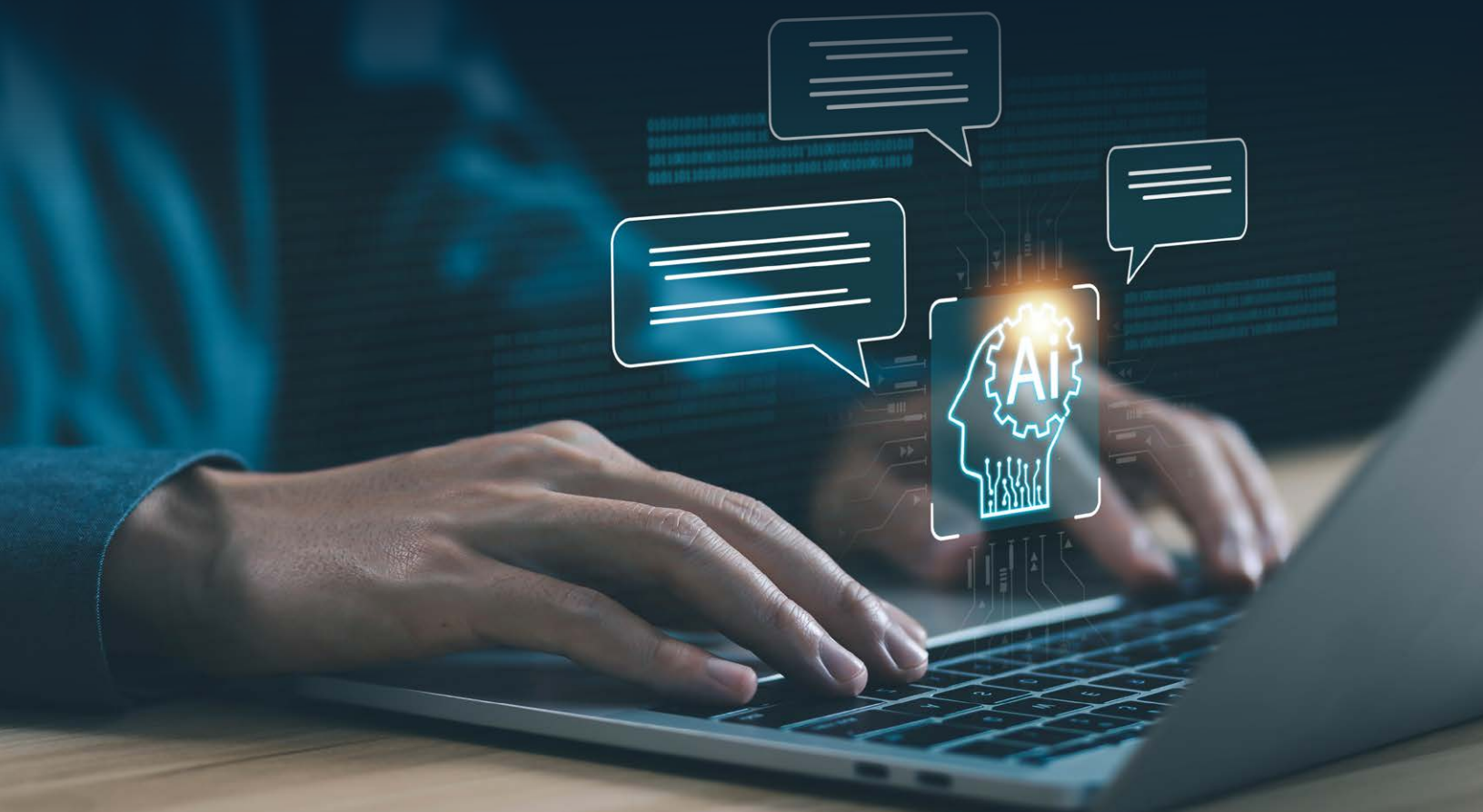
LTIMindtree

Whitepaper

Mastering the Generative AI Project Lifecycle

A Strategic Guide for Enterprises



Introduction

Today, every industry has witnessed the transformative effects of Generative AI or Gen AI. The market leaders have also tested some niche use cases and seen improvements in productivity, efficiency, and cost-effectiveness. They are now looking to scale these use cases across the organization. At the same time, many organizations are still struggling to implement Gen AI and are unsure where to start. Considerations around cost, skill set, and regulations are also holding them back.

