THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME XXV STLR.ORG SPRING 2024

FOCUSING ON FINE-TUNING: UNDERSTANDING THE FOUR PATHWAYS FOR SHAPING GENERATIVE AI

Paul Ohm[*]

Those who design and deploy generative AI models, such as Large Language Models like GPT-4 or image diffusion models like Stable Diffusion, can shape model behavior in four distinct stages: pretraining, fine-tuning, in-context learning, and input and output filtering. The four stages differ among many dimensions, including cost, access, and persistence of change. Pretraining is always very expensive and in-context learning is nearly costless. Pretraining and fine-tuning change the model in a more persistent manner, while in-context learning and filters make less durable alterations. These are but two of many such distinctions reviewed in this Essay.

Legal scholars, policymakers, and judges need to understand the differences between the four stages as they try to shape and direct what these models do. Although legal and policy interventions can (and probably will) occur during all four stages, many will best be directed at the fine-tuning stage. Fine-tuning will often represent the best balance between power, precision, and disruption of the approaches.

[*] Professor of Law, Georgetown University Law Center. Thanks to the Columbia Science & Technology Review for inviting me to take part in this symposium and in particular to Shyam Balganesh, Penina Cohen, and Emily Davidson. For useful feedback on this Essay, thanks to Zartosht Ahlers, Chinmayi Arun, Shyam Balganesh, Jack Balkin, Stefan Bechtold, Julie Cohen, Feder Cooper, James Grimmelmann, Dan Klerman, Filippo Lancieri, Alicia Solow-Niederman, Holger Spamann, Ayelet Gordon-Tapiero, Danny Wilf-Townsend, Tim Wu, Christopher Yoo, and Angela Zhang and workshop participants at ETH Zurich, the Silicon Flatirons Center at the University of Colorado Law School, Yale Internet Society Project, Georgetown Law, and the Technology Law Scholars program in the Institute for Technology, Law, and Policy, at Georgetown.

I. INTRODUCTION

How do humans control, shape, influence, or steer the outputs of generative AI models such as text-producing large language models (LLMs) or image-generating diffusion models? What techniques do engineers use to shape or change the way these models perform? What can those outside the companies that create these models (these are almost all produced by companies) ask, nudge, or command these engineers to do to reshape these models?

The answers to these questions are central as we respond to the impact these powerful technologies will have on society and the economy. We can shape these models to make them more useful: chatbots that are more polite and helpful or image generation tools that create more aesthetically pleasing art. We can also shape these models to make them less harmful: less likely to defame, hallucinate, train terrorists, or infringe copyright.

The main contribution of this Essay is to help legal scholars and policymakers understand that engineers and users shape the outputs of generative AI models in three distinct stages: *pretraining*, *fine-tuning*, and *in-context learning*.[1]

[1] Rishi Bommasani et al., ARXIV, *On the Opportunities and Risks of Foundation Models* 4-5 (2021), https://perma.cc/V24A-GPPN; JESSICA JI, JOSH A. GOLDSTEIN & ANDREW J. LOHN, ISSUE BRIEF: CONTROLLING LARGE LANGUAGE MODEL OUTPUTS: A PRIMER 4-5, Center for Security and Emerging Technology (2023), https://perma.cc/J776-B7WY; Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*,

During pretraining, engineers train immense neural networks on massive datasets using hundreds of millions of dollars (or more) of compute, producing what is called a *base model* or a *foundation model*.[2] Although these base models are capable of extraordinary feats of computation and representation—such as the capability of producing human-like text fluently—without more they are often unacceptable or unusable for many applications, for example because they are prone to producing texts with racist sentiment. Engineers then fine-tune these models to address these shortcomings, producing *fine-tuned models* that can be deployed for use by end users.

Once deployed, many fine-tuned models accept prompts from users, permitting what is known as in-context learning (ICL).[3] ICL, the most prominent form of which is prompt engineering, is the ability of users to shape a fine-tuned model's outputs through prompting. Unlike pretraining and fine-tuning, ICL does not result in permanent change to the model; in technical terms, ICL does not alter the weights and other parameters of the model.

In addition to these three stages, generative AI models can be placed within software systems that provide *input and output filters*, a fourth approach for shaping their behavior. Input filters detect problematic user inputs and prevent models from seeing them. Output filters detect problematic model outputs and prevent users from seeing them.

All four of these approaches can shape the outputs of generative AI, yet they vary widely along many dimensions, including how much they cost, the amount and type of data they require, the permanence of the changes they effect, and the technical know-how they demand.

The second major contribution of this Essay is to encourage legal scholars, policymakers, regulators, and judges to focus on the fine-tuning stage as the best stage to address many (but not all) problems with generative AI. Reshaping one of these models during fine-tuning represents a better balance between power, precision, and disruption of outsider interventions than acting during any of the other stages. Fine-tuning gives those outside the companies that create these models a targeted ability to change them without the massive expenses and worse precision of shaping a model during pretraining.

Focusing on fine-tuning will also upend some emerging conventional wisdom about AI and policy. As one example, many legal commentators have come to understand that with AI models that predate the advances of the past few years, a generic "training" stage is the best opportunity to root out bias in a model, often

___ J. COPYRIGHT SOC'Y ___, 32-33 (forthcoming 2024), https://perma.cc/YDY7-ZAS6 (Nov. 14, 2023 version) (listing eight stages of the "generative-AI supply chain").

[2] *See id*. The term "foundation model" was coined by researchers at the Stanford Institute for Human-Centered Artificial Intelligence. *Id.* at 6-7. *See* Lee, Cooper & Grimmelman, *supra* note 1, at 41 n.194 (noting controversy in the ML community about possible confusion over the "foundation model" term and opting to call these "pretrained models" or "base models" instead).

[3] Bommansani, *supra* note 1, at 5.

pointing to the age-old computing maxim, "garbage in-garbage out."[4] Some evidence suggests that for generative AI models, removing the "garbage" during pretraining might be the worst of all worlds for attacking bias, resulting in models that are both less powerful and still biased.[5] Although there is much we still don't know, this research suggests that we should focus our bias mitigation strategies on fine-tuning instead.[6]

This Essay proceeds in three parts. Part II explains the differences between pretraining, fine-tuning, in-context learning, and input and output filtering, focusing in particular on the relative costs of data, compute, and money at each stage. Part III explains why fine-tuning presents the best mix of cost versus capability for shaping the outputs of generative AI in many cases. Finally, Part IV looks at what these insights mean for legal and policy interventions.

II. PRETRAINING, FINE-TUNING, IN-CONTEXT LEARNING, AND FILTERING

Those who train and deploy generative AI models do so in four distinct stages: pretraining, fine-tuning, in-context learning, and filtering. The differences between these stages are presented in this Part, with a focus on five characteristics: when the approach takes place; what kinds of behaviors can be instilled in the model in each stage; how much each stage costs in terms of data, money, or compute; who is able to take advantage of the stage; and what is produced. The varying answers to these questions will suggest different types of interventions for each stage, which will be developed in Part III.

A. Preliminary Thoughts: Transfer Learning and the Durability of the Four-Stage Model

Before turning to the survey of the stages, focus first on *why* model builders train models in this way, and whether this is likely to change in the near future. The four-stage model serves a goal known as transfer learning, an approach for building powerful and flexible machine learning models. There is good reason to believe that we will continue to train large generative AI models using these four stages—and in particular the first two stages of pretraining and fine-tuning—long enough for these stages to be of interest to policymakers.

Pretraining and fine-tuning take advantage of what is called *transfer learning*, an approach to machine learning in which a model trained to perform a particular

[4] Sandra Mayson, *Bias-In, Bias-Out*, 128 YALE L.J. 2218, 2224 (2019); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 683 (2016).

[5] Michael Gira et al., *Debiasing Pre-Trained Language Models via Efficient Fine-Tuning, in* PROCEEDINGS 2D WORKSHOP LANGUAGE TECH. FOR EQUALITY, DIVERSITY AND INCLUSION 59 (B. Chakravarthi et al. eds. 2022).

