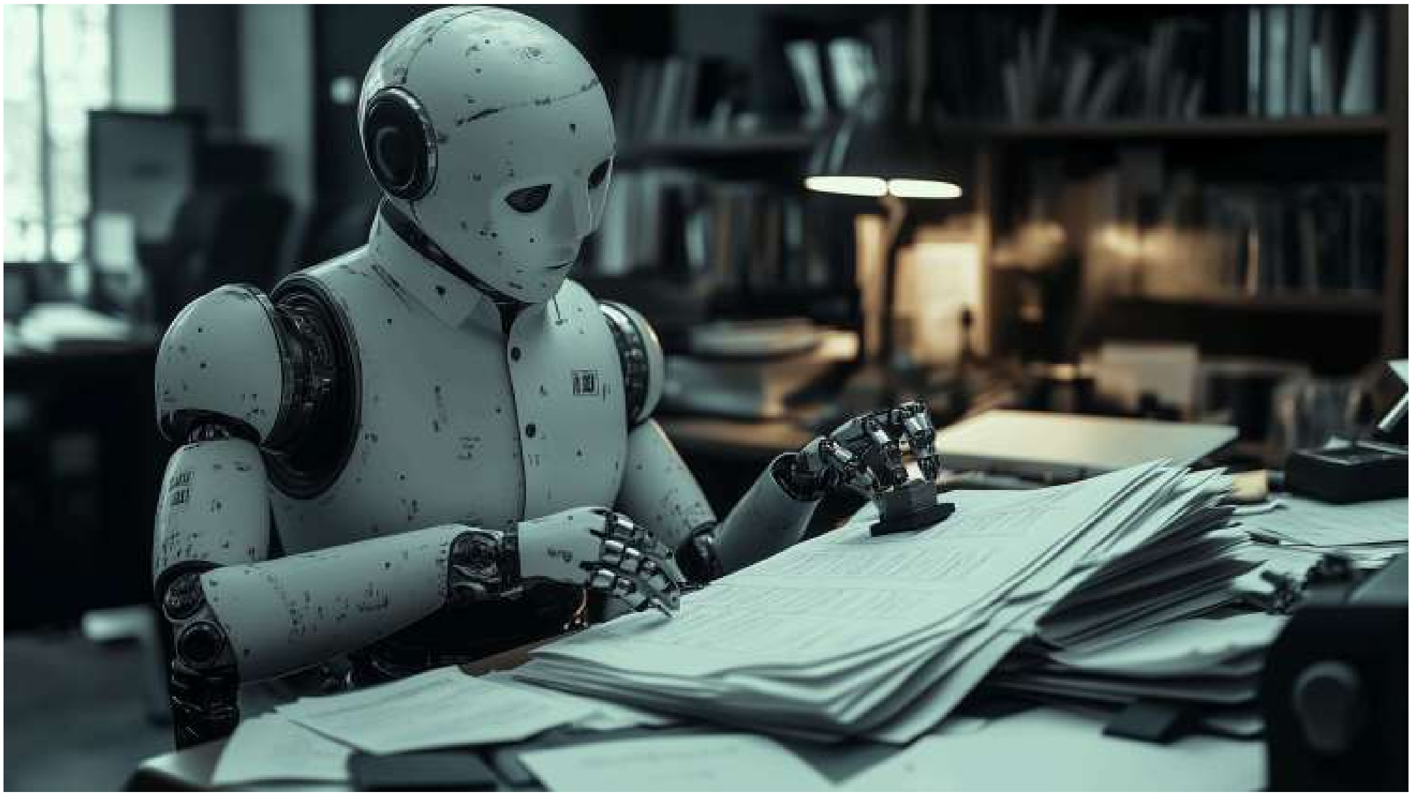
VentureBeat

# Fine-tuning vs. in-context learning: New research guides better LLM customization for real-world tasks

Ben Dickson

@BenDee983

May 9, 2025 5:23 PM



Credit: VentureBeat made with Midjourney

Two popular approaches for customizing large language models (LLMs) for downstream tasks are fine-tuning and in-context learning (ICL). In a recent study, researchers at Google DeepMind and Stanford University explored the generalization capabilities of these two methods. They find that ICL has greater generalization ability (though it comes at a higher computation cost during inference). They also propose a novel approach to get the best of both worlds.

The findings can help developers make crucial decisions when building LLM applications for their bespoke enterprise data.

## Testing how language models learn new tricks

Fine-tuning involves taking a pre-trained LLM and further training it on a smaller, specialized dataset. This adjusts the model's internal parameters to teach it new knowledge or skills. In-context learning (ICL), on the other hand, doesn't change the model's underlying parameters. Instead, it guides the LLM by providing examples of the desired task directly within the input prompt. The model then uses these examples to figure out how to handle a new, similar query.

The researchers set out to rigorously compare how well models generalize to new tasks using these two methods. They constructed "controlled synthetic datasets of factual knowledge" with complex, self-consistent structures, like imaginary family trees or hierarchies of fictional concepts.

To ensure they were testing the model's ability to learn new information, they replaced all nouns, adjectives, and verbs with nonsense terms, avoiding any overlap with the data the LLMs might have encountered during pre-training.

