disc
Data Integration Support Center
WestEd

2025
May

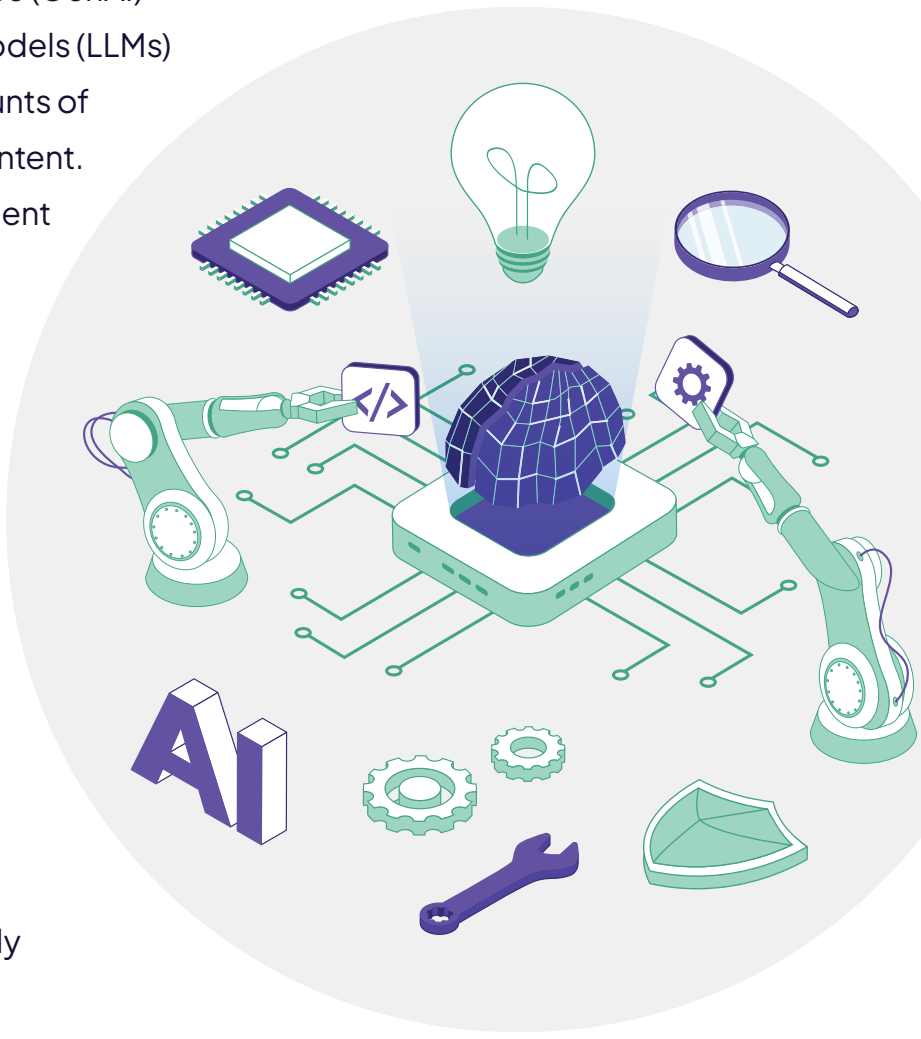
SPOTLIGHT

Fine-Tuning Models in a Secure AI Environment

TECHNICAL IMPLEMENTATION

AUTHORS:
Brandon LeBeau
Mitchell Clarke
Baron Rodriguez
Sarah Quesen

Modern generative artificial intelligence (GenAI) relies on pretrained large language models (LLMs) that learn associations from vast amounts of text to perform tasks and generate content. When public agencies want to implement AI capabilities, they typically start by evaluating commercially available models, such as those from OpenAI, Google, or Microsoft. These models are pretrained on open-access, publicly available information found on the web. Such off-the-shelf models often perform well on tasks for which information on the web is abundant. However, these models can be less reliable when handling specialized tasks, context-specific information, proprietary data, or rapidly changing information.



These limitations can result in several problematic behaviors that affect the quality and reliability of outputs, especially for public agencies:

HALLUCINATIONS occur when the model generates content or makes assertions that are nonexistent, nonsensical, or inaccurate, often by incorrectly combining or extrapolating from its training data.

For example, medical centers reported incidents in which an AI tool used to transcribe patient interactions invented fictitious medical details, risking misdiagnosis (Burke & Schellmann, 2024).

In another instance, an AI chatbot created to help small businesses repeatedly provided erroneous advice, including advice that would lead to the businesses breaking the law (Offenhartz, 2024).

REPEATING PROMPT INFORMATION occurs when the AI model fixates on specific phrases or elements from the input prompt, incorporating them repeatedly or too literally in its outputs.

