TREDENCE

Beyond Possible

The Acceleration Advantage:

Why Enterprise Teams Should Race to Develop LLMOps Capabilities

Ravindra Patil
Senior Director, Data Science

Enterprise leaders have captured the vision of how generative AI can transform their businesses. When ChatGPT was released in late 2022, chief data officers, chief artificial intelligence (AI) officers, and heads of data science immediately began brainstorming on how to apply generative AI to their business.

Over **70%** of all organizations are experimenting with large-language model (LLM) capabilities.[1] Top use cases today include summarizing text, analyzing and structuring input, generating content, and answering questions in chat, among others.

However, comparatively few have moved beyond experimentation to deploy generative AI in their operations. Why is this? Teams are still striving to gain experience with this new technology while mitigating data privacy risks, security gaps, model hallucination and bias.

Yet, the most significant risk is that enterprises will get mired in endless experimentation and piloting while other organizations set up the structure, they need to accelerate with LLMs. If this occurs, the gap between the haves and have-nots will rapidly increase as AI leaders deploy task- and domain-specific models, extend them across use cases and business functions, and redesign processes at scale.

Generative AI: Who's Piloting and Who's Productionizing Models

Most enterprises have taken initial steps with generative AI and have the budgets and teams to explore this technology.

Nearly 3 in 4 72.8% of all companies globally have experimented with generative AI, but only 18.2% of this group is already implementing it.[2]

4 in 10 41.2% of teams have the budget and resources to create value and drive AI transformation and generative AI across their organizations. [3]

Nearly all 87.7% of organizations want to standardize on a single AI/ML platform. [4]

Only 1 in 10 Organizations have implemented a formal policy governing employees' generative AI use,[5] which could increase data privacy and security risks.

Choosing different options LLM options today include widely used models provided by OpenAI, Meta, and Anthropic; those developed by cloud providers; closed machine learning (ML) models; and proprietary LLMs built by enterprises.

1 "More than 70% of companies are experimenting with generative AI, but few are willing to commit more spending," online article VentureBeat, https://venturebeat.com/ai/more-than-70-of-companies-are-experimenting-with-generative-ai-but-few-are-willing-to-commit-more-spending/

2 VentureBeat, ibid.

3 Enterprise Generative AI Adoption, report, AI Infrastructure Alliance, page 9, August 2023, https://ai-infrastructure.org/enterprise-generative-ai-adoption-report-aug-2023/

4 Ibid, page 8.

5 AI Policies are Low, Use is High, and Adversaries are Taking Advantage, Says New AI Study," press release, ISACA, October 25, 2023, https://www.isaca.org/about-us/newsroom/press-releases/2023/ai-policies-are-low-use-is-high-and-adversaries-are-taking-advantage-says-new-ai-study

LLMOps = MLOps + Specialized Processes

Developing a large-language model (LLMOps) operation in action is essential if enterprise teams are to develop, productionize, and deploy generative AI models and scale them across roles, business functions, and regions. LLMOps processes include traditional MLOps processes, such as:

Designing models

Using LLMOps processes and best practices, teams track experiments, register and archive models, and tune hyperparameters.

Testing models

Teams test models to ensure accuracy, deliver desired performance, and use the best foundational models.

Deploying models

Standardized model deployment processes, such as implementing CI/CD versioning, using PaaS and IaaS infrastructure, and leveraging container and release management, make it easier, faster, and cheaper to deploy models.

Monitoring models

Teams then monitor models to ensure they meet business KPIs and are accurate while responding to alerts and notifications about performance anomalies.

However, LLMOps processes are needed to address the unique engineering, architecture, security, ethical, and operational challenges of productionizing generative AI solutions. They include:

Creating a scalable infrastructure and architecture

Teams must set up infrastructure, cloud, and data layers to build their generative AI models. Modules require training, finetuning, and contextualization before they can be used.

Protecting data and models

Setting up models in closed, protected environments on-premises or in the cloud enables teams to enforce controls and mitigate data privacy risks, protecting data and models against misuse, manipulation, and theft.

