Deliver better large language model performance

Customize a pre-trained model with your own data on a full-stack solution for Generative Artificial Intelligence (GenAI) large language models

Why customize a pre-trained model?

One of the strengths of large language models (LLMs) is they contain a broad amount of information and knowledge, thanks to the substantial amount of text data used to train them. However, this also means the models often struggle to maintain accuracy on topics or items that were not used within the initial training dataset, which is why it's so important to fine-tune models with your own proprietary data.

Pre-trained model customization is the process of retraining an existing generative AI model for task-specific or domain-specific use cases. This is often more efficient than to train the model from scratch on a new dataset.

This layered training approach—in which specialized information is added to a pre-trained model—is called transfer learning or model fine-tuning. This creates application-specific parameters on top of pre-trained LLMs, with the purpose of making the models perform better.

One of the main challenges for organizations when using public GenAI applications (like ChatGPT) is that the value is limited for organizational use, as these models aren't built or paired with your own business data.

The way forward to maximize the value of these pre-trained large language models (LLMs) is to combine and train on your own domain-specific data, based on the use case. The goal through the customization process is to create interfaces between the models and your downstream applications to make them even more powerful.

Over 340k engineering hours spent on design, development and validation on GenAI solutions[1]

- Create new value with a secure infrastructure for your business-critical operations
- Deploy an improved workload experience
- Make your data more valuable through customization and tuning
- Lower costs around optimization activities with proven guidance

Reduce time-to-results with a proven solution and professional services

Quickly build on-premises infrastructure for your application needs using pre-tested solutions made to simplify adoption. This proven approach helps reduce the complexity and risk of implementation, so you can accelerate the customization of pre-trained models based on specific use cases.

And with Dell Services experts, you can realize the value of GenAI for your data more quickly with a portfolio of services to assist you at every stage of your journey – from planning to scaling your applications, with managed services to help where needed.

Example of customization process: using one, or multiple, LLMs (in a secure on-premises environment) to either:

Learn more

- See Design Guide
- Efficient Fine-tuning Using Low-Rank Adaptation (LoRA) on Single GPU

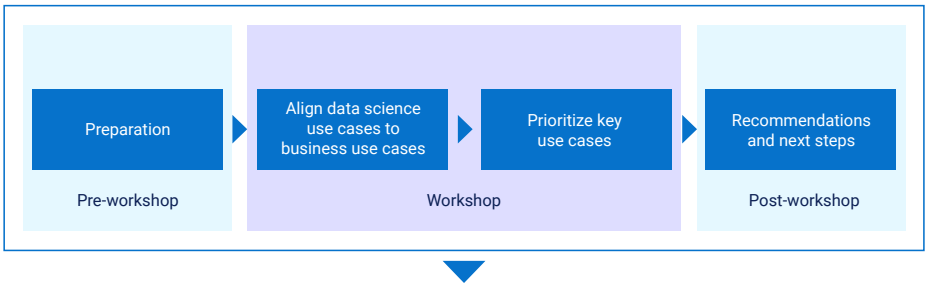
What are common customization techniques?

Prompt learning focuses on crafting effective input prompts to elicit desired responses from the LLM. It involves experimenting with different prompts and refining them based on the model's responses to improve its performance on specific tasks.

Prompt tuning refines prompts through an iterative process to achieve better task-specific performance. Prompts are adjusted based on model-generated outputs, and the process is repeated until the desired results are achieved.

P-tuning (parameter tuning) combines prompt engineering with fine-tuning to further customize the LLM. It involves both adjusting prompts and fine-tuning the model on task-specific data to achieve optimal performance.

Transfer learning leverages knowledge gained from pre-training on one task or domain to enhance performance on another related task or domain. The model is pre-trained on a diverse dataset, and then fine-tuned on a smaller, task-specific dataset, allowing it to transfer learned features to the target task.



1. fine-tune a LLM on proprietary domain-specific data
2. or use a language model to further refine proprietary domain-specific data for use in downstream LLM fine-tuning.

Technical Specifications

The Validated Design configurations are based on the newest, AI-acceleration-optimized Dell PowerEdge XE and rack servers, leveraging the latest NVIDIA GPUs and NVIDIA AI Enterprise, with Triton Inference Server and the NeMo framework. Fast, ample data lake storage for GenAI and LLMs is provided by Dell PowerScale all-flash or hybrid storage arrays.

Compute	Networking	Software
• PowerEdge XE9680 server equipped with eight NVIDIA H100 SXM GPUs with NVSwitch • PowerEdge XE8640 server equipped with four NVIDIA H100 SXM GPUs with NVLink • PowerEdge R760xa servers supporting up to four NVIDIA L40S PCIe GPUs • Management: PowerEdge R660 servers	NVIDIA Networking, Dell PowerSwitch S5232F-ON or S5248F-ON **Storage** Supported by Dell PowerScale F600, F710, F900 storage	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise with NeMo Framework for LLMs and Triton Inference Server; NVIDIA Base Command Manager Essentials

Dell Technologies and NVIDIA

Dell Technologies and NVIDIA work together to enable and accelerate GenAI adoption, deliver engineering-validated hardware and software to accelerate AI, ML and DL workloads to meet customer needs across all businesses and verticals. With this Validated Design for LLM customization, you can accelerate your digital transformation with solutions optimized for rapid time to value from your AI initiatives.

Learn more about Dell solutions

Contact a Dell Technologies Expert

View more resources

Join the conversation with #PowerEdge @DellTech

[1] Based on internal analysis, October 2023

DELL Technologies

NVIDIA