The models were then tested on various generalization challenges. For instance, one test involved **simple reversals**. If a model was trained that “femp are more dangerous than glon,” could it correctly infer that “glon are less dangerous than femp”? Another test focused on **simple syllogisms**, a form of logical deduction. If told “All glon are yomp” and “All troff are glon,” could the model deduce that “All troff are yomp”? They also used a more complex “semantic structure benchmark” with a richer hierarchy of these made-up facts to test more nuanced understanding.

“Our results are focused primarily on settings about how models generalize to deductions and reversals from fine-tuning on novel knowledge structures, with clear implications for situations when fine-tuning is used to adapt a model to company-specific and proprietary information,” Andrew Lampinen, Research Scientist at Google DeepMind and lead author of the paper, told VentureBeat.

To evaluate performance, the researchers fine-tuned Gemini 1.5 Flash on these datasets. For ICL, they fed the entire training dataset (or large subsets) as context to an instruction-tuned model before posing the test questions.

The results consistently showed that, in data-matched settings, ICL led to better generalization than standard fine-tuning. Models using ICL were generally better at tasks like reversing relationships or making logical deductions from the provided context. Pre-trained models, without fine-tuning or ICL, performed poorly, indicating the novelty of the test data.

“One of the main trade-offs to consider is that, whilst ICL doesn’t require fine-tuning (which saves the training costs), it is generally more computationally expensive with each use, since it requires providing additional context to the model,” Lampinen said. “On the other hand, ICL tends to generalize better for the datasets and models that we evaluated.”

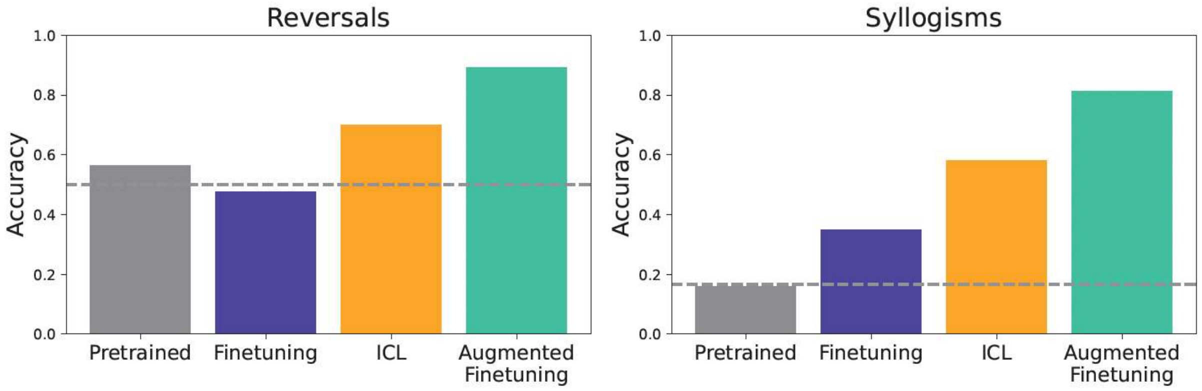
## A hybrid approach: Augmenting fine-tuning

Building on the observation that ICL excels at flexible generalization, the researchers proposed a new method to enhance fine-tuning: adding in-context inferences to fine-tuning data. The core idea is to use the LLM’s own ICL capabilities to generate more diverse and richly inferred examples, and then add these augmented examples to the dataset used for fine-tuning.

They explored two main data augmentation strategies:

1. A **local strategy**: This approach focuses on individual pieces of information. The LLM is prompted to rephrase single sentences from the training data or draw direct inferences from them, such as generating reversals.
2. A **global strategy**: The LLM is given the full training dataset as context, then prompted to generate inferences by linking a particular document or fact with the rest of the provided information, leading to a longer reasoning trace of relevant inferences.

When the models were fine-tuned on these augmented datasets, the gains were significant. This augmented fine-tuning significantly improved generalization, outperforming not only standard fine-tuning but also plain ICL.



“For example, if one of the company documents says ‘XYZ is an internal tool for analyzing data,’ our results suggest that ICL and augmented finetuning will be more effective at enabling the model to answer related questions like ‘What internal tools for data analysis exist?’” Lampinen said.

This approach offers a compelling path forward for enterprises. By investing in creating these ICL-augmented datasets, developers can build fine-tuned models that exhibit stronger generalization capabilities.

This can lead to more robust and reliable LLM applications that perform better on diverse, real-world inputs without incurring the continuous inference-time costs associated with large in-context prompts.

"Augmented fine-tuning will generally make the model fine-tuning process more expensive, because it requires an additional step of ICL to augment the data, followed by fine-tuning," Lampinen said. "Whether that additional cost is merited by the improved generalization will depend on the specific use case. However, it is computationally cheaper than applying ICL every time the model is used, when amortized over many uses of the model."

While Lampinen noted that further research is needed to see how the components they studied interact in different settings, he added that their findings indicate that developers may want to consider exploring augmented fine-tuning in cases where they see inadequate performance from fine-tuning alone.

"Ultimately, we hope this work will contribute to the science of understanding learning and generalization in foundation models, and the practicalities of adapting them to downstream tasks," Lampinen said.

Press Releases Contact Us Advertise Share a News Tip