[6] Angelina Wang & Olga Russakovsky, *Overwriting Pretrained Bias with Finetuning Data* 3957, 3965 (2023), https://perma.cc/3984-7Y4Y.

task can be further trained to perform a different task.[7] Although transfer learning has been around for almost fifty years,[8] in the modern approach, massive models are first *pretrained* on massive datasets to provide a baseline of capability, such as the ability to represent human language or generate images fluently. These pretrained models, called base or foundation models, can then be *fine-tuned* through additional training to provide more targeted, specific, and useful capabilities, such as a chatbot capable of communicating conversationally. Transfer learning is seen as a significant advance on older approaches of machine learning model development, which produced single-purpose models that could not be reused for other purposes.

Unlike transfer learning, which has been around for decades, the distinction between pretraining and fine-tuning is a relatively recent innovation. This raises an important question about this project: is this Essay focused on a technological phenomenon that will be replaced by something better tomorrow? Legal scholars and policymakers must strike a balance when relying on observations about new technology and how it is made. On the one hand, the technological details matter immensely, and responsible participants must develop detailed and accurate understandings of the details of technology in the past, at present, and in the likely future. On the other hand, we must take care not to overfit our policies and intuitions on features of today's technological landscape that will be gone soon.

Given these concerns, is the pretraining/fine-tuning/in-context learning/filtering four-step framework both general and durable enough to deserve the attention of non-technical audiences? It is hard to say for certain, but the four-step framework for the development of generative AI systems seems stable enough to put at the center of our short-term and mid-term policy responses, for at least three reasons. First, the models that have attracted the most attention of legal scholars and policymakers over the past five years—including GPT-n (underlying ChatGPT), Bard, Claude, LLaMA, Bert, DALL-E, and many more—were developed using this approach.[9]

Second, model builders are strongly incentivized to continue to use the pretraining/fine-tuning two-step, as it provides significant efficiencies. It economizes on the allocation of the raw resources required—compute, data, and time—and in turn the way these resources can be paid for and shared. The kind of unprecedented power the world has been gobsmacked to observe in systems like Stable Diffusion and ChatGPT are the direct result of the massive amount of compute and data processed during pretraining. The fine-tuning step, in turn, allows

[7] Lorien Pratt & Sebastian Thrun, *Machine Learning - Special Issue on Inductive Transfer*, July 1997.

[8] Stevo Bozinovski & Ante Fulgosi, *The influence of pattern similarity and transfer learning upon the training of a base perceptron B2*, PROC. OF SYMPOSIUM INFORMATICA 3-121-5 (1976); Stevo Bozinovski, *Reminder of the First Paper on Transfer Learning in Neural Networks, 1976*, INFORMATICA 44 at 291-302 (2020).

[9] *See* Bommasani et al., *supra* note 1, at 5 (noting the rapidly developing homogenization of natural language processing around one set of designs).

companies to create specialized and tailored versions of these models at relatively low cost.

Third, these features of industrial production describe the world at an appropriately generalized level. Consider just the past decade of work on language generation. Ten years ago, the state of the art focused on recurrent neural networks. Within a few years, the focus had shifted to "long short-term memory."[10] Only five years ago, the focus shifted again to transformer architectures built on self-attention mechanisms.[11] Each of these advances dethroned the prior advance.[12] All rose to prominence and then faded probably too quickly to have justifiably been the focus of policymaker attention.[13] Yet all three relied on the same stable base of pretraining and fine-tuning.

Given the unprecedented speed of innovation in generative AI, it is foolhardy to predict exactly how long pretraining and fine-tuning will be the predominant approach for creating generative AI. It seems highly unlikely that something will replace it in less than a year; and it seems equally unlikely that we'll be using exactly these terms to describe these two steps in twenty years. The gap between one and twenty years seems like an appropriate time frame of stability and durability for legal scholars and policymakers to take into account. Consider that the European Union took about two years to bring the AI Act from proposal to final enactment, which seems about as quick as possible for such a comprehensive and important piece of legislation. Two years thus seems like a realistic if ambitious yardstick (meter stick?) for minimum durability, and it falls within my (admittedly speculative) time horizon for the relevance of the pretraining/fine-tuning two-step.

B. The Four Stages of Shaping Generative AI

There are four distinct phases during which generative AI models can be shaped: pretraining, fine-tuning, in-context learning, and filtering. In Parts II and III, I will argue that fine-tuning provides the best opportunities for outside entities such as policymakers and judges to influence the shape and outputs of these models. Fine-tuning presents the best tradeoffs of power and precision against costs and the risks of unintended consequences for advancing most policy goals. To understand these tradeoffs better, we must first understand the many and subtle differences between the four stages.

[10] *See* Ganesh Prasad Bhandari, *From RNN to Transformers: The NLP Evolution*, MEDIUM, Sept. 28, 2023, https://medium.com/ai-innovations-digest/from-rnn-to-transformers-the-nlp-evolution-5aeb194ca148 (tracking the history of NLP advances).

[11] *Id.*

[12] *Id.*

[13] *See* Yan Maksi, The Evolution of Transformers and LLM (Large Language Models): 💡Roadmap, Impact, and Future Prospects, KAGGLE, (2023) https://www.kaggle.com/discussions/general/458973 (showing timelines of developments from RNNs to LSTMs to transformer-based networks).

Below is a table outlining the aspects of the four stages I will cover in greater detail in the coming pages, to orient the reader, especially those who might want to skip some of this detail to get to the law-and-policy implications in later Parts:

Table 1: Comparison of the four stages used to shape generative AI

	Pretraining	**Fine-Tuning**	**In-Context Learning**	**Filtering**
When	First	After pretraining, repeatedly	After fine-tuning, repeatedly	Either after pretraining or fine-tuning
Goal	General capabilities	Outputs that are tailored, useful, and aligned	Better outputs	Fewer harmful inputs and outputs
Costs	Tens to hundreds of millions of dollars, or more	Wide range depending on goals, from dollars to millions	Negligible cost of processing a prompt	Wide range depending on goals
Who	Wealthiest tech companies and countries	Anyone with access to a pretrained model, data, and technical knowhow	Any generative AI model user	Model deployers
Product	Trained model (weights, hyperparameters)	Trained model (weights, hyperparameters)	Single set of outputs	No change to underlying model
Examples	GPT-4-base, LLaMA	GPT-4-final, Claude, ChatGPT, LLaMA2	Chain-of-thought prompting[14]	ChatGPT's refusal to answer some prompts

Consider the four stages of AI model development in turn:

1. Pretraining

The first phase of developing an AI model is known as *pretraining*. Pretraining is the process of building what is sometimes referred to as a "base" model or a "foundation" model.[15] Pretraining is accomplished by training a neural network (or

[14] Jason Wei et al., ARXIV, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (2022), https://arxiv.org/abs/2201.11903.

[15] Bommansani, *supra* note 1, at 6-7.

more likely, a complex system of neural networks interconnected in precise and complicated ways) on massive amounts of training data.

A quick note on terminology: AI tends to be even more jargon-laden than computer science generally, and the words used to describe aspects of generative AI are especially volatile. The line between pretraining and fine-tuning is one example. I am restricting pretraining to the steps taken to create the bare base model. Others might include some of the additional steps designed to make the base model more useful or less harmful to be part of pretraining as well. My usage pushes any training that happens after the base model is produced to the fine-tuning stage.

When It Occurs. Pretraining happens at the beginning. It is literally the first step that brings the foundation model into being. Because pretraining tends to be expensive and time-consuming, it ideally happens infrequently. Once a foundation model has been trained through pretraining, it tends not to be retrained without good reason.

The emerging industrial practice is to replace foundation models with newer versions, as opposed to retraining the existing base model (indicated with version numbers like other forms of complex software) that represent significant increases in size and power. Thus we see iterations such as GPT-2, 3, 3.5, and 4, or LLaMA and LLaMA2.

What Can be Accomplished. The goal during pretraining is to build powerful and general capabilities into the model. For a model trained on text, this might include the ability to produce language fluently or the ability to build internal representations of semantic meaning; for a model trained on images, this might include the ability to generate new images. The emphasis is on the generality of capability, meaning model designers may steer away from pretraining models that are overly specialized, for fear that such training might limit what the model can be fine-tuned to do.

