disc
Data Integration Support Center
WestEd

2025
June

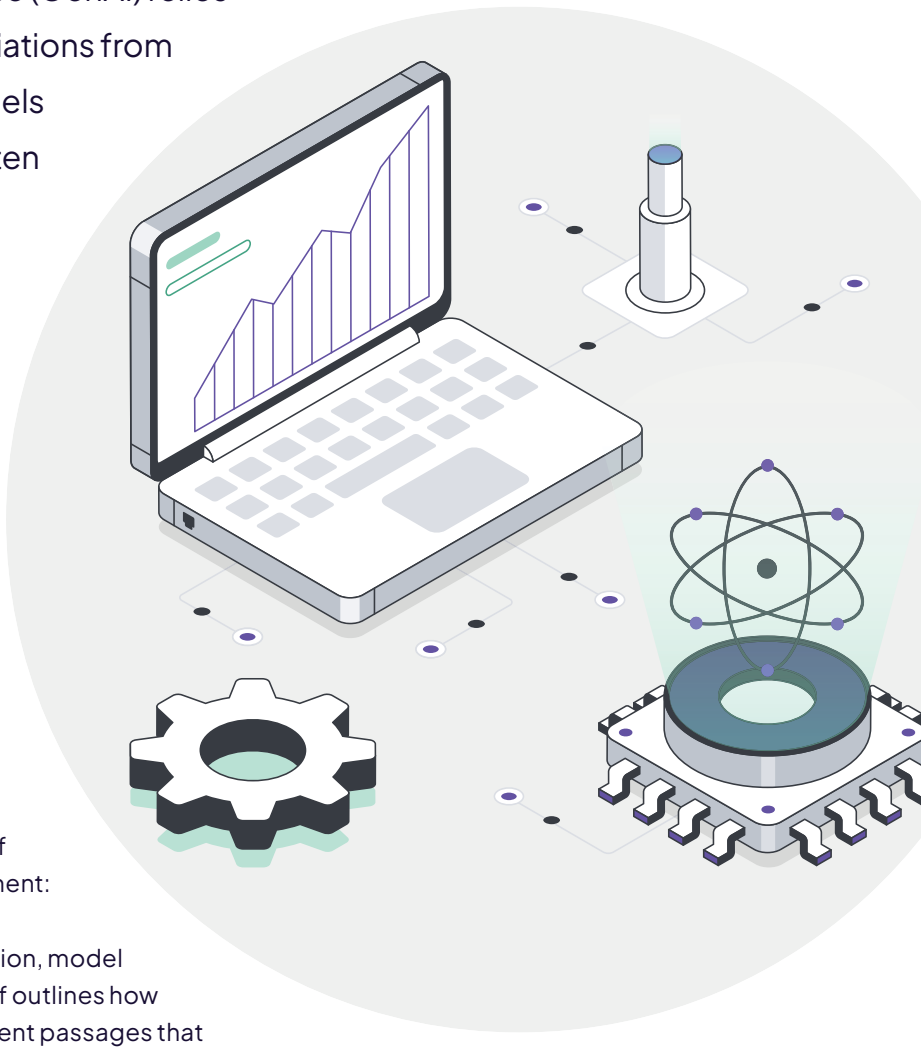
SPOTLIGHT

Fine-Tuning Models in a Secure AI Environment

PRACTICAL CONSIDERATIONS

AUTHORS:
Brandon LeBeau
Mitchell Clarke
Baron Rodriguez
Sarah Quesen

Modern generative artificial intelligence (GenAI) relies on pretrained models that learn associations from vast amounts of text. While these models perform well on general tasks, they often struggle with specialized applications requiring domain expertise or contextual understanding. Fine-tuning—the process of adapting pretrained models with additional targeted training data—offers organizations a powerful method to enhance model performance for specific use cases.



WestEd recently implemented a model fine-tuning project within its secure AI environment to improve the generation of English language arts (ELA) assessment content. A companion brief titled "Fine-Tuning Models in a Secure AI Environment: Technical Implementation" details the technical implementation process, including data preparation, model training, and evaluation methodologies. That brief outlines how WestEd fine-tuned a model to generate assessment passages that better align with grade-level standards and incorporate appropriate complexity while maintaining technical quality.

