Deliver better large language model performance

Dell Generative AI Solutions delivers everything you need for Large Language Model Customization

Public GenAI applications (like ChatGPT) and their value is limited for your own organizational use, as these models aren't built or paired with your own business data.

They can help in some areas, but to maximize the value of these pre-trained large language models (LLMs) they need to be combined and trained 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.

- Deploy an improved workload experience
- Make your data more valuable through customization and tuning
- Lower costs around optimization activities with proven guidance

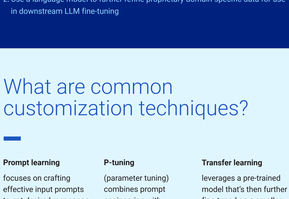
Why customize a pre-trained model?

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 GenAI model for task-specific or domain-specific use cases. This layered training approach—in which specialized information is added to a pre-trained model—is called transfer learning or model fine-tuning.

This process creates application-specific parameters on top of pre-trained large language models, with the purpose of making the models perform better.



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

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

What are common customization techniques?

Prompt learning

focuses on crafting effective input prompts to get desired responses from the LLM. It involves experimenting and refining prompts based on the model's responses to improve its performance

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 a pre-trained model that's then further fine-tuned on a smaller, task-specific dataset, allowing it to transfer learned features to the target task

Dell Validated Design for Generative AI with NVIDIA - Model Customization

Accelerate deployment and reduce risk with pre-tested and proven solutions designed to help you avoid design, planning, and adoption pitfalls. Components can be mixed and matched, and independently scaled depending on your application needs.

Generative AI Framework

Conversational AI: NVIDIA NeMo

End-to-end enterprise framework for developers to build, customize, and deploy generative AI models with billions of parameters.

AI Ops and ML Ops Platforms

NVIDIA AI Enterprise

Partner AI Ops software for a smooth end-user experience, including interactive notebooks, experiment management, pipelines, and more.

Software Infrastructure

NVIDIA Base Command Manager Essentials

Orchestration & scheduling layer for running AI training, including multi-node jobs, and scaling inference.

Infrastructure Management

OpenManage Enterprise OneFS | CloudIQ

Familiar Dell management tools delivering proactive monitoring and predictive analytics simplifying infrastructure operations.

Hardware Infrastructure

Dell Servers, Networking, Storage | NVIDIA GPUs

Compute with Accelerators
Dell PowerEdge XE9680 and PowerEdge XE8640 servers with NVIDIA H100 GPUs

Networking
Dell PowerSwitch S5232F-ON or S5248F-ON

Storage
Support for Dell PowerScale, ECS, and ObjectScale

Deliver outcomes faster with our help

Dell Services experts can assist you at every stage of your GenAI journey:

Strategize	Implement	Adopt	Scale
Build your roadmap to achieve the innovation objectives of your IT and business stakeholders	Establish your platform, leveraging Dell Validated Designs to implement GenAI inferencing hardware and software	Accelerate the value of your use cases by implementing a pre-trained inferencing model	Manage your innovation portfolio with resident technical experts and training offers to develop the GenAI skills of your team

Dell Technologies and NVIDIA

Dell Technologies and NVIDIA work together to enable and accelerate GenAI workloads, 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

https://www.dell.com/ai
https://www.dell.com/en-us/lp/contact-us
https://infohub.delltechnologies.com/