Costs. Pretraining is expensive.[16] It may qualify as among the most expensive discrete industrial actions taken today. It is expensive along all three measures of cost: compute, data, and price. GPT-3 was trained on about 570 gigabytes of training data,[17] and required 3.14 x 10^{23} (one hundred sextillion or one hundred billion trillion) computer operations of compute.[18] Back-of-the-envelope calculations estimate the time costs of the compute alone at roughly $5 million.[19] OpenAI has not been transparent about the amount of resources it expended training

[16] Lee, Cooper & Grimmelmann, *supra* note 1, at 40 ("Altogether, the dollar cost [of pre-training] can range from six to eight figures . . . .").

[17] Tom Brown et al., OPENAI, *Language Models are Few-Shot Learners* 8-9 (2020), https://perma.cc/D7XW-CYVC.

[18] *Id.* at 46 tbl.D.1.

[19] Chuan Li, *OpenAI's GPT-3 Language Model: A Technical Overview*, LAMBDA LABS BLOG, June 3, 2020.

GPT-4, but estimates and rumors suggest it was trained on about one petabyte of data and cost more than $100 million.[20]

These expenses are just the baseline. One of the most important debates in generative AI is about scale. Many theorize that the power of generative AI models is directly related to the scale of the model—meaning the amount of training data used and the number of 'nodes' and 'weights' in the model.[21] If true, it means that progress in these pursuits will require more data and larger models. So relative capability increases which cost $100 million today could soon require billions.

Who Can Do It. Given the costs, pretraining large-scale generative AI models can only be done by some of the largest and wealthiest entities in the world. The massive compute economies of scale required have been achieved only by massive cloud computing companies, giving the big providers—Amazon, Microsoft, and Google—a competitive advantage. In turn, smaller start-up firms like OpenAI and Anthropic must seek partnerships with one of these giants relatively early in their business lifecycles.

Although the massive datasets that are used in pretraining tend to be made available for free or for a much lower cost than the cost of compute, they come with the hidden cost of litigation risk, as many lawsuits have been filed alleging that these datasets have been assembled, distributed, and reproduced in violation of laws including copyright law, privacy and data protection laws, and the Computer Fraud and Abuse Act.[22] All of these technical and litigation costs suggest that pretraining

[20] Will Knight, *OpenAI's CEO Says the Age of Giant AI Models Is Already Over*, WIRED (Apr. 17, 2023), https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/ (quoting OpenAI CEO Sam Altman for $100 million figure).

[21] *See* Lee, Cooper & Grimmelmann, *supra* note 1, at 31 ("Today's generative-AI models are able to produce incredible content, in large part because of their large scale."). A "node" is the basic processing unit in a neural network, one of millions or billions in large-scale foundation models. Larry Hardesty, *Explained: Neural Networks*, MIT NEWS, April 14, 2017, https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414. A "weight" is the strength of the connection between two nodes. *Id.*

[22] Blake Brittain, *Lawsuits Accuse AI Content Creators of Misusing Copyrighted Work*, REUTERS (Jan. 17, 2023), https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/ (describing suit alleging violations of copyright, right of publicity, unfair competition, and breach of contract); *Getty Images v. Stability AI*, BAKERHOSTETLER, https://perma.cc/8H68-6GJ3 (discussing copyright, trademark, unfair competition, trademark dilution, and deceptive trade practices); Michael M. Grynbaum & Ryan Mac, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*, N.Y. TIMES (Dec. 27, 2023), https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (discussing copyright, unfair competition, and trademark dilution); Eugene Volokh, *First (?) Libel-by-AI (ChatGPT) Lawsuit Filed*, VOLOKH CONSPIRACY (Jun. 6, 2023), https://perma.cc/S56F-LX7W (discussing libel challenges); D. Reed Freeman Jr. et al., *Data Scraping, Privacy Law, and the Latest Challenge to the Generative AI Business Model*, NATIONAL LAW REVIEW (Jul. 17, 2023), https://perma.cc/TE77-66FV (discussing challenges under the Computer Fraud and Abuse Act); Isaiah Poritz, *OpenAI Hit With Class Action Over 'Unprecedented' Web Scraping*, BLOOMBERG LAW (Jun. 28, 2023), https://perma.cc/Q4VV-6GAW (discussing challenges under the Electronic Communications Privacy Act, Computer Fraud and Abuse Act, California Invasion of Privacy Act, Illinois Biometric Information Privacy Act, various tort privacy laws, and more).

state-of-the-art generative AI models is an activity restricted to a handful of companies and nation-states.[23]

What is Produced. Pretraining generates a pretrained model, meaning the hyperparameters, parameters (weights and biases), and other design details needed to use the model. All of these pieces can be *deployed*, meaning configured to accept input data (e.g., a prompt for a large language model) and generate output data (e.g., a response to the prompt).[24] These pieces may also be *shared*, transferred to any third-party to deploy on their own.[25] When shared, all of the general capabilities of the model can be used by the recipient, who does not need any of the training data (nor knowledge of what was in the training data) to utilize the full power of the base model.

2. Fine-Tuning

The second phase of training generative AI models is known as *fine-tuning*. Fine-tuning is the process of refining a pretrained base model. A fine-tuned model has the same architecture created during pretraining—the neural network or system of neural networks.. Fine-tuning starts where pretraining leaves off, meaning it begins with the parameters (weights and biases) derived during pretraining.

When It Occurs. By definition, fine-tuning happens after pretraining; one needs a base model to fine-tune. Fine-tuning, however, is not the same one-and-done process as pretraining. A single base model can be fine-tuned repeatedly and successively to serve different goals.[26] Models can be fine-tuned and discarded in an iterative trial-and-error process of improvement. Fine-tuned models can be further fine-tuned, leading to an ancestral family tree of parent-child-grandchild relationships between models.

What Can be Accomplished. As the name suggests, fine-tuning produces models that build or extend upon the capabilities of the base model, creating models with capabilities that can be more targeted, specialized, or nuanced than the generic state of the base model.[27] As a prominent set of examples, the base model of OpenAI's GPT-4 is a language generating machine, meaning it predicts the next word to complete a given prompt. It does not, however, automatically understand how to hold a conversation—meaning how to respond to an ongoing back-and-forth of questions-and-answers, building on what has been said so far. OpenAI had to

[23] *See* Craig S. Smith, *What Large Models Cost You – There Is No Free AI Lunch*, FORBES (Sept. 8, 2023), https://perma.cc/3GTZ-5Y8A (summarizing costs of training large language models).

[24] Lee, Cooper & Grimmelmann, *supra* note 1, at 45-49 (describing generative-AI deployment).

[25] *Id.*

[26] *Id.* at 42-45 (describing how fine-tuned models can be further fine-tuned).

[27] JI et al., *supra* note 1, at 7 ("Supervised fine-tuning can be quite powerful in the right context when the right kind of data is available, and is one of the best ways to specialize a model to a specific domain or use case.").

fine-tune its non-conversational base model to understand the art of conversation, creating a fine-tuned model that could act like chatbot, namely ChatGPT.[28]

In the example in the preceding paragraph, fine-tuning was used to make a general language model a more useful but still general model. Fine-tuning can also be used to make a general model more useful but highly specialized. A base model fine-tuned on Python source code will be very good at generating Python source code but maybe worse at other language tasks than the base model was. A base model fine-tuned to translate English to French might be very good at translation, but not at other tasks, such as writing code.

Fine-tuning is also where *alignment* often happens.[29] When model builders identify behaviors in a base model that do not accord with their desired human values, they try to better align the model using fine-tuning.[30] For example, the base model of GPT-4 will generate text containing instructions for building bioweapons, racist speech, conspiracy theories, advice on committing suicide, and a host of other behaviors we would find dangerous, illegal, or otherwise unacceptable in an ordered society.[31] OpenAI invested millions of dollars and many months of effort into fine-tuning the base model to try to avoid as many of these behaviors as possible.

Fine-tuning can be used in myriad other ways, ranging from the profound to the trivial. OpenAI's webpage on fine-tuning gives the example of fine-tuning GPT-4 to be more sarcastic.[32] Others have fine-tuned language models to mimic a particular style of speech or image models to emphasize photorealistic imagery.