Building on that implementation experience, this brief captures key insights through a series of questions and answers with members of the project team. Their responses address common questions about fine-tuning considerations, implementation challenges, and lessons learned throughout the process. These insights can help technical leaders evaluate whether fine-tuning aligns with their organization's needs and guide implementation planning for those pursuing similar projects.

Before adopting any AI capabilities, organizations should consider several factors. The Data Integration Support Center (DISC) recommends having a clear understanding of the risks, governance, benefits, challenges, staffing, training, and specific use cases for the organization. These topics are covered in a previous brief titled "Building a Secure Generative Artificial Intelligence Environment for Research Use." Additionally, DISC offers support to public agencies in facilitating discussions and forming strategies to create a strong foundation for AI implementation.

Q What are the shortcomings of pretrained models?

A Like many technology solutions, pretrained models are designed to serve a **BROAD RANGE OF GENERAL PURPOSES**. Like the standard software packages organizations often start with, these models provide a foundation that works adequately for many basic needs, particularly in low-stakes applications. However, just as organizations frequently need to customize standard software for specialized needs, limitations become apparent when the pretrained model must handle complex tasks that require specialized knowledge or have significant consequences.

Our experience with assessment development illustrates these shortcomings. While the pretrained model met basic requirements, it **STRUGGLED WITH TASKS REQUIRING DEEPER CONTEXTUAL UNDERSTANDING**. This type of struggle stems from limitations in the model's training data, which often lack the specific content needed for specialized applications. When asked to perform tasks beyond its training data, the model may produce outputs that appear plausible but could contain misleading, inaccurate, or false inferences ("hallucinations").

Q Why should organizations consider fine-tuning a pretrained model?

A Fine-tuning offers organizations a powerful method for **ENHANCING MODEL PERFORMANCE FOR SPECIALIZED TASKS** (Radford et al., 2018). In our case, WestEd wanted more thoughtful, engaging reading passages, but the basic model did not meet our expectations. By introducing carefully selected examples to the model, particularly from content not available in public datasets, organizations can significantly improve the quality of the outputs and reduce problematic behaviors such as hallucinations. In our example, providing better examples of reading passages helped the model develop new, task-specific connections that improved its performance for our particular use cases.

Fine-tuning is particularly useful in the following circumstances:

- The task is complex and highly specialized, such as making accurate and engaging reading passages for specific grade-level ELA assessments.
- The task includes quality standards or output requirements, such as meeting specific evaluation criteria.
- The environment is controlled, such as a secure AI environment that limits what new information the model can access and where the data are stored.

Q What are the drawbacks of fine-tuning?

A First, fine-tuning requires **SIGNIFICANT INVESTMENT IN RESOURCES, INFRASTRUCTURE, AND EXPERTISE**. The fine-tuning process requires substantial computing power and storage capacity. Furthermore, models may need to be updated and retrained to maintain performance as new data become available or requirements change.

Second, fine-tuning carries **TECHNICAL RISKS**. The model may struggle to adapt to the new information. It may become overly specialized, performing well on the specifics but failing to generalize to new situations or prompts. Finding the right balance requires careful testing and validation.

Third, the success of fine-tuning depends on the **QUALITY OF THE TRAINING DATA**. If the fine-tuning data are not sufficiently representative of the target task, the process may yield limited improvements. Organizations must think carefully about whether they have the training data that will enable fine-tuning to meet their needs.