For example, a senior technologist on the WestEd AI team was working on a document about a multi-state collaboration and created a prompt asking for a separate image of each state's geographical shape encompassing a unique state-specific feature. The image generated was an unrecognizable/unnatural version of each state. The technologist then simplified the prompt to ask for one state and one feature, but the morphed state outline consistently reappeared. The results created a feedback loop in which the AI tool repeatedly reinforced certain phrases or concepts from the prompt rather than generating novel, appropriate content. This behavior led to unusable outputs.

INACCURACY occurs when a model provides incorrect results or makes erroneous decisions because it was trained on data that are flawed, out of date, skewed, or incomplete. This behavior often manifests when the model learns and amplifies associations that perpetuate stereotypes or result in invalid conclusions that do not align with expert human judgment.

For example, child welfare agencies discovered that an AI tool identified a disproportionate number of Black children for mandatory investigations, conflicting with experienced social workers' assessments (Ho & Burke, 2022). The model had learned and amplified problematic associations, leading to systemically inaccurate risk predictions.

Public agencies may be able to mitigate these limitations using strategies such as fine-tuning.

This brief outlines the technical process used to fine-tune a pretrained model within WestEd's secure AI environment, which limited what information the model could access and controlled where data were stored, ensuring protection of sensitive information. The implementation details and results provide a foundation for understanding the fine-tuning process. For organizations thinking about their own fine-tuning projects, a companion brief titled "Fine-Tuning Models in a Secure AI Environment: Practical Considerations" addresses common questions about model selection, data requirements, evaluation approaches, and practical considerations to help guide decision-making.

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 companion 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.

WestEd's Use Case

1 Develop Initial Passages

The Assessment Research and Innovation (ARI) team at WestEd wanted to explore if they could develop text passages for an English language arts (ELA) assessment using GenAI. These passages had to be appropriate for the grade level and accurately assess ELA skills specified by grade-level standards. The ARI team also recognized that incorporating contexts that are relevant to students' backgrounds and experiences can effectively engage them in the assessment (Parsons & Taylor, 2011).

When assessments are developed to intentionally consider a range of students' backgrounds and experiences, researchers (e.g., Landl, 2021; Evans et al., 2021) claim that such assessments will be seen as more effective and are likely to have a positive impact on academic and social outcomes. For example, suppose a reading passage mentions an elevator. Many students will think of a machine that transports people between floors in a building, while some may envision a facility that stores grain. Similarly, when a reading passage describes ice fishing, students from warmer climates may struggle to visualize the activity, having never experienced frozen lakes substantial enough to support people.

Additionally, students can have difficulty maintaining engagement during assessments, especially when ELA text passages require several minutes to read, which affects their performance (Elleman & Oslund, 2019).

ELA passage development takes substantial amounts of time for content specialists. The work includes drafting a passage catered to standard-specific questions. For example, passages that include inference-type questions are written differently than passages that include questions that are meant to identify specific facts about the passage. In addition, cross-referencing and verifying facts to ensure that the passage is factually accurate can take significant time.

The ARI team expected that AI could help (Arslan et al., 2024), but they also wondered if it would have difficulties with this task. First, existing LLMs lacked exposure to assessment-specific content, as ELA passages used in secure standardized tests are not publicly available and thus were not part of the models' training data. Second, the requirement for context-rich passages represented a novel approach in assessment development, suggesting the models' training data might not include sufficient examples of this type of content. In fact, many assessment developers aim for "neutral" passages devoid of context.

2 Create Structured Prompts

To begin, the ARI team worked closely with the Assessment Design and Development team to develop a structured prompt framework for generating text passages for the ELA assessment. The prompt included the targeted grade level, word count, reading complexity, required content standards, and topic to consider. The initial focus was on informational ELA passages that could support specific grade-level standards and questions based on those standards. The GenAI had access to WestEd's passage and question writing standards developed by content experts. The GenAI also had access to commonly used state standards to aid in aligning passages to support these standards.

3 Establish Success Criteria

The team then established evaluation criteria to rate the quality of the model-generated passages. The criteria included factors such as the following:

- **technical requirements:** grade level-appropriate vocabulary and syntax, word count and Lexile text complexity statistics, alignment with content standard, and appropriate length and structure for assessment use
- **content quality:** accurate and verifiable facts, clear internal logic and coherence, and appropriate complexity and depth
- **content alignment:** support for matching the passage with questions aligned to specific ELA grade-level content standards
- **student engagement:** relevant context for the targeted student population, age-appropriate themes and scenarios, and clear connection to the student's real-world experiences and backgrounds

4 Evaluate Output

Next, with support from ELA content experts, the team evaluated the model's output against the criteria. The results revealed that the pretrained model performed adequately on many criteria but struggled with deeper contextual elements. The passages had many issues that prevented them from being used in an assessment. For example, the GenAI consistently produced text that was too difficult for the intended grade level and generated word counts that were excessively high. The ELA passages also contained awkward phrasing that did not sound human, and the parameters from the prompt were not always adhered to closely. Finally, in reviewing the informational content, errors or exaggerations were identified that would need to be edited manually. The team concluded that the model lacked exposure to the specific types of contextually rich assessment questions that are necessary for effective student engagement. This result highlights the importance of background and context in the pretraining phase needed to generate engaging assessment questions that can accurately measure the student's knowledge.