At this stage, companies focus on two major styles of fine-tuning. The first is *supervised fine-tuning*.[33] In supervised fine-tuning, a pretrained model is shown additional data to steer its behavior. To fine-tune an LLM to engage in conversation, the model can be shown examples of conversation.

The second broad category is known as *reinforcement learning with feedback*. Here, the training data used in fine-tuning is generated by having a human being (typically) react to the outputs of the model being fine-tuned. For example, a human (probably an expert in bioweapons) might be assigned to judge different possible

[28] Fine-tuning a model to improve its ability to respond to particular kinds of instructions is often referred to as "instruction tuning." *See* JI et al., *supra* note 1, at 7.

[29] In their work on generative-AI and copyright, Lee, Cooper, and Grimmelmann refer to "alignment" as a separate stage of the "generative-AI supply chain" from fine-tuning. Lee, Cooper & Grimmelmann, *supra* note 1, at 53-55. I treat alignment more as an overarching goal—one that happens to take place very often during fine-tuning—rather than a separate stage.

[30] *See* BRIAN CHRISTENSEN, THE ALIGNMENT PROBLEM 13 (2022).

[31] Bommasani, *supra* note 1, at 9 ("For example, a problematic model capable of generating toxic content might be tolerable if appropriate precautions are taken downstream. The extra application-specific logic is crucial for mitigating harms.").

[32] *See Fine-tuning*, OPENAI, https://platform.openai.com/docs/guides/fine-tuning (last visited Jan. 29, 2024).

[33] JI et al., *supra* note 1, at 7-9.

outputs to a prompt and flag as undesirable the outputs that tread too closely to teaching someone how to build a bioweapon.[34]

This distinction between supervised fine-tuning and reinforcement learning with feedback has very important public policy implications and could easily be the focus of an entire essay.

Costs. The costs of fine-tuning can range widely depending on the goals of the fine-tuner. Just like pretraining, fine-tuning requires training data, compute, and money. The cost of fine-tuning can range from trivially inexpensive to rivaling the cost of pretraining. Given the wide range of possibilities, it is difficult to talk about costs generally, but let us make some observations about what tends to drive these costs up or down.

Cost Observation 1: Fine-Tuning Can Rely on Much Less Data Than Pretraining. This is one of the most important observations for public policy that this Essay makes. Computer scientists seem to observe that a key advantage of breaking the process leading to generative AI into pretraining and fine-tuning is to economize on data and compute. A pretrained model with powerful general capabilities (e.g., language representation or image generation) can be fine-tuned with a relatively miniscule amount of training data. Research has shown fine-tuning has the ability to effect major changes in capabilities using only dozens or hundreds (rather than millions or billions) of pieces of new training data.[35] At the low end, some studies demonstrate how to profoundly alter a model's behavior based on as few as ten new examples.[36]

Comparisons to human thought and learning are fraught, but I will broach one here. A hallmark of human intelligence is our ability to do what AI researchers call "zero shot" learning.[37] Because our brains are "trained" to think generally, we can be instructed on a novel task we have never thought about before and perform it without seeing any examples. We can follow an instruction from a manager to

[34] Within reinforcement learning with feedback, there are at least two important subcategories, depending on the entity providing the feedback. In reinforcement learning with *human* feedback (RLHF), human beings provide the feedback. Alternatively, the feedback on specific examples can come from another model trained on general precepts of good or bad, right or wrong, safe or unsafe, ethical or unethical, etc. This approach is known as reinforcement learning with *artificial intelligence* feedback (RLAIF). *See* Stephen Casper et al., ARXIV, *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback* (2023), https://perma.cc/4S35-WWGD.

[35] *See* Teven Le Scao & Alexander M. Rush, *How Many Data Points is a Prompt Worth?*, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (finding that a single well-designed prompt can have the same effect as 280-3500 training examples).

[36] Haokun Liu et al., ARXIV, *Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning* (2022), https://arxiv.org/abs/2205.05638.

[37] *See* Takeshi Kojima et al., ARXIV, *Large Language Models are Zero-Shot Reasoners* 1 (2022), https://perma.cc/X93R-66KU ("The success of large language models (LLMs) is often attributed to (in-context) few-shot or zero-shot learning. It can solve various tasks by simply conditioning the models on a few examples (few-shot) or instructions describing the task (zero-shot).").

generate a particular kind of spreadsheet or from a professor to write a particular kind of essay with no need for a model to work from. In addition, we are fantastic "few shot" learners.[38] If that manager gives us a model spreadsheet with a few examples or if the professor gives us a few model thesis statements, we can perform as well as if they had given us dozens or hundreds of such examples.

Foundation models exhibit similar behavior. With their base model capabilities, owing to having been trained on millions or billions of examples during pretraining, all it takes is a tiny number of new examples to fundamentally change the model's apparent behavior.[39] By building atop the general capabilities instilled during pretraining, the model can learn to do what seem like very different skills based on relatively few additional examples.

Cost Observation 2: Fine-Tuned Data Can Be Very Expensive. Offsetting the fact that fine-tuning can be based on very little data is the fact that fine-tuning data is often fundamentally different than pretraining data. In pretraining, data is largely a numbers game. The capability of a base model tends to scale with the amount of data. A base model trained on billions of examples will be much more powerful than one trained only on millions.[40] In addition, pretrained data can be less processed data, too. An important insight that has helped lead to modern advances in generative AI is that with enough training data and compute, we can save ourselves the need to do so much human cleaning of training data. The billions or trillions of model weights in modern neural networks are enough to separate the good stuff from the mess.

In contrast, fine-tuning is often about prioritizing data quality over quantity. At the fine-tuning stage, there is more of a premium placed on cleaned and labeled data, which takes time and money to produce. To create a sarcastic version of GPT-4, we need samples of sarcastic answers to questions.[41] To create a diffusion model that does not generate racist imagery, we need human beings to mark racist outputs.[42]

[38] *See* Tom B. Brown et al., ARXIV, *Language Models are Few-Shot Learners* 3-4 (2020), https://perma.cc/5U8P-A7NZ (“[H]umans do not require large supervised datasets to learn most language tasks – a brief directive in natural language ⋯ or at most a tiny number of demonstrations ⋯ is often sufficient to enable a human to perform a new task to at least a reasonable degree of competence.”).

[39] *Id.*

[40] *See* Jared Kaplan et al., ARXIV, *Scaling Laws for Neural Language Models* 4-6 (2020), https://perma.cc/MYC3-LH8X .

[41] *See* OPENAI, *supra* note 32.

[42] Stephen Casper et al., ARXIV, *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback* 2 (2023), https://perma.cc/B8PH-CWHJ ("We use RLHF to refer to methods that combine three interconnected processes: feedback collection, re-ward modeling, and policy optimization. Figure 1 (top) illustrates this setup. The feedback process elicits evaluations of model outputs from humans. The reward modeling process uses supervised learning to train a reward model that imitates these evaluations. The policy optimization process optimizes the AI system to produce outputs that receive favorable evaluations from the reward model.".)

The need to hire skilled human labor as a source for some kinds of high-quality fine-tuning training data is a relatively recent development.[43] Many have rightly pointed out the way companies exploit low-wage and low-skilled labor who are asked to interact with disturbing content as part of reinforcement learning with human feedback. These critiques raise justifiable and important concerns about workers' rights, worker exploitation, and (given the tendency for the tech industry to turn to workers in less developed countries) entrenched patterns of colonial exploitation.[44] These are all valid and important concerns that policymakers must attend to. But a parallel trend is a growing price war for trained human labelers. Literature majors are needed to write new texts to train bots to sound like Shakespeare. (Alas, our original supplier of Shakespearean texts is no longer available!) Experts in bioweapon proliferation, hate speech, election misinformation, and cybersecurity, to name only a few of countless examples, are finding themselves in sudden demand for reinforcement learning with *human* feedback (RLHF) and labeling work. And given the salaries commanded by experts in many of these areas, this labor does not come cheap. These costs mean that the price of this bespoke data, labeled by skilled-labor, used for fine-tuning will often surpass the usually significantly larger datasets used for pretraining.

Cost Observation 3: Fine-Tuning Requires a Lot Less Compute. Finally, fine-tuning generally requires much, much less compute than pretraining. This is related to observation 1: training on smaller datasets requires many fewer training rounds, leading to a much smaller compute need.