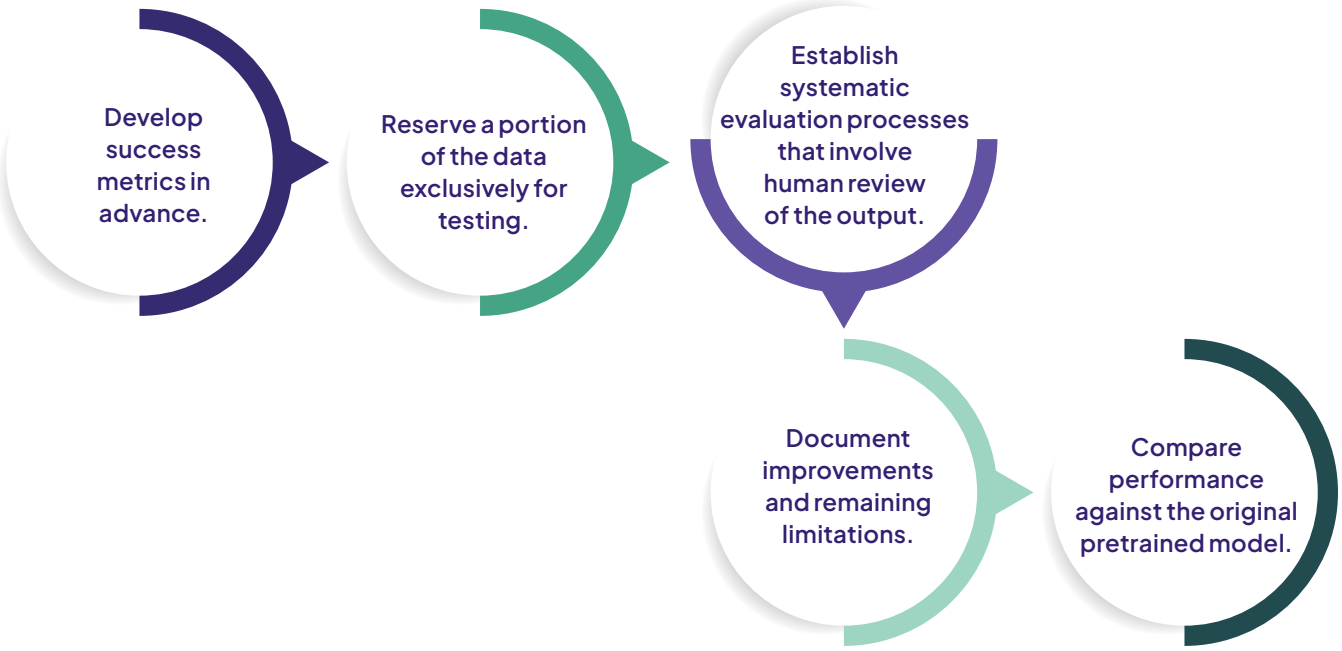
Fourth, fine-tuning with proprietary or sensitive data introduces **SECURITY RISKS**. Organizations must implement robust safeguards to protect the training data, control access, and manage how the model's outputs are used. Mitigating the risks often requires establishing a secure environment with the appropriate controls and monitoring.

Recognizing that these GenAI models **CONTINUOUSLY EVOLVE** is also important. Pretrained models' ability to generate accurate outputs should improve as the models are trained on more up-to-date data and the models' complexity increases. This improvement may make fine-tuning unnecessary. If fine-tuning does not yield the desired results today, it might work on other models or perhaps not be needed in future versions. Some would argue that large language models (LLMs) may be too large and cumbersome for specific tasks, preferring small language models trained on very specific examples as the best option for better performance.

Q Why is evaluating the model important even after fine-tuning?

A While fine-tuning often enhances the model, it LIKELY DOES NOT SOLVE ALL PROBLEMS. Evaluating the output helps determine whether the improved model is truly performing better and uncovers any new issues that may have been introduced. It also provides clarity on whether additional fine-tuning is needed to achieve the intended goals.

Effective evaluation requires careful preparation before fine-tuning. Organizations should follow a structured evaluation process, which will help them understand the true impact of their fine-tuning efforts and guide decisions about further model refinements:



Q What is the difference between training and testing?

A Fine-tuning requires splitting the available data into two sets: one for training the model and one for testing it. This separation is needed to evaluate whether the model has truly learned to generalize patterns rather than memorize specific examples. Without it, organizations have no way to tell if the model will perform reliably on new, similar tasks. This approach is similar to how teachers use one set of problems to teach a concept and a different set to check for understanding. Success is not about repeating known answers; it is about applying knowledge in new situations.