Gen AI application lifecycle management deals with the selection, customization and fine-tuning, evaluation, guardrails, and deployment of generative AI. Any enterprise looking to adopt Gen AI solutions at scale to benefit its customers and employees must understand every aspect of this lifecycle.

This whitepaper will help enterprises understand the Gen AI project lifecycle and provide solutions to some of the most common challenges, borrowing from our experience of implementing LTIMindtree's in-house Gen AI platform.

The Gen AI project lifecycle

The lifecycle of a typical Gen AI project is divided into the following steps:

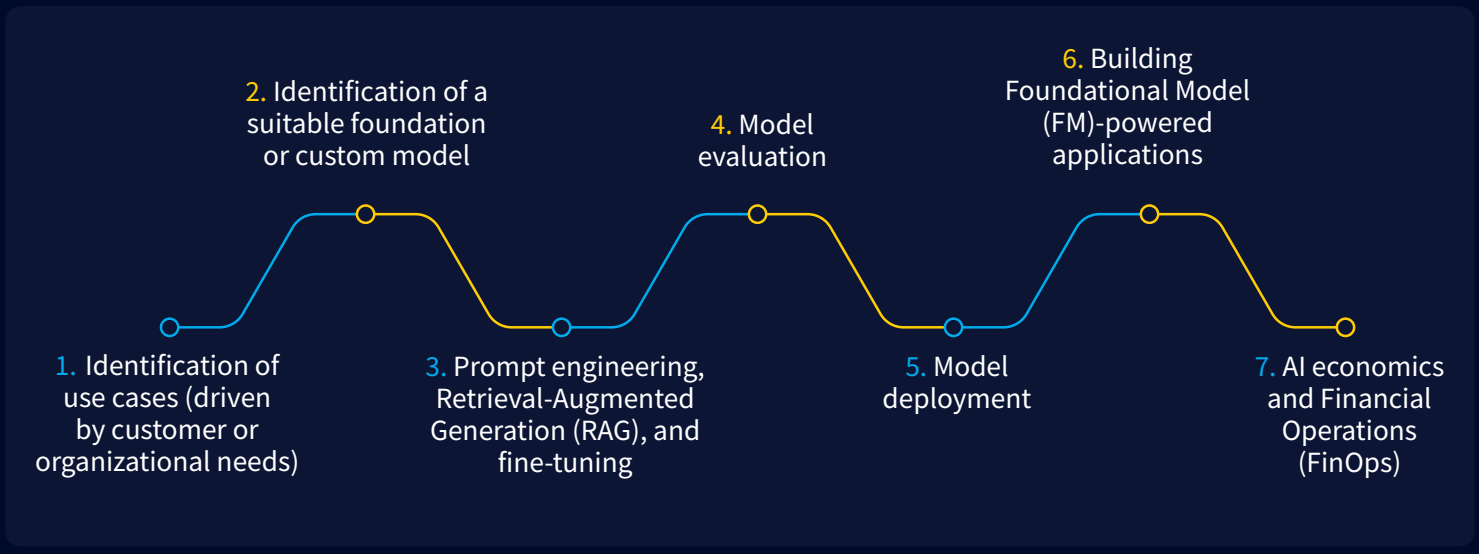


Fig 1: Stages of the Gen AI project lifecycle

1. Identification of use cases (driven by customer or organizational needs)

To effectively identify a use case, an organization needs to categorize and weigh several priorities appropriately. CXOs need to drive design thinking sessions with business and technology SMEs to arrive at a consensus on shortlisting the use case based on the weightage pattern.

Below is a suggested weightage pattern we used that could be tailored to contemporary needs.

Category	Score
Productivity improvement	20%
Customer experience	20%
Data accessibility	10%
Ease of integration and implementation	15%
Product and service quality improvement	10%
Return on Investment (ROI)*	25%
Total	**100%**

*Please note that the weightage given to ROI depends on factors like business goals and the industry to which the organization is aligned.

Business goals: If the primary goal is to improve efficiency or reduce costs, ROI will likely be a major consideration. However, if the focus is on innovation or gaining a competitive edge, the organization may accept a lower ROI in the short term.

Industry alignment: Organizations that rely heavily on data analysis or creative content generation will give more weightage to ROI when evaluating Gen AI use cases.

A typical use case for the manufacturing Industry is shown here as an illustration.