Costs Bottom Line: Fine-tuning can be much, much, *much* cheaper than pretraining.[45] Although the costs of obtaining or creating datasets for fine-tuning can rival or exceed the costs of pretraining, the smaller datasets, more modest behavioral change goals, and smaller compute requirements can require much smaller investments of money, other capital, and labor. To give a sense of scale, OpenAI allows anyone to produce their own bespoke fine-tuned versions of GPT-3.5 (and they promise to offer the same service for GPT-4 soon).[46] As of the time of this publication, fine-tuning the top-end GPT-3.5 model costs $0.008 (8 tenths of one cent) for 1000 tokens.[47] So assuming a fine-tuning dataset has examples with

[43] *Id*. at 8 ("[C]orrections are relatively high effort and depend on the skill level of the evaluator."); Bommasani et al., *supra* note 1, at 86 ("[A]nnotating MRI data requires expert medical knowledge, whereas labeling sentiment for English texts requires only common sense judgement."). A company well-known for providing these kinds of services is Scale AI. *See* Josh Dzieza, *AI is a lot of work*, THE VERGE (Jun 20, 2023, 8:05 AM), https://perma.cc/PV3G-UP2S.

[44] *See* Billy Perrigo, *OpenAI used Kenyan workers on less than $2 per hour*, TIME (January 18, 2023 7:00 AM), https://time.com/6247678/openai-chatgpt-kenya-workers/.

[45] *See* Sachin Kumar et al., ARXIV, *Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey* 7 (2023), https://perma.cc/44EU-JTPH ("Designing and training models from scratch to mitigate harms can incur heavy environmental and resource costs. In contrast, an alternative branch of work has developed methods for modifying the model parameters of already-trained LMs, which requires much fewer resources.").

[46] OPENAI, *supra* note 32.

[47] *OpenAI Pricing*, OPENAI, https://perma.cc/7BUA-ENYQ (last visited Jan. 29, 2024).

long sentences averaging, say, 100 tokens each, then training 10,000 example sentences would cost $8.[48]

Who Can Do It. Unlike pretraining, fine-tuning is available to any person or entity with three things: the right level of access to a pretrained model, a dataset, and a bit of technical know-how. The first requirement means that those who pretrain models get to decide who can fine-tune their models. They serve as the initial gatekeepers to fine-tuning. OpenAI allows anyone with an account to fine-tune GPT-3.5 for a small fee. Meta has released the weights of the LLaMA and LLaMA2 models, meaning anybody with an internet connection, a laptop, and enough disk space satisfies the first requirement. But regardless of whether a company uses a gatekept or open-sourced model, fine-tuning is a relatively democratized activity compared to pretraining. Anyone can fine-tune. Policymakers without any technical training can learn enough to produce a fine-tuned version of GPT-3.5 in a weekend for the cost of a few cups of coffee. Middle schoolers with a laptop and broadband Internet connection can fine-tune LLaMA for the cost of their time and some electricity.

What is Produced. Fine-tuning produces a fully-fledged model, complete with newly tuned parameters. Just as with a pretrained model, a fine-tuned model can be deployed or shared. The only difference between a pretrained and fine-tuned model is the difference in the behavior that has been fine-tuned. A fine-tuned large language model might be less prone to certain racist outputs (or more prone to certain racist outputs), better at answering questions or speaking sarcastically, or more capable of producing certain outputs (say, poetry or Python source code). A fine-tuned model might be trained on a customer's proprietary documents, or a government agency's internal memos, increasing its ability to produce text about non-public information.

To put it plainly, pretraining tends to focus on increasing power, general performance, and the capability for transfer learning. Fine-tuning accomplishes a much more diverse and tailored set of goals, much more coextensive with all of the goals the many actors in society might want to imbue into their models. The goals in fine-tuning might be broad and general, narrow and specific, and everything in-between.

[48] OpenAI's example is, "For a training file with 100,000 tokens trained over 3 epochs, the expected cost would be ~$2.40 USD." *Fine-tuning, Preparing Your Dataset*, OPENAI, https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset (last visited Jan. 29, 2024).

3. In-Context Learning

The next phase is known as *in-context learning* (ICL).[49] ICL describes methods typically used by end users rather than by model designers or software developers.[50] A well-known form of ICL is called *prompt engineering*. Prompt engineering describes the set of practices users can apply to the prompts they provide to an LLM or image diffusion model to produce better outputs.

When it Occurs. In-context learning happens after a trained model (pretrained or fine-tuned) has been deployed. Since most deployed models have been fine-tuned at least a little, most ICL examples that have been discussed in the literature operate on fine-tuned models and rarely on base models.

What Can be Accomplished. Computer scientists have repeatedly surprised themselves (and outsiders) by how much can be accomplished through ICL without any additional training. Early papers revealed that LLMs like GPT-3 are one-shot and few-shot learners for many tasks.[51] One of the most famous ICL results to date has been the discovery that simply adding the phrase "let's think step-by-step" to a prompt can make LLMs like GPT-2 more adept at solving math word problems, a technique called "chain-of-thought reasoning."[52]

What ICL can accomplish appears to be more limited than fine-tuning. There are likely tasks that can be accomplished with fine-tuning that cannot be accomplished through ICL. One reason is what is known as the "context window size bottleneck." LLMs are designed to allow a particular maximum size of user input, called the context window. The standard GPT-4 context window fits 8,000 tokens—a token is a complete word or part of a longer word—and after 8,000 tokens have been received, the model begins to "forget" or remove from the prompt the oldest information. That said, increasing the context window sizes of LLMs is a very active area of development, and the Claude LLM from Anthropic provides a context window of roughly 100,000 tokens.

The fixed size of a context window places an upper limit on the level of detail possible with ICL. ICL techniques cannot today, for example, process all of the cases ever decided by the Supreme Court nor even all of the words of the U.S. Code, both of which far surpass 8,000 tokens in length. In contrast, one can fine-tune on

[49] Tom B. Brown et al., ARXIV, *Language Models are Few-Shot Learners* 4 (2021), https://perma.cc/5U8P-A7NZ (coining and defining the term 'in-context learning').

[50] *See* George Musser, *How AI Knows Things No One Told It* (2023), https://perma.cc/E57X-MPX6; Model deployers sometimes use in-context learning too, by adding what is known as a "system prompt." Sunil Ramlochan, *System Prompts in Large Language Models*, PROMPT ENGINEERING & AI INST. BLOG, March 8, 2024, https://promptengineering.org/system-prompts-in-large-language-models/.

[51] Qingxiu Dong et al., *A Survey of In-Context Learning*, ARXIV (2022), https://arxiv.org/abs/2301.00234.

[52] Jason Wei et al., ARXIV, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (2022), https://arxiv.org/abs/2201.11903.

as much new training data as one chooses, meaning one could fine-tune GPT-3.5 on the Supreme Court's cases or the U.S. Code.

Costs. The only costs in ICL are the costs of processing a prompt or set of prompts. This is negligible compared to cost of fine-tuning even a small dataset. Most providers of LLMs provide a tier of free and unlimited prompting access, making ICL a free service. Access to more powerful models may require a small fee.[53]

Who Can Do It. Anybody with a web browser and free account can learn basic ICL techniques within minutes. In fact, prompt engineering appears to be emerging as a valuable technical skill. Employers are hiring prompt engineers and universities are teaching classes on it. Like many technical skills, one's skill at prompt engineering improves with study and experience.

What is Produced. The biggest limitation of ICL is that it does not persist outside a single session. ICL does not alter the weights of the model (the use of the word "learning" in ICL is a bit of a misnomer for this reason), meaning ICL does not affect the experiences of other users nor can an ICL-impacted model be shared directly.[54]

4. Input and Output Filtering

The fourth way of controlling the output of generative AI is to filter the inputs sent to and the outputs emerging from the model.[55] Unlike the prior three approaches, this approach does not directly shape the behavior of the model itself; it situates the model is a larger system of related software components.[56]

An *input filter* acts on user prompts. It attempts to detect certain prompts and either block or revise the prompt to achieve a particular goal. The model sees only what the input filter permits. An *output filter* acts on model outputs. It attempts to detect certain outputs and either block or revise those outputs to achieve a particular goal. The model can output only what the output filter permits. Input and output filters can be used in tandem, affecting both the inputs seen and the outputs generated by the model.