Leveraging new skills

Teams of LLMOps engineers and architects, prompt engineers, and SecMLOps experts work closely with business stakeholders to gather use case requirements; propose the best foundational models; and develop, deploy, and manage models.

LLMOps processes are designed to ensure the optimal performance, scalability, and efficiency of LLM models so that they can accomplish business goals. These goals include personalizing the customer experience, driving workforce productivity, increasing innovation, and gaining speed to market, among others.

Up to

40%

Savings

That enterprises can achieve when they use LLMOps processes that include effective resource provisioning, machine learning (ML) workspace management, and cost controls. In addition, enterprises can reduce API calls by up to 60%.

What's Holding Enterprises Back from Deploying Generative AI Models

So, what's holding enterprises back from productionizing models?

Setting up data and LLM processes

Enterprise teams need help selecting the suitable foundational models for key use cases; implementing frameworks; designing new architectures; storing and securing data; developing, managing, and monitoring models; and tuning models to improve accuracy. To manage models, they must acquire or develop new skills and implement processes, such as engineering prompts, fine-tuning models, and engineering agents and applications. In addition, teams will want to monitor and optimize models to ensure they are available; meet latency objectives; and deliver relevant, high-quality outputs.

While enterprise teams can build these capabilities, they can also work with partners who provide model development, deployment, and monitoring. Partners also offer domain-specific models that can be deployed quickly, providing rapid value.

Ensuring compliance and security

OWASP has identified the top 10 LLM application risks as prompt injection, insecure output handling, training data poisoning, model denial of service, supply chain vulnerabilities, sensitive information discovery, insecure plugin design, LLMs with excessive agency, overreliance on LLMs, and model theft.[6]

Enterprises can mitigate these risks and threats by implementing frameworks with appropriate governance and security guardrails, including quality checks. While some jobs may be able to be fully automated, others will require review by trained humans.

Trusting output

Generative AI challenges, which include hallucination, bias, data privacy, and security, have been well-documented in the press. Enterprises can address these issues by implementing frameworks with security and governance to benefit from generative AI's ability to create software, text, and images at scale while controlling the quality of outputs and protecting enterprise data.

Managing model time and costs

Without LLMOps, model training and deployment costs can quickly skyrocket. Oversizing LLMs leads to unnecessary spending as jump factors increase significantly from one model to the next. For example, the cost to train GPT-3XL is currently around $2,500 or less than $2 per parameter, while the cost to train GPT175B, which is multiple models ahead, is almost $1M or more than $62 per parameter.[7]

6 OWASP Top 10 for LLMS Applications, v1.1, report, page 4, October 16, 2023, https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf

7 Internal Tredence presentation

How to Progress Against the LLMOps Journey

Enterprise teams can use this four-step roadmap to adopt LLMOps and increase their maturity with generative AI.

1 Create an LLMOps landing zone

Enterprise teams will work to understand LLM capabilities and infrastructure by creating a business case for new generative AI capabilities, designing initial prompts, and pre-selecting foundational models. Next, they should set up infrastructure, storage, serving, and security. Then, they should test the foundational models and log results to see how they perform against cost, speed, precision, security, and scalability goals.

2 Create repeatable LLMOps processes

After developing a repository of foundational models, teams will do prompt engineering or fine-tune foundational models to improve outputs. Then, they can train teams on new processes and set up a serving architecture. They'll chain prompts to automate processes, register and version applications, log data inputs and outputs, and analyze and filter results. Teams will also develop rating mechanisms to validate output quality and perform security and privacy checks.