You may also refer to Google cloud for 101 real-world Gen AI use cases.[i]

Use Cases	**Product Design**: Generative AI can assist in generating innovative designs for new products, optimizing their aesthetics, functionality, & manufacturability.
Productivity Improvement	Generative AI can significantly reduce the time required for product design iterations. This metric measures the reduction in design time compared to traditional design processes. A higher design time reduction indicates improved productivity by enabling faster exploration of design alternatives, reducing time-to-market, and increasing the efficiency of the design process.
Productivity Improvement Score (20%)	5
Customer Experience	Generative AI can foster design innovation by generating novel and creative design alternatives. This metric evaluates the level of innovation and uniqueness achieved in the design process. It can be measured through user surveys, feedback, or assessments of the product's novelty in the market. Higher design innovation metrics indicate improved customer experience through differentiated and innovative product designs.
Customer Experience Score (20%)	3
Product & Service Quality Improvement	Generative AI can help validate design concepts and identify potential design flaws before physical prototyping. This metric measures the accuracy of design validation performed using Generative AI compared to physical testing or empirical data. Higher design validation accuracy indicates improved product quality by identifying and addressing potential issues early in the design process, leading to fewer design errors and better overall quality.
Product & Service Quality Improvement score (10%)	4

You may use one of the many products and/or feature prioritization frameworks while tweaking them to ensure effective Gen AI component selection.

For example, the RICE framework is a prioritization framework used to evaluate and prioritize features or projects. RICE stands for:

R - Reach (How many users will this feature impact?)

I - Impact (What is the expected impact on the user or business?)

C - Confidence (How confident are we in the estimates and assumptions?)

E - Effort (What is the estimated effort (time, people, cost) required to complete the feature?)

The RICE score is calculated by multiplying the Reach, Impact, and Confidence scores, then dividing by the Effort score.

Application of RICE on Gen AI engagements

1. **Feature prioritization:** Prioritize features or capabilities to be developed in a Gen AI system based on their potential impact, reach, confidence, and effort required.
2. **Model selection:** Evaluate and compare different AI models or approaches, considering factors like accuracy, scalability, and computational resources.
3. **Data selection:** Prioritize data sources or datasets for training Gen AI models based on factors like data quality, relevance, and availability.
4. **Research directions:** Evaluate and prioritize research directions or hypotheses by considering factors like potential impact, feasibility, and resource requirements.
5. **Resource allocation:** Allocate resources (e.g., personnel, computational resources) to different aspects of a Gen AI project based on prioritization.
6. **Value alignment:** Evaluate and prioritize value alignment in Gen AI systems, considering factors like ethical considerations, societal impact, and user needs.
7. **Explainability and transparency:** Manage efforts to improve explainability and transparency in Gen AI systems, considering factors like model interpretability, feature importance, and user trust.

2. Identification of suitable foundation or custom models

Most organizations prefer to use a base foundation model and build the application leveraging their own data. They also seek a scalable, reliable, and secure player who will keep their data private.