With this limitation in mind, the ARI team set out to fine-tune the model using WestEd's secure AI environment. The secure environment offered one critical advantage: It allowed the team to incorporate proprietary assessment content into the training process without risk of exposure.

5 Fine-Tune

A few key staff from the assessment team and WestEd's information technology team played important roles in fine-tuning, including the following:

- An AI specialist was responsible for the technical process of fine-tuning. This work included formatting training data, preparing prompts and responses for fine-tuning, and managing other data processing. This specialist also understood model parameters and training requirements and implemented the fine-tuning process.
- Assessment experts evaluated and validated training examples, ensured the quality of the fine-tuning data, and tested the model's outputs.
- Content experts evaluated and validated the output from a content perspective to support the writing of questions for specific grade-level standards.

To fine-tune the model, the key staff followed a precise technical protocol for the data preparation and submission. For Open AI's GPT model, training data must be formatted as JSONL (JavaScript Object Notation Lines), with each line containing three critical components:

- a system message that defines the model's goal
- an example input prompt that represents a typical user request
- the model's desired response that demonstrates the correct output

Each JSONL line represents a complete training example, teaching the model proper response patterns. While models can begin showing improvement with as few as 10 well-crafted examples, significant enhancements are typically observed with approximately 50 examples. The number needed for fine-tuning depends on the task's complexity and the output's desired specificity. For this assessment development use case, the team prepared 48 carefully curated training examples for fine-tuning the model.

6 Compare Results

Once the fine-tuning was done, the team conducted a comprehensive evaluation comparing the original model outputs with those of the fine-tuned model. Testing revealed substantial improvements in three critical areas. First, the fine-tuned model generated passages with the potential for questions aligned to grade-level standards. Second, the passages aligned better with grade-level technical requirements, including appropriate word counts and Lexile text complexity statistics. Third, the generated text more closely matched the desired format for the specific use case for content developers.

These results demonstrated that targeted fine-tuning could effectively adapt a general-purpose model for specialized purposes, such as assessment content creation. Although the fine-tuned model performed better in some aspects, it still had difficulty incorporating accurate student contexts in the final output. Additional data may be needed to aid the model in producing accurate final text predictions.

Training Examples
Used to teach the model

Testing Examples
Used to evaluate the model

LEARN MORE:
Practical Implementation Guidance

For detailed responses to common questions about implementing fine-tuning, see the companion brief titled "Fine-Tuning Models in a Secure AI Environment: Practical Considerations." This resource includes expert answers to questions such as the following:

- What are the shortcomings of pretrained models?
- Why should organizations consider fine-tuning a pretrained model?
- What are the drawbacks of fine-tuning?
- Why is evaluating the model important even after fine-tuning?
- What is the difference between training and testing?
- Besides fine-tuning, are there other approaches to improving model performance?
- What ongoing maintenance is required for fine-tuning?
- How long did implementing the fine-tuning take? What was the estimated cost?
- What additional advice should agencies consider when exploring the fine-tuning process?

Organizations considering similar projects will find insights to support their decision-making process and 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 with 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

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence, 7*. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1460651/full

Burke, G., & Schellmann, H. (2024, October 26). *Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said*. AP News. https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14

Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences, 6*(1), 3–11. https://journals.sagepub.com/doi/abs/10.1177/2372732218816339

Evans, C. M., Landl, E., & Thompson, J. (2021). What do we know about the implementation and outcomes of personalized, competency-based learning? A synthesis of research from 2000 to 2019. *CompetencyWorks Blog*. Aurora Institute. https://aurora-institute.org/cw_post/what-do-we-know-about-the-implementation-and-outcomes-of-personalized-competency-based-learning-a-synthesis-of-research-from-2000-to-2019/

Ho, S., & Burke, G. (2022, April 29). *An algorithm that screens for child neglect raises concerns*. AP News. https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1

Landl, E. (2021, August 25). *Ensuring assessment systems meet the needs of students with disabilities*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/blog/ensuring-assessment-systems-meet-the-needs-of-students-with-disabilities/

Offenhartz, J. (2024, April 3). *NYC's AI chatbot was caught telling businesses to break the law. The city isn't taking it down*. AP News. https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21

Parsons, J., & Taylor, L. (2011). Improving student engagement. *Current Issues in Education, 14*(1).

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