Beyond Finetuning:
A Billion 1M Models Can Do More Than a Million 1B Models

Dan Jacobellis and Neeraja J. Yadwadkar
University of Texas at Austin

Overview Major cloud providers now provide "model-less" platforms, such as AzureML [2], SageMaker [1], and VertexAI [3], that streamline the ML development workflow. These platforms perform much of the heavy lifting required for inference by automatically calibrating models to the target hardware. However, the streamlined process afforded by these "model-less" abilities does not extend to training, with providers only offering a handful of heuristics-based AutoML tools. As a result, it remains prohibitively difficult and expensive to develop ML models for many niche applications, including biomedicine, remote sensing, and robotics [5, 15]. General-purpose foundation models and the pretrain-finetune paradigm [12] offer a powerful solution, but at the cost of opaque models with high inference costs [11]; for example, inference of a fine-tuned GPT-4 model is 4.5 times more expensive than the base model [16]. To meaningfully aid the model development and training process, we should leave behind finetuning and the baggage of billion-parameter foundation models and embrace alternative training strategies that manifest the full benefits of specialization—better performance, smaller models, and more efficient inference.

The need for accessible training Despite the rapid adoption of ML across industries and disciplines, many civil, medical, and scientific applications have yet to benefit from learning-based approaches for three main reasons [15].

1. When serious outcomes are at stake, transparency, explainability, and generalizability become much more important than test-set accuracy; billion parameters models are often a nonstarter.
2. Data are often collected with proprietary sensors, in formats scrutible only to domain-experts, and encumbered by various privacy restrictions.
3. Enlisting teams of ML engineers and securing sufficient compute to build and maintain custom models is not feasible under tight budgets or dynamic operations.

The limitations of finetuning The impressive performance of foundation models like GPT are the result of semi-supervised learning (SSL)— training strategies like masked prediction [8] that distill petabytes of unlabeled web data into a general-purpose predictive model. Parameter efficient finetuning (PEFT) techniques [10], such as LoRA [13] and IA3 [14], allow exposure to user-provided data with minimal extra training. Although it costs millions of dollars to train a single foundation model via SSL, platforms like AutoTrain [9] and Modal [4] automatically apply PEFT, allowing users to repurpose one model into many. However, this approach to adaptation results in models that are necessarily larger and more computationally expensive than their general-purpose parents, since it requires freezing the vast majority of weights and adding additional trainable parameters. Finetuning is also prone to catastrophic forgetting, leading some pioneers of the technique to abandon it altogether [11].

Scaling (down) foundation models Without finetuning, how can we impart a model with knowledge beyond the primary training data? In the fine-tuning paradigm, the pre-trained model encapsulates the knowledge of the large-scale dataset it was trained on. However, techniques exist to distill this knowledge into a new, tiny dataset [19] which can be efficiently trained on with a smaller model. Another challenge is replicating the incredible capabilities and generalizability that emerge as a result of performing SSL at scale. Novel SSL strategies are one possible solution. For example, utilizing compression-based tasks for pretraining—reconstructing the input from a learned, low-dimensional representation—has shown promising results with modest dataset and model sizes [21, 6, 7, 18, 17, 20]. Tightly integrating these techniques into our ML platforms to cultivate an ecosystem of leaner, specialized models will allow more applications to benefit from learning-based approaches.

References

[1] Amazon SageMaker. https://aws.amazon.com/sagemaker/, 2018.

[2] Azure Machine Learning. https://docs.microsoft.com/en-us/azure/machine-learning/, 2018.

[3] Vertex AI. https://cloud.google.com/vertex-ai, 2021.

[4] Modal AI. Fine-tune an llm in minutes. https://modal.com/docs/examples/llm-finetuning, 2023.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Hugging Face. Hugging face autotrain. https://huggingface.co/autotrain, 2023.

[10] Hugging Face. Peft: Parameter-efficient fine-tuning of billion-scale models on low-resource hardware. https://huggingface.co/blog/peft, 2023.

[11] Jeremy Howard. The end of finetuning. https://www.latent.space/p/fastai, 2023.

[12] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[14] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[15] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, et al. Artificial intelligence index report 2023. *arXiv preprint arXiv:2310.03715*, 2023.

[16] OpenAI. Gpt-4 finetuning. https://openai.com/gpt-4-ft-experimental-pricing, 2023.

[17] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[18] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[19] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[20] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[21] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.