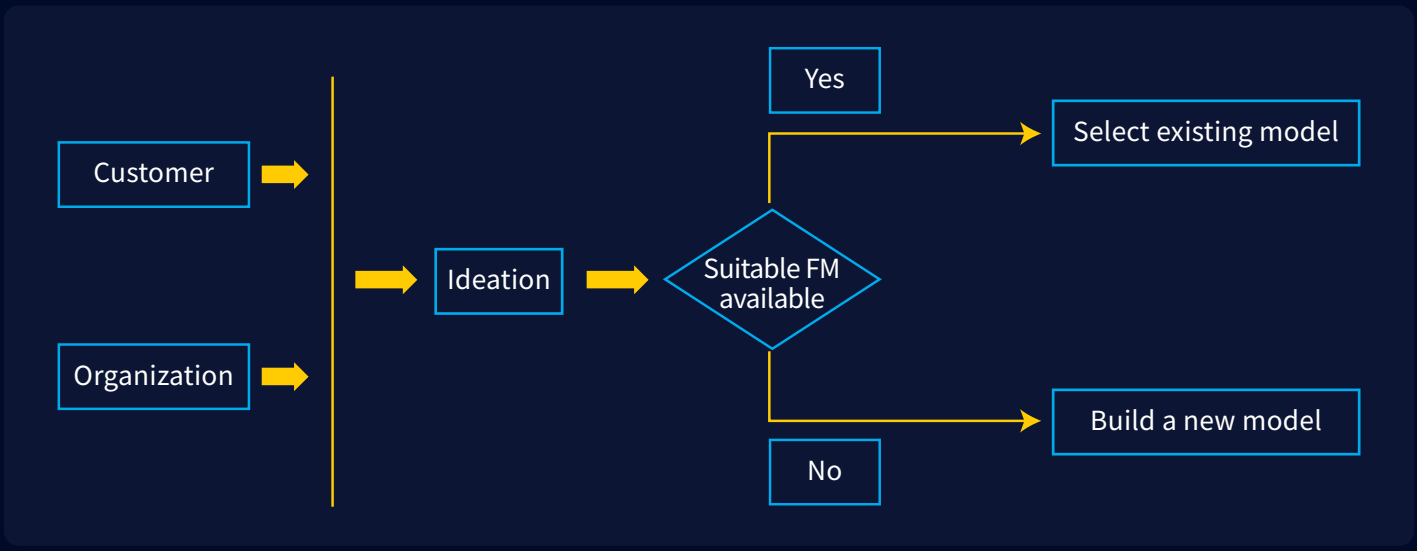


Fig 2: Process of selecting a foundational model

Organizations that have already set up their Machine Learning (ML) and data science algorithms will prefer to jumpstart their application development with any foundation model provider.

Many organizations that want to leverage FMs to transform their business functions and operations will look for a fully managed FM service provider who can completely manage infrastructure, integration, and data security.

Meanwhile, organizations who want to shrink development cycles, improve their coding standards, and help their development teams will look for FMs with a Copilot feature to improve productivity.

LTIMindtree has already built an enterprise Gen AI enablement platform with inbuilt navigator apps (CoPilot), available to all employees and customers.

An important and handy artifact for model selection is the model card, which organizations can use to gain information on the model's purpose, performance metrics, training data, and potential biases.[ii]

3. Prompt engineering, RAG, and fine-tuning

Prompt engineering[iii] helps FMs communicate effectively and get more relevant and accurate results. Generally, every organization has some means of responding to employees' queries, be it through chatbots, ticketing mechanisms, or a few insightful dashboards, such as Prohance or Ultima.

The quality of the prompts, the expected outcome, and the way it is put forward to the FM all go into the art and science of prompting. Here are the various prompt engineering techniques:[iv]

1. Zero-shot prompting: This technique leverages the generative AI model's abilities without providing additional data. It's useful for tasks where you want the model to perform a task it hasn't been explicitly trained on.

2. Few-shot prompting: By adding examples that the Gen AI model can use, you can enhance its performance by guiding its output for specific tasks.

3. Chain-of-Thought (CoT) prompting: Break down complex tasks into smaller sub-tasks, helping the model reason step-by-step and improving reliability.

4. In-Context Learning (ICL):[v] Integrate task demonstrations directly into the prompt in a natural language format. These demonstrations serve as examples for the model to learn from. Unlike traditional fine-tuning, ICL allows pre-trained Large Language Models (LLMs) to address new tasks without adjusting model parameters. Instead, the model leverages the provided examples within the prompt to perform the task. In summary, ICL empowers LLMs to learn new tasks from natural language prompts, making it a powerful technique for adapting models without extensive fine-tuning.

Retrieval-Augmented Generation (RAG)

RAG is a technique that allows FMs to access and process information from external sources before generating a response. This helps them include factual details from external sources to create more accurate responses that may not have been part of the pre-trained data.

This external source can be a vast collection of text documents, code, or any structured information that FMs can access and process. It needs vector databases, where the data is converted into numerical representations called vector embeddings that capture the semantic meaning and relationship between words and concepts.

The vector database uses algorithms to index and query vector embeddings, which enable Approximate Nearest Neighbor (ANN) search through hashing or graph-based search.

The input to the FMs is our Query + Prompt + Enhanced Context.

The key takeaway is that while prompts instruct FMs on what kind of response to be generated, RAG provides the factual context to support these responses.

RAG with DB connectors, or advanced RAG, is a way of integrating document context embedded with database connections so that the LLM can interpret and present consolidated insights. For example, we could integrate SQL, SAP, and Apache Hadoop data to augment document context with database context in the prompts. Creating automatic task-specific agents and automating the workflows will be the obvious next step, albeit with the right guardrails and audit trails in place.

LLM fine-tuning

FMs that are trained on a generic, time-specific corpus of information are less effective for domain-specific tasks. Also, because they are trained offline, there is always a need to catch up with the delta. Keeping the model's data up to date is an ongoing activity. Earlier, models were updated after years, but now they are updated quarterly or even monthly.

This can be addressed by fine-tuning, which is a supervised learning technique where a pre-trained model is adapted to a new task while its parameters are adjusted using a smaller task-specific dataset.

There are four types of fine-tuning methods

1. **Self-supervised:** Train models on unlabeled data, teaching them to predict part of the input from other parts (predicting missing words).
2. **Supervised:** Train models on labeled data (input-output pairs), which allows the model to learn task-specific patterns, which can help in improved performance and cost efficiency.
3. **Reinforcement learning:** Train models through trial-and-error, where the model learns to take actions to maximize a reward signal. This is widely used in robotics and autonomous vehicles.
4. **Proximal Policy Optimization (PPO):** Fine-tune the model's parameters to optimize specific goals. This reinforced learning algorithm is used across applications, including Generative Pre-trained Transformer (GPT).

Let's look at Parameter Efficient Fine Tuning (PEFT), where the original model remains unchanged, but a smaller Low-Rank Adaptation (LoRA) is created. This adapter is loaded into the pre-trained model and used for inference.

This LoRA is particularly useful for adapting models to new tasks or domains without requiring extensive training. This approach significantly reduces the computational resources required compared to traditional fine-tuning methods.

Some examples of fine-tuned FMs for industry verticals are:

1. Manufacturing - SymphonyAI (Anomaly detection and process optimization)
2. Manufacturing - GE Aviation (FM for engine performance)
3. Finance - BloombergGPT
4. Medicine - Med-PaLM

4. Model evaluation and governance

Model evaluation and governance are used to check the accuracy, relevance, drift, toxicity, and hallucination of Gen AI models. Models such as Watsonx.governance provide metrics for different use cases, including content summarization, content generation, RAG relevance, QA, entity classification, etc.

Let's look at a few of the important metrics and benchmarks:

Metrics

1) Perplexity: Perplexity measures a model's ability to predict the contents of a dataset. The higher the likelihood the model assigns to the dataset, the lower the Perplexity.

2). ROUGE (Recall Oriented Understudy of Gisting Evaluation) score is used in text summarization tasks to objectively assess the quality of machine-generated summaries in comparison to reference summaries. ROUGE score closer to 0 indicates poor similarity and a score close to 1 indicates strong similarity.

The following ROUGE scores indicate good and moderate values:

ROUGE 1 - Unigram (scores around 0.4 to 0.5 are moderate, and above 0.5 are considered good)

ROUGE 2 - BIGRAM (scores around 0.2 to 0.4 are moderate, and above 0.4 are considered good)

ROUGE L - Longest words match (scores around 0.3 to 0.4 are low, and around 0.4 is considered good)

Benchmarks

BLEU and GLUE: Bilingual Evaluation Understudy (BLEU) and General Language Understanding Evaluation (GLUE) are both evaluation benchmarks used in Natural Language Processing (NLP) to assess the performance of FMs.

RAGAS is another open-source evaluation metric for different use cases that can be considered.

Watson.governance also offers evaluation metrics like Rouge, BLEU, Meteor, Sari, Jaccard, Cosine Similarity, etc., which can help customers better evaluate, engineer, monitor, and improve, their solutions.

Monitoring and explainability features help monitor model drift/faithfulness, bias, and hallucinations on an ongoing basis, and some explainability is an important customer ask.

Holistic Evaluation of Language Models (HELM) is a framework for increasing the transparency of language models. It is also used to evaluate text-to-image models in Holistic Evaluation of Text-to-Image Models (HEIM).[vi]

Governance and responsible AI

Responsible AI involves a set of practices dedicated to the safe, trustworthy, and ethical design, development, and deployment of Gen AI applications.

Guardrails refer to a set of predefined policies and guidelines that regulate and oversee FM behavior and output. These include:

1. Moderation and explainability: Trust, security, explainability, hallucinations, toxicity, drift, bias, etc. [vii] [viii]
2. Risk mitigation: Security, prompt injection, leakages, red teaming, risk, and compliance.
3. Role-based Access Control (RBAC)/FinOps: Quota limits, semantic caching, and sharing of embeddings through the content hub.

5. FM deployment

In the deployment phase, you integrate your model into the production environment. For FMs to be efficient and scalable, we need to optimize them by examining common factors like model size, latency, throughput, and memory footprint.

Here are some key areas to focus on and the corresponding tools and techniques for FM optimization:

Area	Description	Tools & Techniques
FM selection and fine-tuning	As discussed in Step 2, choose the right FM that meets your accuracy and size requirements. Fine-tune the model on a task-specific dataset to improve its efficiency and performance for your use case.	Some of the open-source platforms available in the market are Labellerr, Kili, and Labelbox, which help create high-quality labeled datasets for fine-tuning FMs.
Hardware and infrastructure optimization	Select hardware that aligns with your FM's requirements, such as memory, processing power, and storage. Utilize cloud platforms for scalability and flexibility, or consider on-premise deployment for enhanced data security.	Cloud Service Providers (CSPs) provide resources for deploying and optimizing FMs on specialized hardware.
Prompt engineering and iterative refinement	Craft effective prompts that provide clear instructions and context to the FM. Gather user feedback and iteratively refine prompts to improve the quality and relevance of the FM's outputs.	Factor in the Subject Matter Expert (SME) feedback and iterate.
Performance optimization	Leverage techniques like quantization, pruning, and knowledge distillation to reduce the FM's size and improve inference speed. This can minimize latency and optimize resource utilization.	Tools like FMPerf and Langsmith offer functionalities to benchmark and evaluate the performance of FMs

These days, most deployment teams prefer using HELM, a package manager that automates the creation, packaging, configuration, and deployment of Kubernetes applications. HELM also takes care of rollbacks should a revert be necessary.

6. Building FM-powered applications

With the advent of high-quality open-source models available in the market, organizations prefer to use either RAG or fine-tuning open-source models rather than training their own FMs from scratch.

LangChain, LlamaIndex,[ix] and Triton Inference server are a few open-source frameworks widely used in creating FM-powered applications. Apache Airflow is also used to manage workflows using DAGs.

FM deployment timelines can vary from weeks to months, depending on the nature of the use case or the complexity of the application. Cultivating and nurturing a stack is one of the main tasks when looking to build in-house. Partnering with Gen AI solution providers could be a hybrid approach while we build some pieces and buy some pieces of the whole stack needed.[x]

Talent stack/human resource stack for hiring

One of the decisive outcomes of design thinking sessions is the tradeoff between "Build or Buy" for implementing Gen AI solutions. As technology evolves, many organizations will continue to face an acute shortage of in-house skillsets for executing their use cases with ease.

Large organizations have their captive capability power but it still needs to be realigned and empowered to suit GenAI needs. This calls for talent acquisition in the following areas.

- Data Science engineers (AI/ML)
- Cloud Architects
- Front-end developers (UI, Angular)
- Front-end designers (UX)
- Backend developers (Python, Java)
- DevOps CI/CD engineer (Kubernetes, CKA/CKS)
- Performance engineers (Automated testing)
- Cyber Security engineers
- Data privacy engineers
- Prompt Engineers
- Business analysts

Cutting edge technologies need cutting edge talent and there is no guarantee of being able to find fully trained resources at every level. Maintaining a periodically updated Training regimen and constant sharing of challenging ideas is a must to stay abreast of the ever dynamic Gen AI era

Role of partners - pieces of the Gen AI puzzle

Many open-source software companies help developers build FM-powered applications. One such example is NVIDIA/NeMo guardrails. NeMo guardrails will help FM-powered applications be accurate, appropriate, and secure. The software helps with some illustrations, the code block, and user documentation, which businesses can use to add safety and enable AI apps to generate permissible text.

Another is the Watsonx.governance integration that leverages model lifecycle management, which includes model selection, fine-tuning, training, evaluation, monitoring, deployments, etc., for externally hosted models. Evaluation metrics like Rouge, BLEU, Meteor, Sari, etc., can help customers better evaluate, engineer, monitor, and improve their solutions.

Monitoring and Interpretability features help monitor model drift/faithfulness, bias, and hallucinations on an ongoing basis. Some level of Explainability and Source-attribution are important customer asks.

