISE OF MERGEKIT

2023

Early in 2023, an innovative technique in the field of language model training, known as **Model Merging**, was first documented in several academic papers. This SOTA approach involves the fusion of two or more LLMs into a singular, cohesive model – presenting a novel and experimental method for creating sophisticated models at a fraction of the cost, without the need for heavy training and GPU resources.

Just a few months later, an open source library named **MergeKit** emerged and quickly gained popularity among LLM developers. Created by **Charles Goddard**, an award-winning software engineer known for his pioneering work at NASA, MergeKit is positioned as a toolkit for the fusion of pre-trained language models. It adopts an out-of-core approach, allowing for intricate merges even in resource-constrained environments. Notably, MergeKit enables merges to be executed solely on the CPU or, if desired, can be accelerated with as little as 8 GB of VRAM.

2024

Mergekit

In early 2024, Charles Goddard officially joined the Arcee AI team, bringing with him the MergeKit library – as **MergeKit joined forces with Arcee AI to forge a powerful collaboration.**

Arcee AI is unwavering in our commitment to maintaining the MergeKit library as an open-source powerhouse, as we solidify its position as the best library in the world for merging models. All MergeKit functionality will remain open source for use by the greater AI community as we work to extend the boundaries of general model capabilities.

II.WHY MODEL MERGING IS ESSENTIAL

Enhanced Performance and Synergy:

Model Merging effectively combines the unique strengths of multiple neural networks into one robust model. This not only improves performance on similar tasks, it also enables better out-of-domain generalization, and leads to synergistic effects – enhancing task performance beyond the capabilities of the individual input models.

When merging models trained on the **same tasks**	Better performance Better out-of-domain generalization
When merging models trained on **different tasks**	Synergistic effects can actually boost task performance above input level

Cost-Effective and Efficient:

Training large models from scratch demands enormous compute and time. MergeKit allows developers to combine already-trained existing models, drastically cutting down the training time and compute costs. This makes advanced models accessible even to those with limited resources.

Mitigates Catastrophic Forgetting:

Model Merging effectively balances general and domain-specific knowledge. Training on your specific data and merging back some “brains” preserves essential information. This process allows the weights in the foundational general model to remain frozen, ensuring that previously-acquired knowledge is not lost.

Accelerated Transfer Learning:

Transfer learning is about adapting pre-trained models to new tasks with minimal additional training. MergeKit enhances this by merging models that are already pre-trained and/or fine-tuned for various tasks, creating a highly-adaptable base model that can be fine-tuned further for specific applications with ease.

How MergeKit is Transforming LLM Training & Transfer Learning

- **Democratizing AI Development:** MergeKit empowers developers, researchers, and organizations of all sizes to build state-of-the-art models without the need for massive infrastructure and compute. By lowering the barriers to entry, it fosters innovation and accelerates the development of AI applications across industries.

- **Expediting Research and Innovation:** Researchers can now experiment with multiple model architectures and techniques more efficiently. MergeKit facilitates rapid prototyping and iteration, enabling quicker discoveries and advancements in research.

- **Industry-Specific Applications**: From finance and healthcare to legal and beyond, MergeKit enables the creation of domain-specific models with unparalleled precision. By merging models trained on diverse datasets, it ensures that the resulting model is highly specialized and effective for industry-specific applications.

Broad Applicability:

The techniques and advantages of Model Merging apply to all types of neural networks, not just Large Language Models (LLMs). This versatility allows for the customization of models to meet specific needs using open-source components, enabling a wide range of applications in various fields.

III. MODEL MERGING WITH ARCEE AI

Arcee AI offers a platform that makes both standard Model Merging and Metric-Guided or “Evolutionary” Model Merging easily accesible to users of all technical backgrounds.

Base Model Merging with the Arcee AI platform

Base Model Merging involves combining two or more pre-trained models to create a more robust and capable model.

You start by by defining a config YAML file with the models and parameters associated with the merge.

The Arcee AI platform streamlines this process with:
A simple interface: An easy-to-navigate UI that allows users to define their YAML file or choose the base default with the chosen merge method, and then perform the merge with just a few clicks.

Compatibility checks: Automatic verification to ensure the models being merged are compatible, reducing the risk of errors.

Customizable parameters: Options to tweak the merging process according to specific needs, ensuring optimal performance.

Metric-Guided/Evolutionary Model Merging with Arcee AI

Metric-Guided Merging, also known as Evolutionary Model Merging, is a technique that uses evaluation metrics to guide the merging process of multiple models. You provide the model with a list of **evals** for which you want to optimize the merge; then the model does an **eval > merge > eval > merge** over and over until the most optimal merging parameters are recognized by the model. Think of it as base merging on steroids.

Traditionally, Model Merging has been a manual and exploratory process, requiring numerous trial-and-error attempts and manual evaluations to determine how merging parameters influence the final model's performance.

However, if you're starting with measurable qualities or competencies you want your model to have, you can instead launch an Evolutionary Model Merge. With this process, you select the models to merge and the merging technique, and the algorithm efficiently finds the ideal combination of parameters. This method significantly streamlines the merging process by eliminating the need for individuals to manually define merge configurations, allowing the optimization algorithms to focus on achieving the targeted attributes of the model.

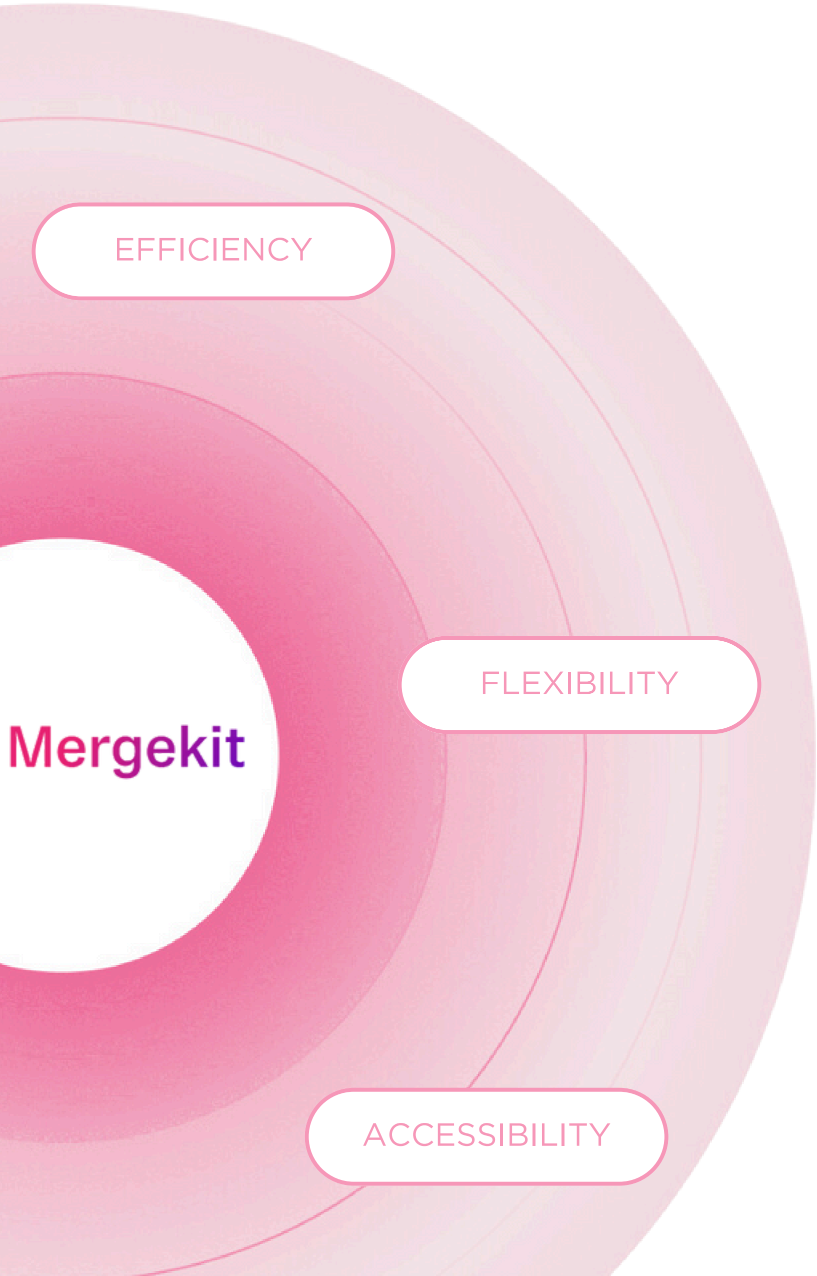
Important note about Evolutionary Model Merging:
this Metric-Guided Merging method produces the best possible model, but it does require more compute than base merging, which can be done on CPUs. That said, **Evolutionary Merging is still extremely cost-efficient compared to traditional full training of models.**

The Arcee AI platform facilitates Evolutionary Model Merging with:

Eval selection: Allowing users to choose relevant eval metrics that will guide the merging process and choose the most optimal merge.

Automated evolution: Using evolutionary algorithms to iteratively merge models and evaluate their performance, automatically selecting the best combinations.

Visualization: Providing graphical representations of the merging process and performance improvements, helping users to understand and analyze the results.



EFFICIENCY

Reduce the time and effort required to merge models, allowing engineers to focus on other critical tasks

FLEXIBILITY

Cater to a wide range of use cases, from simple model enhancements to complex, metric-driven optimizations

ACCESSIBILITY

Make advanced Model Merging techniques available to users without deep technical expertise and lacking GPUs

IV. CUSTOM TRAINING & MODEL MERGING IN ACTION

Expanding on the foundational concepts introduced, we now explore the practical advantages of Model Merging powered by Arcee AI. We provide users the ability to train your own custom model tailored to your specific needs through our training APIs, and then merge it with other high-performing models through MergeKit. This dual approach amplifies the strengths of both models, resulting in an exceptionally powerful and versatile final model.

CUSTOM TRAINING WITH KNOWLEDGE INJECTION

Start by selecting an open-source foundational model such as Mistral or Llama, which serves as our checkpoint for conducting **Continual Pre-Training (CPT**). Leveraging a checkpoint allows us to utilize open-source intelligence, bypassing the need to build a model from scratch. Following this, we proceed with Domain Knowledge Injection, also referred to as Continual Pre-training (CPT). Training a model on your specific dataset is vital as it equips the model with the nuances and details relevant to your particular use case and task. This model becomes highly specialized, capturing domain-specific knowledge that is crucial for its intended use.

Typically, knowledge injection involves embedding extensive domain-specific data into the pre-trained Large Language Model (LLM). It traditionally demands adjusting a massive set of parameters–often scaling into the billions–to effectively integrate vast amounts of domain-specific data.

This complexity is precisely why techniques like LoRA (Low-Rank Adapters) are unsuitable for this type of application. Although LoRA excels in tasks that require minimal addition of new knowledge to the LLM —such as **instruction-tuning** and preference alignment methods like **Direct Preference Optimization (DPO)** — it falls short in the context of CPT, where the objective is to incorporate substantial new knowledge.

SPECTRUM: ENHANCING CONTINUAL PRE-TRAINING EFFICIENCY

At Arcee.ai, our pioneering of Model Merging marked our first major leap forward in creating powerful and efficient models. But our innovations didn’t stop there. We have also built another incredible SOTA technique into the Continual Pre-Training (CPT) step of our platform: it’s called “Spectrum.”

Spectrum improves our training efficiency by 30-42% while ensuring effective knowledge injection. This method involves selectively training specific layer modules based on their signal-to-noise ratio (SNR) and keeping approximately 50% of the model's weights frozen. By doing so, we minimize interference with the model's innate pretraining data, allowing us to inject new domain and business-specific knowledge without compromising the pre-existing general knowledge.

The Spectrum technique allows us to achieve efficient training by:

Selective Layer Training: Targeting layers with the highest potential for learning new information based on their SNR, we focus computational resources where they are most effective.

Maintaining Pre-Trained Knowledge: By keeping a significant portion of the model's weights frozen, we retain the valuable pretraining data, ensuring that the model's general capabilities remain intact.

Efficient Resource Utilization: With Spectrum, we optimize memory and computational usage, leading to faster training times and reduced costs without sacrificing performance.

Minimal Catastrophic Forgetting: The approach helps in preserving previously learned knowledge, reducing the risk of catastrophic forgetting during the training process.

MODEL MERGING MADE EASY

Next, we utilize **MergeKit** to combine the custom-trained model with another pre-trained model that excels in different areas. By employing merging techniques such as Linear, SLERP, TIES, and DARE, we integrate the unique strengths and expertise of both models. This process creates a hybrid that is more capable and versatile than either model alone, leveraging the distinct advantages of each to enhance overall performance and versatility. Our end-to-end platform allows you to do all these steps via a few clicks, in your browser (Arcee Cloud) or in your VPC (Arcee Enterprise)

Benefits of Our Method

- **Domain-Specific Data Utilization**: By employing CPT, we can incorporate proprietary client data, ensuring models are finely-tuned to specific requirements.

- **Efficiency in Model Development**: Utilizing existing models accelerates development, avoiding the need for complex and expensive models.

- **Cost Effectiveness**: Through Model Merging, our approach combines the specialized expertise of custom-trained models with the other pre-trained model, ensuring cost-effective and high-performance language model development.

REAL WORLD IMPACT

Imagine a healthcare system where diagnostic models are continually enhanced by being merged with the latest research models – leading to faster and more accurate diagnoses. Envision financial models that adapt in real-time to market changes by seamlessly merging with new predictive models. Picture educational tools that evolve rapidly, providing personalized learning experiences by merging models tailored to individual learning styles.

Alternatively, envision a world where you can train a small 1B parameter model on your specialized healthcare or finance data, and then seamlessly merge that model with the capabilities of a larger model like LLaMA 70B or Qwen 72B. This approach allows you to achieve a highly powerful, domain-specific model at a fraction of the cost compared to training the entire 70B parameters from scratch. The efficiency and effectiveness of this training and merging process are unmatched, delivering exceptional performance with minimal resources.

By making Model Merging so accessible, Arcee AI has begun to revolutionize the world of LLM training and transfer learning – paving the way for more efficient, and powerful GenAI solutions. By combining the advantages of Model Merging with custom training, Arcee AI is unlocking new potentialing and driving the next wave of AI innovation.

CUSTOMER SUCCESS STORIES

By utilizing the Arcee AI platform and MergeKit, many organizations have successfully merged open-source checkpoints and refined their domain-specific models – achieving significant performance improvements and operational efficiencies.

Here are a few examples of how our customers have benefited:

A Fortune 500 financial services customer uses their model trained with Model Merging and CPT to **provide expert tax advice**. In just their first iteration with Arcee AI, they saw ranking on internal benchmarks jump 23% and deployment costs reduced by 96%.

A top five global P&C insurance customer has been able to boost model performance by 83% while cutting deployment costs by 89%

Guild uses Model Merging and CPT to provide their users with a **dynamic onboarding experience** with their domain-adapted SLM in the loop.

Trusted By

intel aws PIMCO THOMSON REUTERS

activeloop CarbonArc ALEE angel

AngelList ZILO™ workhuman* Guild A

V. COMMON QUESTIONS

Is Model Merging the solution for me?
Merging could be highly beneficial if you're looking to enhance the performance of your models without the extensive resources typically required for training new models from scratch. It's particularly useful if you have models that are effective on their own but could perform even better if combined.

Can I use my own pre-trained model(s) with Arcee AI?
Yes, you can merge your in-house developed models using Arcee AI's platform, which accommodates a variety of model formats and merging techniques.

How does Arcee AI guarantee the security of models throughout the training and Model Merging processes?
With Arcee Enterprise, the training and ongoing enhancement of your custom language models happens fully within your VPC, ensuring the highest end-to-end security.

How does your pricing work?
Arcee has different tiers for API calls in production with a sliding scale for consumption and support tailored to your use case and needs. Please contact sales@arcee.ai for more details.

arcee.ai