To illustrate, consider attempts to prevent large language models from teaching people how to build bioweapons. An input filter might try to detect attempts to learn how to build bioweapons, perhaps sending back, "I'm sorry but I cannot teach you how to do that," without ever involving the LLM at all. An output filter might try to detect LLM outputs that seem to explain how to build a bioweapon, regardless

[53] As of the publication of this Essay, access to OpenAI's GPT-4 model through the ChatGPT interface costs $20/month. *Pricing*, OPENAI, https://perma.cc/57F5-9TA2 (last visited Jan. 29, 2024).

[54] *See* Bommasani et al., *supra* note 1, at 87.

[55] JI et al., *supra* note 1, at 9; *See* Kumar et al., *supra* note 45, at 3.

[56] *See* Lee, Cooper & Grimmelmann, *supra* note 1, at 15 (emphasizing the importance of focusing not only on generative-AI models but also on the larger systems within which these models are placed).

of what the user originally input, sending back the same explanation. The details differ, but these differences may not be visible to the end user.

When it Occurs. Input and output filters are constructed after the pretrained (and possibly fine-tuned) model has been trained and becomes part of the system used in the model's deployment.

What Can be Accomplished. Just about any goal imaginable can be accomplished with input or output filtering. Some have drawn comparisons between filters for generative AI models and content moderation of social media platforms. A content moderation filter that can detect hateful speech can be literally repurposed to serve as an output filter for an LLM to detect hateful speech.

Costs. The cost of input and output filtering is as varied as the goals imaginable. Input and output filters are just software components, and they can be trivially inexpensive, very expensive, and everything in-between. In fact, quite often these filters take advantage of separate machine learning models, so the costs might involve the costs of training or fine-tuning another model.

Who Can Do It. Only one with control over the system in which a model is deployed can take advantage of input and output filtering. One must be situated between the end users and the model to take full advantage of this approach.

What is Produced. Like ICL, input and output filters do not change the underlying model. This means that if the model itself were to be extracted from its surrounding software, it would no longer be subject to the filters. On the other hand, unlike ICL, input and output filters persist from session to session and user to user.

III. WHAT CAN AND SHOULD BE DONE IN EACH STAGE

Part II introduced a method of industrial production of generative AI models—the separation into pretraining, fine-tuning, in-context-learning, and input and output filtering—that presents different opportunities and costs for those who would control the output of massive generative AI models. This Part translates this technical understanding into policy strategy. What do the technical characteristics of each phase make easy or difficult for those who would shape these models both inside and outside the companies that produce them? Are certain kinds of problems best tackled at one specific stage? As the development of these models is such a dynamic area of research, what would we like to know but do not yet know?

This Part comes to a general conclusion: in most cases, anyone who wants to control the behavior of a generative AI model should focus on the fine-tuning stage. ICL and input and output filters do not change the underlying model, making their effects temporary and contingent. Using pretraining to satisfy many goals is likely to be the worst of all worlds: doing so will likely make the base models less powerful, limiting what can be accomplished through transfer learning; require more money, compute, and data to accomplish the goal than would be required in fine-tuning; and possibly do a worse job satisfying the goal. Fine-tuning remains the best alternative in many cases.

A. Changing the Model Itself

Pretraining and fine-tuning change the model itself while ICL and filtering do not. This distinction matters for external actors trying to bring lasting change to the way a model behaves.

ICL changes the behavior of the model in the context of a single set of inputs, but will not affect any other sets of inputs. While ICL can be very important for users of generative AI models, its short-lived effects are unlikely often to be useful to policymakers or legal scholars seeking to change the behavior in a systematic or widespread manner.

Unlike ICL, filtering is built into code itself.[57] An input or output filter will affect the experience of all users who interact with a particular system. For this reason, policy interventions should often focus on filters: for example, judges may order a company to create or change the behavior of a filter, or regulators may require or scrutinize a filter.

There is one reason to prefer filtering over any of the other three stages: it can be more demonstrably foolproof. If one creates an output filter that always blocks a particular individual's name from ever being uttered by the model (for example, as a way to mitigate defamation), that filter can be highly precise and effective.[58]

Although filters can be highly effective tools for achieving policy goals, because they fail to alter the model itself, the protections they offer will often be easier to evade than pretraining or fine-tuning fixes. Models can be separated from the systems of software, including filters, that surround them.[59] A model is just a bundle of parameters and code that can be packaged into a file and transmitted across the internet. Filters are thus literally separable from the model. A toxic model policed by filters that prevent it from outputting toxicity can be made toxic again by the relatively simple act of removing the filters.[60]

For these reasons, the rest of this Essay focuses on the distinction between pretraining and fine-tuning, and says little more about ICL or filtering. That said,

[57] To be clear: some generative AI models permit a more highly integrated form of ICL, through use of so-called "system prompts." Think of a system prompt as a way to embed a useful ICL prompt into every user interaction. This use of ICL begins to be comparably "build into the code" as input and output filters.

[58] In this example, the filter also needs to account for variants of the person's name. So nicknames, aliases, etc. Also, it is hard to create a filter that can detect all possible identifying phrases for a person: "the Republic nominee for President in 2020," will not be blocked by an output filter programmed to look for the phrase, "Donald Trump."

[59] Just because it is possible to remove a model from its surrounding system, that is true only if the company deploying the model permits it. OpenAI, for example, is unlikely to permit most people from using the GPT-4 models without the company's surrounding system of code.

[60] These systems of models plus filters can be very complex and full of interconnected code and dependencies. It might not be easy to extricate a model from a complex system, and it might be possible to design a system to intentionally make it very difficult to extract the model.

filtering may prove to be an especially important target for policymaker intervention.

B. The Lessons of Transfer Learning

The goal of transfer learning is to train a general model that can be reshaped to accomplish specific tasks. The pretraining then fine-tuning sequence gives computer scientists an efficient process for pursuing this goal. This method of transfer learning thus presents a threshold question for those hoping to reshape a generative AI model: are you targeting a behavior that is fundamental and general or applied and targeted? The broadest and biggest goals may best be addressed during pretraining, while all other goals should cause us to focus on what happens in fine-tuning.[61]

It is important to understand the nature of what is meant by "fundamental and general." Such goals relate to basic, elemental, core behaviors and competencies of the model, the kind of attributes that undergird the essential abilities of the model. For large language models, they include goals such as "represent human language." Only goals on par with the fundamental way language is represented are the kind of goals that may cause one to focus on pretraining of an LLM.

As one admittedly charged example, the Chinese government has been critical of LLMs created by Chinese companies for having a pro-Western and pro-democracy slant.[62] Any such slant, if it exists, might be a result of the fact that much of the language data that is readily available is English language text created in the West. If true, then this slant might be the kind of fundamental behavior that is best "fixed" during pretraining, by increasing the relative share of Chinese-language text and decreasing the share of English language text.

At the other end of the spectrum are goals that are targeted to narrow, specific examples or contexts. Shaping an LLM to avoid defaming a particular person or constraining an image diffusion model to avoid infringing a particular artist's style are far too narrow and specific to be the focus of pretraining.

Between these two ends of the spectrum—between fundamental and very specific behaviors—are a host of medium-to-large scoped goals that might be attainable either through pretraining or fine-tuning. For these, we need to look at various tradeoffs.

[61] Bommasani et al, *supra* note 1, at 87 ("[Fine-tuning] is useful whenever the desired use case of a model differs from the relatively general training objective used for foundation model training.").

[62] *See* Cissy Zhou, *China Tells Big Tech Companies Not to Offer ChatGPT Services*, NIKKEI ASIA, Feb. 22, 2023, https://asia.nikkei.com/Business/China-tech/China-tells-big-tech-companies-not-to-offer-ChatGPT-services.

C. Assessing the Trade-offs

Goals that are any narrower than "represent human language" or "avoid a pro-Western slant" may possibly be addressed during either pretraining or fine-tuning. Deciding which of these two can accomplish this goal better is an active area of research into alignment.[63] Although this is a fast-moving field, the emerging consensus suggests that fine-tuning is a far more efficient way of accomplishing most of these goals, even ones that seem fairly broad, such as "be polite" or "be less racist." Let's consider the various tradeoffs that argue against using pretraining for goals like these.