Monitoring and automatic redaction of Personally Identifiable Information (PII) and Hate Abuse Profanity (HAP) using traditional AI methods rather than LLM, is a critical aspect of Prompt Moderation since the risks associated with loss of private and confidential data are high and we would not want these to go outside the Organization Network.

The adoption of IP-indemnified open-sourced IBM Granite and Sandstone models, which offer IP-free training data, Source-attribution and better explainability, is a step in the right direction for enterprises looking for on-prem deployments with better control.

7. AI economics and Financial Operations (FinOps)

AI economics and FinOps help to navigate the complex interplay between technological innovation, resource allocation, and financial sustainability.

AI economics

The primary focus areas of AI economics include:

1. Cost-benefit analysis: Evaluate the cost of developing and training Gen AI models against the potential benefits.

2. Resource allocation: Check the optimal usage of computational resources, data, and personnel.

FinOps

While FinOps typically emphasizes cost management, resource optimization, budgeting, and forecasting, FM Applications must track token telemetry. This involves monitoring token usage, analyzing token embeddings, and monitoring tokenization metrics.

Conclusion

In today's rapidly evolving technological landscape, organizations face the dual challenge of keeping pace with trends and ensuring robust FinOps and cybersecurity. As many embark on their journey with Gen AI, it's crucial to follow best practices that ensure Responsible AI for both employees and customers. Key strategies include anticipating future changes, focusing on trust, security, compliance, and cost transparency throughout the Gen AI lifecycle, and prioritizing deployment areas with the right architectures and models. Talent acquisition and development are critical, as is leveraging diverse platforms and partners like NVIDIA[xi] and AWS. Cultivating a Gen AI-first mindset, driven by leadership and supported by research and training, positions organizations to excel in this fast-paced era. By adopting these strategies, businesses can navigate the complex Gen AI project lifecycle effectively, ensuring both innovation and security.

References

i. 101 real-world gen AI use cases from the world's leading organizations, Google Cloud, 13 April, 2024: https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders

ii. AI ACROSS GOOGLE: Palm 2, Google: https://ai.google/discover/palm2/

iii. What is Prompt Engineering?, AWS: https://aws.amazon.com/what-is/prompt-engineering/

iv. Prompt engineering guide, Github: https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/guides/prompts-adversarial.md

v. How does in-context learning work? A framework for understanding the differences from traditional supervised learning, Stanford AI labs: https://ai.stanford.edu/blog/understanding-incontext/

vi. A reproducible and transparent framework for evaluating foundation models, Stanford CRFM: https://crfm.stanford.edu/helm/

vii. Impact of Generative AI on Content Moderation, Avasant, July 2023: https://avasant.com/report/impact-of-generative-ai-on-content-moderation/

viii. Regulations on the Protection of the Citizens' Fundamental Right to Freedom of Expression and Stringent Scrutiny of Online Content | Market Insights™, Everest Group, February 2023: https://www.everestgrp.com/tag/content-moderation/

ix. Let's talk about LlamaIndex and LangChain, Superwise, November 2023: https://superwise.ai/blog/lets-talk-about-llamaindex-and-langchain/

x. 16 Changes to the Way Enterprises Are Building and Buying Generative AI, Andreesen Horowitz, March 2024: https://a16z.com/generative-ai-enterprise-2024/

xi. Driving Enterprise Transformation: CIO Insights on Harnessing Generative AI's Potential, Nvidia: https://www.nvidia.com/en-us/on-demand/session/gtc24-s62779/?playlistId=playList-87118008-d10b-42f9-8c57-a50bbf006662

About the authors

Kishore Yallapragada

Associate Principal, Enterprise AI Platform

With over two decades of IT consulting and industry experience, Kishore has served several clients across industry verticals. He is currently driving program management activities with a focus on technical hiring and senior leadership portfolio reporting. Kishore has showcased our Gen AI product at multiple summits and forums.

Kiran Chandranna

Director, Enterprise AI Platform

With varied experience as a techie, an MBA and Finance Professional, and a tech Entrepreneur, Kiran now drives marketing, partnerships, and customer experience for LTIMindtree's Enterprise Gen AI platform.

LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by 81,000+ talented and entrepreneurial professionals across more than 30 countries, LTIMindtree — a Larsen & Toubro Group company — solves the most complex business challenges and delivers transformation at scale. For more information, please visit https://www.ltimindtree.com/.