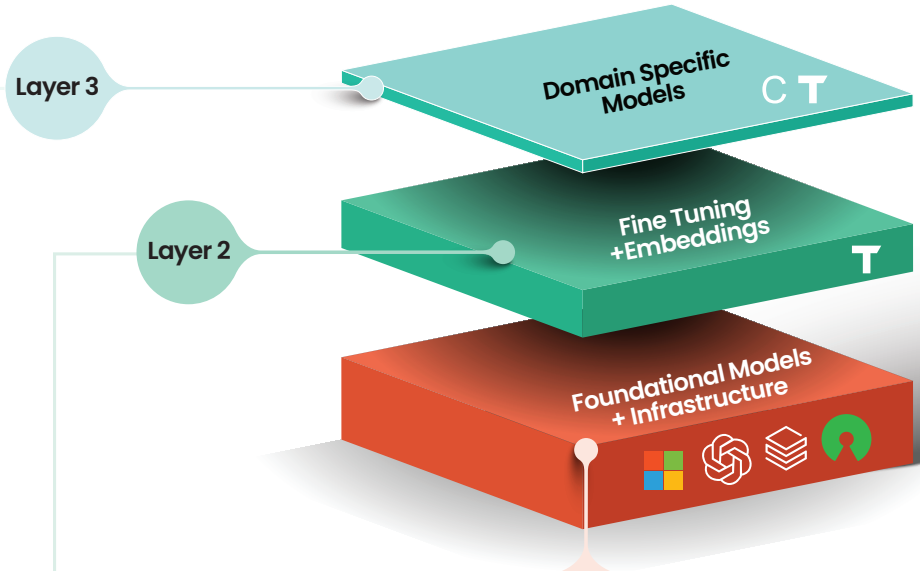
3 Develop reliable LLMOps processes

Next, the goal is to fine-tune personalized LLMs. Teams will manage data dependencies and LLM organizational templates. They'll store and log prompts, test prompt lineage, and transfer learning by reusing models to solve new problems. They will also set up pipelines and automated processes to scale model deployment; standardize serving processes, including API testing, rating, and reliability; and establish tradeoff controls to balance performance and cost objectives. Finally, teams will monitor and analyze to identify and mitigate risks.

4 Scale LLMOps capabilities

By now, teams have developed industrialized generative AI capabilities. They have developed an organizational repository of fine-tuned LLMs and have automated the control flow, agents, and tools for training and serving pipelines. Prompt engineering is well-managed and is enhanced by vector databases and prompt augmentation. Model outputs are robust, reliable, and trusted by users and are continually optimized with monitoring and automated retuning and build capabilities. The organization's LLM governance function uses codified AI principles to ensure models are free from hallucination and bias; ensures data security and privacy; and maintains compliance with data regulations and emerging LLM regulations.

What Enterprise LLMOps Capabilities Look Like



Layer 1: Foundational models + Infrastructure

- Storage and selection criteria of foundational models by use cases.
- Manage security: access control, log activities, etc
- Model lineage: tracks the usage of foundational models. in specific scenarios, such as fine-tuning for particular solution domains.
- Vulnerability and data dependencies management.
- Inclusion and decommission of foundational models

Microsoft databricks OpenAI

Layer 2: Fine Tuning +Prompt eng +Transf learning & Fed learning mgmt

- Fine-tuning techniques, distillation, and quantization.
- Prompting, prompt change management, prompt evaluation, prompt versioning
- Architecture req & design: GPU, CPU or Mixed, distributed, parallel, etc.
- Trans. Learning selection criteria
- Federate learning management
- Testing, optimizing against attacks and response filtering

Layer 3: Domain Specific use cases + Serving +Monitoring + Retraining

- Serving: architecture planning, implementation and optimization(e.g., Mixture of Experts-MoE); load, latency, and API testing;
- Retraining criteria and automation with fine-tuning strategy.
- Monitoring
- Control flow setup and monitoring
- Vector storage management
- Techstack and Framework selection

Most common Business demands

Caption: While foundation models can be used to optimize tasks, most enterprises will create domain-specific models that require fine-tuning to solve specific industry challenges, such as retail content personalization.

Benefits of Implementing an LLMOps Function

Setting up an LLM structure that will scale across use cases, business functions, and regions delivers significant benefits, including:

Support selecting foundation models

Enterprise teams can work with partners to select the most suitable foundation models for key use cases, ensuring that model parameters, token span, pricing, data training cutoff, and optimization techniques meet their needs.