First, the research suggests that pretraining requires many more training data examples than are required during fine-tuning to accomplish the same goal. One could try to instill politeness in a pretrained model by trying to identify and remove all of the impolite examples from the millions or billions of texts in the pretraining corpus. This would be an exorbitantly expensive and difficult task.

In contrast, research suggests that training on very few examples during fine-tuning can shape outputs in seemingly broad and fundamental ways. It may be that showing only a few thousand examples of polite language (labeled positively) and a few thousand examples of impolite language (labeled negatively) might be enough to steer a model's outputs toward politeness. In fact, it may take only a few hundred of each.[64] The basic intuition is that during fine-tuning, the model starts where pretraining left off, a model that already represents human language in great detail. Learning "one new trick"—politeness—during fine-tuning is much more efficient than learning how to represent all of language—plus politeness—all at once during pretraining.

Second, looking for the impolite needles in the pretraining haystack is not only expensive but is very likely to be impossible to do well. At the scale of millions or billions of training data examples, those searching for the impolite (or racist or harassing) examples will miss many impolite examples (false negatives) and mistakenly remove many polite examples (false positives).[65] The result will probably be far less success at meeting the goal of politeness.

As an example of a well-intentioned but poorly implemented attempt, consider the C4 dataset, created by Google and used during pretraining in the creation of

[63] *See* Helen Ngo et al., ARXIV, *Mitigating harm in language models with conditional-likelihood filtration* (2021), https://perma.cc/CEX7-EXXV (testing methods for reducing "harmful views" in large language models during pretraining and fine-tuning stages).

[64] *See* Solaiman and C. Dennison, ARXIV, *Process for adapting language models to society (palms) with values- targeted datasets* (2021), https://perma.cc/5UEL-ZU8N (using 80 handwritten question-answer pairs to reduce propensity to generate non-aligned text).

[65] Kumar et al., *supra* note 45, at 8 (noting that trying to filter out harmful examples from training data "admit[s] many false negatives [failing to detect documents with subtle toxicity] and false positives [erroneously flagging documents that discuss sensitive topics and use hateful speech as examples; additionally, removing data from different dialects like AAE], unintentionally exacerbating risks of marginalization and exclusion.").

many large language models.[66] To make C4, Google started with the "Common Crawl," an archive of web data commonly used to train language models. Google tried to create a filtered subset of the Common Crawl that was both more useful for natural language tasks and less prone to generate offensive text.[67] To accomplish these goals, Google removed "any page that contained any word on the 'List of Dirty, Naughty, Obscene, or Otherwise Bad Words.'"[68] This is a crowdsourced list initially created by Shutterstock to prevent offensive autocomplete suggestions on the Shutterstock website, and today has grown to include more than 400 English words as well as lists in more than two dozen other languages.[69]

This approach appears to have worst-of-both-worlds consequences. Removing crudely identified biased training data may not actually remove much offensiveness from the system. But deleting all documents that include the word "sex" to build foundation models may limit the transfer learning potential of those models to create important fine-tuned applications. Worse, the pages deleted disproportionately included websites focused on LGBTQ topics, because the list contains entries such as "gay sex."[70] The point is not to fault Google for trying to tackle bias at the root. The point is that doing so well is a daunting task, even for one of the wealthiest and most technologically advanced companies in the world.

Third, given the exorbitant costs of pretraining, there is very little room for error in meeting a goal. Although the procedure of pretraining a model involves many initial small-scale experiments to test how things are going, it culminates in a monthslong pretraining marathon, consuming significant compute resources and hundreds of millions of dollars. If one tries but fails to instill politeness or combat racism during pretraining, the only option would be to fix it in fine-tuning. Redoing pretraining is not an economically viable response.[71]

D. The Importance of Vantage Point

There is a less technical reason that legal scholars and policymakers should focus on fine-tuning: their vantage point.[72] The institutions and levers of power

[66] *See* Colin Raffel, et al., ARXIV, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* 12 (2023), https://perma.cc/25EX-FGUF.

[67] *Id.* at 4, 26.

[68] *Id.* at 6.

[69] Tom Simonite, *AI and the list of dirty, naughty, obscene, and otherwise bad words*, WIRED, Feb. 4, 2021, https://www.wired.com/story/ai-list-dirty-naughty-obscene-bad-words/. The list includes "three entries for Klingon . . . and 37 for Esperanto." Raffel et al., *supra* note 66.

[70] *Id.*

[71] Bommasani et al., *supra* note 1, at 88 ("[D]ue to the computationally demanding nature of training foundation models, frequent re-training from scratch might carry unacceptable financial or environmental impacts, or simply take too long to be a viable method for keeping models up to date.").

[72] *See* Kumar et al., *supra* note 45, at 9 ("Different stakeholders are involved in different model development phases with varying access to resources. As a result, intervention strategies are different depending on the stakeholder.").

available in policy, for example laws, regulations, and judicial injunctions, tend to operate in the time between pretraining and in-context learning.

Using a law or regulation to dictate what happens in pretraining is fraught. Statutes that govern technology tend to focus on the effects of built systems rather than dictate the design of new systems. Although a more recent trend sees laws that govern the design of technology, these tend to require adherence to general principles over specific design features. For example, the EU's General Data Protection Regulation (GDPR) requires "data protection by design and by default," but this requires vague and general design principles such as "data minimisation" and use limitation.[73]

Laws governing the design of generative AI models are likely to follow the same general template, requiring models to be "fair" or "unbiased" or "truthful." None of these general exhortations provides enough specific detail to obligate a model designer to do anything specific during pretraining.

In contrast, law and regulation are well-aligned to govern fine-tuning. Because fine-tuning tends to be the locus of "alignment," and because fine-tuning can happen repeatedly, laws and regulations can more specifically govern what happens at this stage.

Another way of putting this is to imagine the pretraining/fine-tuning two-step like a funnel. At the big end of the funnel—during pretraining—policymakers and judges can impose big changes on generative AI models that will fundamentally change everything the model does. At the narrow end of the funnel—fine-tuning—policymakers and judges can operate much more narrowly, precisely, and surgically, and their changes will have fewer collateral effects.

Preferring fine-tuning thus also advances humility and minimalism in policymaking. For example, judges may be especially inclined to minimize the unintended consequences of the changes they impose. A judge may want to stop a model from defaming a particular person or infringing a particular author's work, while worrying about causing other changes outside of their power and purview.

E. Focusing on Fine-Tuning and Unleashing Pretraining

Taking all of this into consideration, the policy interventions that would address a majority of the problems that have been raised about generative AI are better implemented during fine-tuning and not pretraining. Quite often, trying to fix a problem in pretraining will be much more costly and also less effective.

[73] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation ("GDPR")), OFF. J. EUR. UNION L. 119/1 (Apr. 5, 2016), Art. 25.

This suggests that we should be reluctant to focus on pretraining for most of our goals. Unless our goal is to target something fundamental to the model, effecting change during pretraining is probably inefficient and unlikely to work well.

This also means that law and policy may be relatively hands-off about what happens during pretraining. Aside from shaping the broadest, most fundamental goals, external actors may have less cause to dictate what happens during pretraining.

This will have important implications for competition policy. The costs of pretraining mean that only the largest and most powerful technology platforms can pretrain the largest models. A relatively hands-off policy toward influencing the design of pretrained models may entrench this power and exacerbate anticompetitive harms.

IV. IMPLICATIONS FOR LAW AND POLICY

The pretraining/fine-tuning distinction has arisen without much direct regulatory intervention. It is a product of market forces and computer engineering decisions, but as stated earlier, it seems to be a durable feature of the way these models will be created for at least the near future. This final Part considers some of the implications of this distinction for law and policy.

A. Challenges to the Legitimacy of Training Itself

Although this Essay urges policymakers to focus on fine-tuning to accomplish most of their goals, one important category of goals probably cannot be accomplished through fine-tuning alone: challenges to the very legitimacy or legality of pretraining itself. Two prominent examples are the prohibitions of copyright and data protection law. Owners of the copyrights in data believed or known to have been used to train generative AI models have sued companies like OpenAI and Stable Diffusion alleging that the act of training itself is copyright infringement and thus should not have occurred without prior permission.[74] Similarly, plaintiffs have sued these companies in Europe, alleging violations of the rights given to data subjects under the GDPR. For example: because data obtained for one purpose cannot be reused for a second purpose without the data subject's express consent.

Without taking sides in the roiling debates over whether the training that has occurred violates copyright law or the GDPR, these theories present problems that most likely would not be remediable in fine-tuning. In these cases, the very process of pretraining constitutes the alleged copyright or data protection violations, and both copyright law and the GDPR give judges and enforcement officials the power

[74] These plaintiffs usually also allege that the outputs of these models are separate acts of infringement. These are assertions of harms that do not question the legitimacy or legality of the act of producing outputs; they question specific outputs. These harms can better be addressed by reshaping the outputs, suggesting fine-tuning.

to prevent or enjoin future infringements. It is likely not sufficient to say that we should "focus on fine-tuning" to address these harms, and they need instead to be addressed during pretraining.

B. The Possibilities of Fine-Tuning

With a focus on fine-tuning, legal scholars and policymakers will need to modify bits of received and conventional wisdom that have accumulated in discussions over pre-generative AI systems.

First, shaping generative AI can be inexpensive. We must disabuse ourselves of the idea that harm-avoidance and alignment happens only at (or even mostly at) the exorbitant pretraining phase. We need to stop saying things like, "asking the company to retrain their model would be an unacceptably expensive step."

Second, because fine-tuned models can be further fine-tuned, we can shape generative AI systems step-by-step. One fine-tuning step can make a language-generating model better at answering questions, while a second can make it better at engaging in back-and-forth conversations, and a third can make it better at talking to children.

This makes shaping generative AI an iterative process. It removes the pressure to solve all problems at once. We can layer solutions atop one another, shaping models over time.

This also gives us the ability to experiment. We can apply a fix and then observe the result, testing the fine-tuned model to see if the fix worked, and checking to see if it broke something else as a result.

C. The Limits of Fine-Tuning

At the same time, fine-tuning has limits that policymakers and legal scholars must consider. In general, controlling an LLM in any stage—including fine-tuning—is an imprecise and not well-understood undertaking. "[R]esearchers have yet to fully understand exactly how to manipulate data in a way that will have meaningful impacts on the resulting model while minimizing performance loss."[75] The best we can do is make best efforts to shape the outputs, but we will often discover that our best efforts have succeeded only partly.[76] This is by necessity an iterative process of trial-and-error, which the economics of pretraining do not permit.

Fine-tuning to accomplish a particular goal will also often give rise to an arms race, as people try to find ways to circumvent the last fine-tuned change. We will be forever locked in a sequence of fine-tune, circumvent, fine-tune, circumvent, etc.

[75] JI et al., *supra* note 1, at 6.

[76] *Id.* ("[I]t is extremely difficult to predict how changing their training data will affect their performance or their propensity to output certain types of content.").

Fine-tuning will also have side-effects on the way the model operates.[77] The state of the art does not permit surgical changes to a single behavior without potential changes to the model's other capabilities. LLMs and image diffusion models are complex black boxes. GPT-4 comprises more than a trillion parameters, and fine-tuning will make miniscule changes to some of them.

The nature and extent of these side-effects cannot at present be anticipated, well-characterized, or easily measured. Fine-tuning a model to stop defaming someone, say, might make the model worse at solving some forms of math problems, or better at translating French to English, or slightly more racist.[78] Some have reported that ChatGPT's math performance has degraded over time, perhaps due to fine-tuning unrelated to mathematical reasoning.[79]

The side effects of fine-tuning is an area of active research. Some of this research has already been identified as important by computer scientists, but it should also be encouraged for its policy importance.

Finally, fine-tuning may not work. Computer scientists do not have a deep understanding of why giant generative AI models work, so they also cannot say to what extent a particular fine-tuning attempt will have the intended effect. It may work during testing but fail when subject to end-user testing. It may work most of the time but fail under certain conditions. It may work most of the time but fail randomly and unpredictably.

D. Case Study: When to Address Bias

As an important case study, focusing on fine-tuning and not pretraining may shift attitudes and recommendations about how to deal with biased generative AI systems. In this discussion, "bias" refers to legally actionable and socially problematic forms of bias—such as discrimination based on race, ethnicity, religion, gender, or sexual orientation—as opposed to the more technical definitions of bias sometimes used in the technical literature.

Researchers have proposed and are testing approaches for dealing with bias during both the pretraining and fine-tuning stages, with various studies suggesting pros and cons of each.[80] Some studies suggest that bias can be effectively addressed

[77] *See* Kumar et al., *supra* note 45, at 7 ("[F]inetuning or prompt-tuning on a small dataset may lead to overfitting reducing the general purpose utility of LMs."); Helen Ngo et al., ARXIV, *Mitigating harm in language models with conditional-likelihood filtration* 1 (2021), https://perma.cc/SQU9-GTMM (finding that filtering harmful examples from training data led to a "marginal decrease in performance" on some language benchmarks).

[78] Ngo et al., *supra* note 77, at 1; JI et al., *supra* note 1, at 6.

[79] Lingjiao Chen et al., *How is ChatGPT's Behavior Changing Over Time?*, ARXIV (2023) (suggesting ChatGPT's performance to do some math problems has degraded over time). *But see* Arvind Narayanan and Sayash Kapoor, *Is GPT-4 Getting Worse Over Time?,* AI SNAKE OIL BLOG, JULY 19, 2023, https://www.aisnakeoil.com/p/is-gpt-4-getting-worse-over-time (taking issue with the idea that the Chen paper should be interpreted as "degredation" as opposed to more neutral behavior change).

[80] Bommasani et al., *supra* note 1, at 134.

at the fine-tuning phase for much less cost and in a much more targeted way than during pretraining. For example, work by Wang and Russakovsky confirmed that although biased pretrained models can lead to biased fine-tuned models, "this bias can be relatively easily corrected" during finetuning, with little to no loss of accuracy or other measures of performance.[81]

Other studies suggest that removing bias during pretraining may produce fewer unintended side effects. Work suggests that language models tend to perform better on some widely used metrics for assessing the capabilities of such models when they are pretrained to avoid bias than when they are fine-tuned.[82]

Additional studies suggest that addressing bias during pretraining, by identifying and removing examples of biased data from the pretraining corpus, can make it harder to further address bias during fine-tuning. Research suggests the counterintuitive notion that the best way to root out biased behavior is to intentionally retain biased data at the pretraining stage, in order to provide an efficient target for elimination at the fine-tuning stage.[83] Addressing problems like bias at the fine-tuning stage provides more opportunities for tuning and improvement. Instead of the "one and done" expense of pretraining, during fine-tuning, we can train, test, and iterate to make the model better over time.

It is too early to state definitively whether bias is best addressed during pretraining or fine-tuning, and the most important thing we can do is continue researching the pros and cons of each approach. It is likely that the best approach will be a combination of both—we can remove some of the worst data during pretraining, perhaps shaping the fundamental nature of the model without sacrificing the model's capability. But there will always be an important space for fine-tuning to shape and hone and develop the operation of the model into the future.

V. Conclusion

Generative AI models have begun to transform our social, political, and economic lives. The challenge for all of us is to shape these new technologies to cause less harm and do more good. A key insight is to understand the relative costs and benefits at intervening during the pretraining, fine-tuning, in-context learning, and filtering stages. In many cases, the fine-tuning stage presents an opportunity to shape these models to be less harmful with the best tradeoffs between power and efficacy. Policymakers and legal scholars should focus on fine-tuning for much of what they hope to accomplish.

[81] Angelina Wang & Olga Russakovsky, *Overwriting Pretrained Bias with Finetuning Data*, Proceedings of the IEEE/CVF International Conference on Computer Vision (2023), at 3958, https://perma.cc/WA6A-Q5QN; Ji et al., *supra* note 1, at 7.

[82] Kumar et al., *supra* note 45, at 7.

[83] *See* Hyung Won Chung, et al., arXiv, *Scaling instruction-finetuned language models* 6 (2022), https://perma.cc/TT6W-FFAF.