The optimal split between training and testing data depends on what data are available. In this assessment use case, we had 60 ELA passages available for fine-tuning and used an 80/20 split: 48 passages for training and 12 for testing. We had multiple passages aimed at different grade levels and types of passages (literary or informational). We randomly selected one passage within each grade level and passage type for the testing data. While a small test set is better than none, having 12 examples limited both the generalizability of the results and the precision of performance estimates.

The ultimate guide for establishing the training and testing datasets is to ensure that they represent excellent examples for the final use case of the GenAI. In the assessment context, this means including a range of passage types, grade levels, and levels of complexity that matches the demands of the task. If the training data are too narrow, such as including only literary texts, the model may not learn to generate strong informational passages. Likewise, if the test set does not reflect the variety the model will encounter, it cannot provide a reliable measure of performance. Assessing the training and testing data to understand the key characteristics and how they align with the desired usage is an important step before using these data for training. While random sampling often works well, the inherent randomness of sampling does not guarantee that any single sample will be representative.

Q Are there other approaches to improving model performance?

A If fine-tuning is not feasible due to cost, time, or data constraints, organizations can use other strategies.

Retrieval-augmented generation (RAG) is one alternative strategy (Lewis et al., 2020). While both RAG and fine-tuning aim to improve model outputs using additional knowledge, they work differently. Fine-tuning modifies the model by updating its parameters through additional training. RAG, on the other hand, keeps the model unchanged but provides it with access to external knowledge that has been indexed (i.e., tokenized, or converted into a searchable format) for quick retrieval during generation. Think of fine-tuning as teaching the model new skills, while RAG is more like giving the model a reference library to consult.

In our assessment development case, we tested RAG by providing exemplary passages as reference material. Unfortunately, this approach did not yield significant improvements in the passage quality compared to the base model. Simply retrieving examples may not have

Fine-Tuning
Modifies the model by updating its parameters through additional training

Few-Shot Learning
Similar to fine-tuning but requires only a handful of exemplars

Retrieval-Augmented Generation
Provides access to external knowledge that has been indexed

Prompt Engineering
Optimizes how organizations frame questions to the model

provided enough structure or guidance for the model to internalize and replicate these features. This result suggests that our use case demanded deeper model adaptation, which RAG alone was not able to support.

Another type of enhancement is few-shot learning. Few-shot learning is similar to fine-tuning but requires only a handful of strong examples. While it may not significantly improve the pretrained LLM, it can help the model focus on specific design features. For example, if the output must be a numbered list to be processed by another system, few-shot learning could provide the needed structure for the LLM to generate the output in a numbered list. Additionally, few-shot learning does not demand the storage or computational resources required for fine-tuning, making it a cost-effective solution when the pretrained LLM performs the task well.

Prompt engineering can complement other model customization strategies by optimizing how inputs are framed, regardless of how the model was trained or fine-tuned. Even when an organization uses fine-tuning to adapt a model to a specific task, effective prompt design can further improve results by guiding the model toward desired formats, tones, or structures.

In a setup that includes fine-tuning, few-shot learning, and RAG, prompt engineering plays a coordinating role. For example:

- With fine-tuning, prompt engineering ensures that the model's outputs remain consistent with how it was trained, especially when applied to slightly new tasks.
- With few-shot learning, prompts include a handful of examples that reinforce formatting, style, or domain-specific conventions without retraining the model.
- With RAG, prompts help the model know how to incorporate retrieved content appropriately—such as asking it to summarize, extract, or compare information from the context provided.

Each approach has distinct advantages and limitations. Prompt engineering is generally more accessible, few-shot learning can be cost-effective to learn specific output structures, and RAG enables knowledge integration without model modification. However, these approaches may not be as effective as fine-tuning for specialized applications such as our assessment development use case.

Q What ongoing maintenance is required for fine-tuning?

A We recommend **REGULARLY EVALUATING THE MODEL'S OUTPUT** to ensure that the model continues to perform as intended over time. If the task remains stable, such as extracting data from consistently formatted tables, additional updates may not be necessary. However, if the task changes, better training data become available, or newer base models offer improved capabilities, revisiting fine-tuning can help maintain or enhance performance. Ongoing monitoring also helps detect performance issues early and should be part of an organization's broader governance plan for managing AI systems.

Q How long did implementing the fine-tuning take? What was the estimated cost?

A The fine-tuning process timeline **DEPENDS PRIMARILY ON DATA PREPARATION** rather than the actual model training. In our experience, preparing high-quality training data—including collection, cleaning, formatting, and validation—took approximately 2 weeks of staff time, while the actual fine-tuning computation required only 3 to 5 hours.

Cost is determined by the **NUMBER OF TOKENS PROCESSED**, not simply the count of examples. Tokens are the text fragments the model reads internally (often parts of words). Most providers charge by token count, not file size or example count. For example, OpenAI's GPT-4o model currently costs $3.75 per million tokens for fine-tuning (as of February 28, 2025).

To estimate costs beforehand, an organization can tokenize its training data before submission. Our reading and writing assessment project with 48 examples converted to about 62,000 text units for the AI to process (tokens), which we ran through the system three times (three epochs), using approximately 186,000 text units total for training. This total resulted in a fine-tuning cost of around a dollar.

Of course, organizations should factor in not just the fine-tuning costs but also the staff time for data preparation and evaluation when budgeting for a fine-tuning project. Finally, since LLMs continue to evolve quickly, the return on investment for fine-tuning may change as newer or more efficient models become available.

Q What additional advice should organizations consider when exploring fine-tuning?

A

START WITH A CLEAR PROBLEM DEFINITION. Before investing in fine-tuning, thoroughly document where the organization's pretrained model falls short, and ensure that fine-tuning is the most appropriate solution. Not every performance issue requires fine-tuning to resolve. Often, starting with few-shot learning, prompt engineering, or a RAG pipeline before fine-tuning is cost-effective.

PREPARE THE DATA CAREFULLY. The quality of fine-tuning depends entirely on the training data. Invest time in collecting high-quality examples, ensuring that they reflect the intended use case, maintaining consistent formatting, establishing clear evaluation criteria, and setting aside enough data for testing.

BEGIN SMALL AND ITERATE. Starting with a modest set of examples allows organizations to test early and make adjustments before scaling up. However, starting too small may not yield meaningful improvements. As a general guideline, having around 50 high-quality, representative examples is often a useful starting point.

PLAN FOR THE FULL LIFECYCLE. Fine-tuning is not a one-time effort. Be prepared for ongoing evaluation, security monitoring, infrastructure costs, staff training, and future retraining as tasks evolve or better models become available.

LEARN MORE:
Technical Implementation Guidance

For detailed information about the technical process of implementing fine-tuning, see the companion brief titled "Fine-Tuning Models in a Secure AI Environment: Technical Implementation." This resource provides comprehensive guidance on the following issues:

- data preparation and formatting requirements
- training procedures and parameters
- testing and evaluation methodologies
- implementation challenges and solutions
- results and performance improvements

Organizations considering similar projects will find practical technical guidance to support their implementation planning.

How DISC Can Help

DISC at WestEd can facilitate productive conversations and assist in the development of an AI strategy for an organization's integrated data system. DISC offers technical assistance to public agencies free of cost, including the following services:

- providing expertise on the scope and variety of AI tools to support data integration efforts
- developing use cases that leverage AI tools and aligning those use cases to the state's strategic priorities
- conducting neutral, external assessments of the maturity of the integrated data system and its capacity to support and scale AI technologies
- facilitating structured discussions to develop a foundation for the use of AI in the integrated data system

Citations

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Suggested citation: LeBeau, B., Clarke, M., Rodriguez, B., & Quesen, S. (2025, June). *Fine-tuning models in a secure AI environment: Practical considerations.* WestEd.