Managing data and context windows

More powerful GPTs offer larger context windows. For example, GPT-4 Turbo offers a window size of 128K tokens or about 300 book pages. However, tokenizing text and encoding and embedding output at this scale takes extensive processing power – and costs. LLMOps processes will set prompt and response token limits to match jobs to the desired output.

Optimizing downstream tasks and deployment

With LLMOps processes, teams gain standardized, automated processes that are easy to repeat and scale. This enables teams to focus on innovating and generating business value by deploying new capabilities.

Ensuring effective security and governance

By developing reliable and scalable LLMOps processes, teams mitigate data and model risks.

Accelerating time to market

Industrializing LLM capabilities and launching new models in days to weeks is critical to achieving business transformation goals.

Reducing costs

With well-controlled processes, enterprises avoid run-away model training and deployment costs, ensuring they receive ROI on new initiatives.

Case Study: How a Top-Three CPG Firm Set Up LLMOps Capabilities

A global CPG firm sought to use LLMs to automatically summarize documents, making content findings more readily available to teams. Offering this capability globally would speed users' time to insight and drive organizational productivity.

As data and AI teams experimented with LLMs, they experienced numerous challenges, including disorganized models, growing costs, a lack of governance, and an inability to scale. In addition, users didn't trust model outputs due to their hallucinations and security vulnerabilities.

After implementing LLMOps processes, the CPG firm was able to:

Optimize LLM infrastructure

Teams could adjust instance types based on monitoring results, reducing costs by 15%.

Simplify model design

By harnessing a modular and parameterized approach, the CPG firm was able to simplify and refactor the overall LLM model design.

Leverage best practices

Teams now have best practices to use across every execution checkpoint, including data quality, orchestration, and DevOps processes. As a result, code bugs have been reduced by 20%, and execution times have decreased by 80%.

Gain versatile metrics

The team can flexibly leverage evaluation metrics, such as accuracy and semantic similarity, to determine output effectiveness.

Automate pipelines

The teams have automated all execution and data validation pipelines.

Improve governance

The LLM program is well-controlled with robust governance and security measures.

Deploy LLMOps to Speed New Generative AI Capabilities. Talk to Tredence.

Generative AI applications offer great promise, but only if you address LLM complexity and costs by implementing effective LLMOps capabilities. They can help you speed and scale LLM processes while improving model performance, output quality, and costs.

When you work with Tredence, you:

Gain end-to-end services

Tredence provides generative AI enterprise and technology strategy development, functional expertise, human-centric design, industrialized technology delivery, platform engineering and implementation, managed services, change management and adoption, and governance and compliance services, all packaged in easily consumable LLMOps services. Engineering models, frameworks, and delivery will enable you to optimize performance and cost.

Streamline the building of proprietary models

We use LLMOps processes to build generative AI models on LLMs. We provide our intellectual property (IP) to train and finetune your models while your teams use your IP to contextualize models to key use cases.

Access domain-specific models

Tredence provides prebuilt, pre-trained models that speed time to value. You can develop their embeddings and store the model in protected environments. We provide model training and finetuning while you use your IP to contextualize models to your use cases.

Capitalize On This No-Risk Offer

Ready to get started?

Contact us to schedule a 60-minute discovery call, where we will learn about your needs; discuss how to deploy LLMOps capabilities at your company; and quantify the value you will achieve by implementing standardized, automated processes.

Seize the acceleration advantage. Move forward now before your peers and competitors to galvanize growth and reap new ROI from generative AI capabilities.

About Tredence Inc.

Tredence is a global data science solutions provider focused on solving the last mile problem in AI. The 'last mile' is the gap between insight creation and value realization. Tredence is a Great Place to Work-Certified and as a 'Leader' in the Forrester Wave: Customer Analytics Services. Tredence is 2000+ employees strong with offices in San Jose, FosterCity, Chicago, London, Toronto, and Bangalore, with the largest companies in retail, CPG, hi-tech, telecom, healthcare, travel, and industrials as clients.

Want to know more about us?

Please visit: www.tredence.com

Follow us at: