

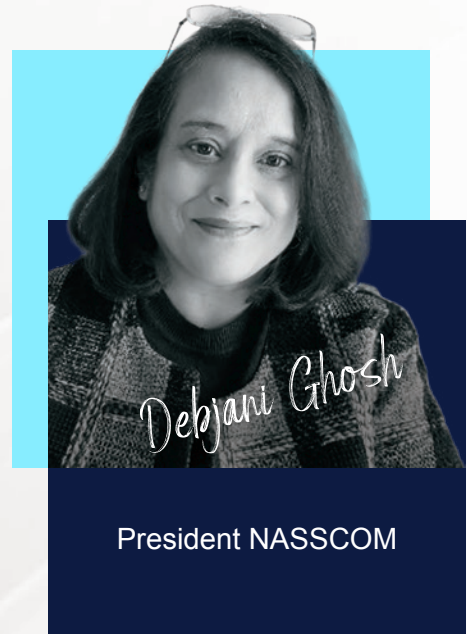
FMOps

The
GenerativeAI
Imperative
for Production

“

Generative AI's role in economic growth is multifaceted and undeniable. Today, as most countries are grappling with the challenges of a slower economic growth, driven by slowdown in growth of labour and productivity due to ageing population, harnessing AI's prowess is not just a choice but a necessity for survival. AI optimizes resource allocation, enhances decision-making, and fuels innovation, unlocking unprecedented economic potential. Generative AI offers multitude of opportunities to accelerate growth across sectors, promises a new era of productivity, efficiency, and competitiveness, provided we leverage it right.

”



President NASSCOM

Foreword

Technological innovation, with its captivating rhythm and resonance, has consistently and profoundly reshaped our understanding of what's possible. Time after time, we've witnessed these transformations, and today, we are on the brink of yet another monumental shift. Generative AI, along with its intricate underlying operations, represents this forthcoming wave of change. As we stand poised to not only witness but also play pivotal roles in this new era, the responsibility and opportunity to embrace, comprehend, and optimize this technological marvel becomes paramount.

This report, meticulously curated, offers a deep and insightful dive into the vast oceans of Generative AI. Designed with both the seasoned expert and the eager novice in mind, it aspires to illuminate the pathways of an AI-driven world. This document is our humble attempt to serve as a compass, guiding your journey through the vast potential of tomorrow's AI landscape.

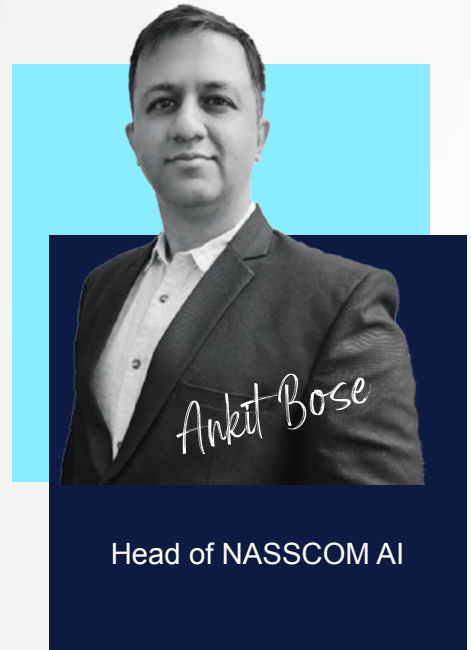




Table of contents

EXECUTIVE SUMMARY

STRATEGIC VIEWPOINT

- 01.** Introduction to Generative AI
- How is Generative AI Different from Traditional AI?
 - What is FMOps in Generative AI?
 - Why FMOps?
 - What Does LLMOps Mean in the Context of FMOps and Generative AI?
 - Differences Between MLOps and LLMOps

- 02.** Why Should Enterprises Embrace GenAI?

- 03.** How Should Enterprises Get Started with Gen AI?
- Finding the Right Gen AI Strategy
 - Embedding Strategies with Responsibility

PRACTITIONER'S VIEWPOINT

- 04.** What Does it Take to Make a Winning GenAI Solution?
- Finding the Right Model
 - Finding the Right Technique
 - Finding the Star Team
 - Finding the Right Tech Stack
 - LLMOps Current Landscape
 - LLMOps Implementation RACI Matrix
 - Operationalizing LLMs with LLMOps

- 05.** GenAI: 2023 & Beyond
- GenAI Trends of the Future

- 06.** Glossary

EXECUTIVE SUMMARY

This compendium intends to usher enterprises into the world of generative AI by offering an industry perspective into how to build successful generative AI solutions. It is an attempt to encapsulate the best practices to guide organizations on effective set-up, management, and scaling of Gen AI operations. We believe that this playbook will help stakeholders gain a clear understanding of foundation model operations and enable them in anticipating and mitigating the challenges that are associated with it at different stages.

In the context of Generative AI applications and solutions in production, it becomes critical to have a streamlined approach to develop, deploy, run, monitor, and manage language model applications. LLMOps is a practice and approach to overseeing the lifecycle of LLMs from training to maintenance using tools and methodologies. By operationalizing technology at scale, LLMOps aims to make the path to adopting Generative AI easier. LLMOps focuses on the operational capabilities and infrastructure required to fine-tune existing foundational models, capture the prompt engineering, build monitoring pipelines to capture experiments and deploy these refined models in production.

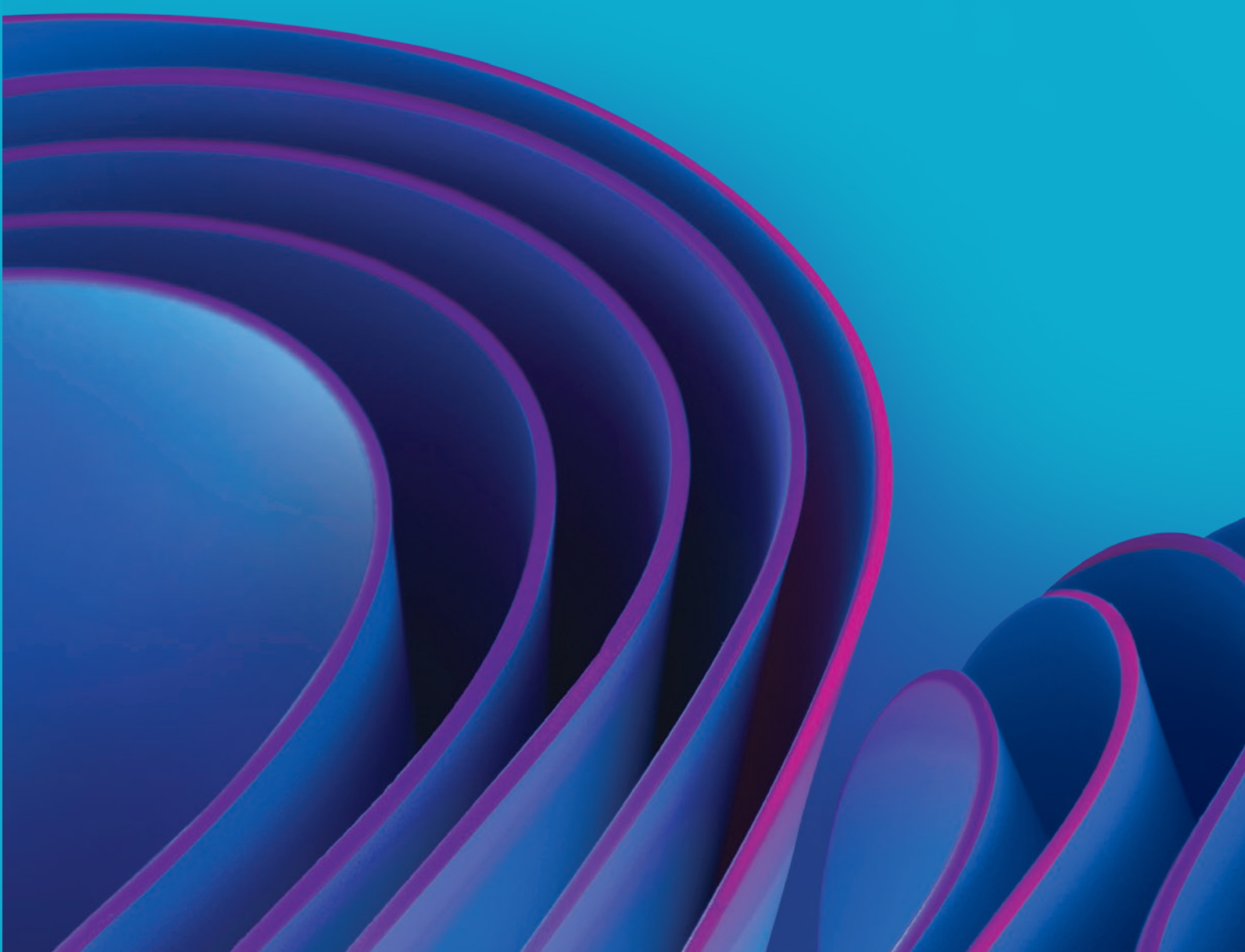


Strategic Viewpoint

01

Introduction to Generative AI

- How is Generative AI Different from Traditional AI?
- What is FMOPs in Generative AI?
- Why FMOPs?
- What Does LLMOPs Mean in the Context of FMOPs and Generative AI?
- Differences Between MLOps and LLMOPs



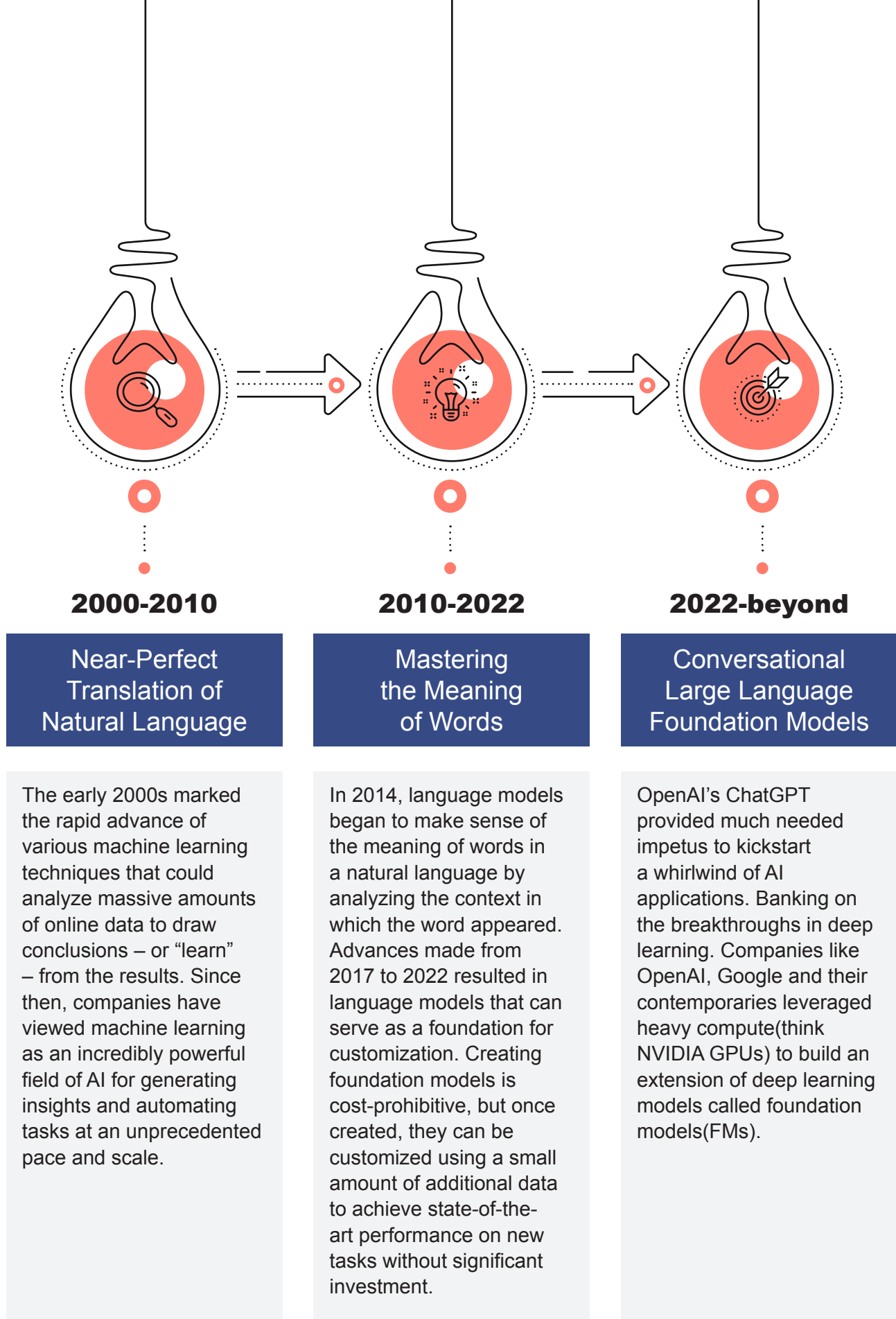
Introduction to Generative AI

The historical progression of transformative technological innovations reveals intriguing patterns. Consider the light bulb, introduced in 1879; it wasn't until 1930 that 70% of American households enjoyed the benefits of electricity. Fast forward to 2017: the revolutionary transformer architecture emerges and within six years, it is part of every modern-day AI application. ChatGPT has become a household name. Large language models, unlike any other innovation, violates all the patterns of the past with its unprecedented acceleration.

OpenAI's ChatGPT provided the much-needed impetus to kickstart a whirlwind of AI applications. Banking on the breakthroughs in deep learning, companies like OpenAI, Google, and their contemporaries leveraged heavy compute (think NVIDIA GPUs) to build an extension of deep learning models called foundation models (FMs). Coined by Stanford researchers, the term "Foundation Model" refers to models trained on comprehensive data using self-supervision at scale, enabling adaptation to a wide array of downstream tasks. FMs are

large AI models developed through extensive training on vast troves of unannotated data. These models are designed to capture general patterns and relationships in the data and are typically used as a starting point for a wide range of downstream tasks, such as text classification, question answering, and language generation in case of natural language processing (NLP).

Large language models (LLMs) like OpenAI's GPT-4, which power conversational interfaces like ChatGPT are a type of foundation model. These interfaces powered by LLMs are capable of generating new content in text, image, video and audio format by taking instructions in natural language. The multi-modal foundation models are collectively called as generative AI (GenAI), which refers to the generative capabilities of the FMs that can be applied to narrow tasks. GenAI is a subset of AI in which models are trained to generate new original content based on natural language input.



Generative AI: A timeline

Source: Genpact, NASSCOM

How is Generative AI Different from Traditional AI?

In light of the projected growth of the global AI market, set to approach nearly \$4 trillion¹ by 2030, enterprises are poised to unleash the advantages of conversational interfaces to solve narrow use cases, which were mostly untouched by traditional AI approaches. At the heart of GenAI advancement sits LLMs like OpenAI's GPT-4, Google's PaLM, or Meta's Llama, which are built upon the Transformer architecture, where multi-head attention layers are stacked in a very deep neural network. Existing LLMs mainly adopt similar model architectures (i.e., Transformer) and pre-training objectives (i.e., language modeling) as small language models. As the major difference, LLMs largely scale the model size, pre-training data, and total compute (orders of magnification). They can understand the natural language better and generate high-quality text based on the given context (i.e., prompts). Such a capacity improvement can be partially described by the scaling law, where the performance roughly follows a substantial increase with respect to the model size. Foundation models, including generative pretrained transformers (which

drives ChatGPT), are among the AI architecture innovations that can be used to automate enterprise solutions, augment humans or machines, and autonomously execute business and IT processes.

Unlike traditional AI, training GenAI models is resource-intensive. For instance, to train GPT-3, OpenAI had to collaborate with Microsoft for their supercomputers, which cost them millions of dollars. In addition to this, OpenAI assembled a team of human experts who flagged inaccurate responses from the model. The combination of a large corpus of data, efficient transformer architecture, funding, and human-in-the-loop (RLHF) made ChatGPT the success it is today.

Today, LLMs can be adapted via natural language prompts to do a passable job on a wide range of tasks despite not being trained explicitly to do many of those tasks. The Centre for Research on Foundation Models says that the significance of foundation models can be summarized by two words: emergence and homogenization.

What is FMOps in Generative AI?

FMOps deals with the operational capabilities required for the efficient alignment, deployment, optimization, and monitoring of foundation models within the framework of an AI system. The term Foundation model (FM) was first coined in 2021 by researchers at Stanford University. Unlike conventional task-specific models, FMs are gigantic, and feature billions of parameters.

This gives them an inherent trait of 'emergent capabilities,' such as reading comprehension and artistic creativity, as they learn to reconstruct data. Multiple variations of FMs have arisen, addressing tasks such as text-to-text, text-to-image, and speech-to-text, each offering distinctive levels of control and accessibility.

1. [The data dividend: Fueling generative AI by McKinsey](#)

Why FMOps?

FMOps helps enterprises foster collaboration, reduce conflicts, and hasten release cycles in their LLM pipelines.

Here are few reasons that make FMOps the right choice for building scalable GenAI solutions:

1

Enhanced Efficiency

FMOps empower data teams to accelerate model and pipeline development, creating high-quality models and faster deployment in production settings.

2

Seamless Scalability

With extensive scalability and management capabilities, FMOps allow for overseeing and monitoring multiple models within a continuous integration, delivery, and deployment environment.

3

Reduced Risk

FMOps catalyzes transparency and swift responsiveness to regulatory requests, particularly as LLMs are often under regulatory scrutiny. This ensures better adherence to organizational or industry policies, enhancing risk management and mitigating potential challenges.

4

Integration with DataOps

FMOps can seamlessly integrate with DataOps practices, facilitating a smooth data flow from ingestion to model deployment. This integration promotes data-driven decision-making and accelerates value delivery.

5

Faster Iteration and Feedback Loop

FMOps shorten iteration cycles and facilitate quick feedback loops. This agility is vital for adapting models to changing business needs.

6

Improved Security and Privacy

FMOps prioritizes safeguarding sensitive information and data privacy, ensuring protection against vulnerabilities and unauthorized access.

7

Better Resource Allocation

FMOps ensures access to suitable hardware resources like GPUs for efficient fine-tuning while also monitoring and optimizing resource usage.

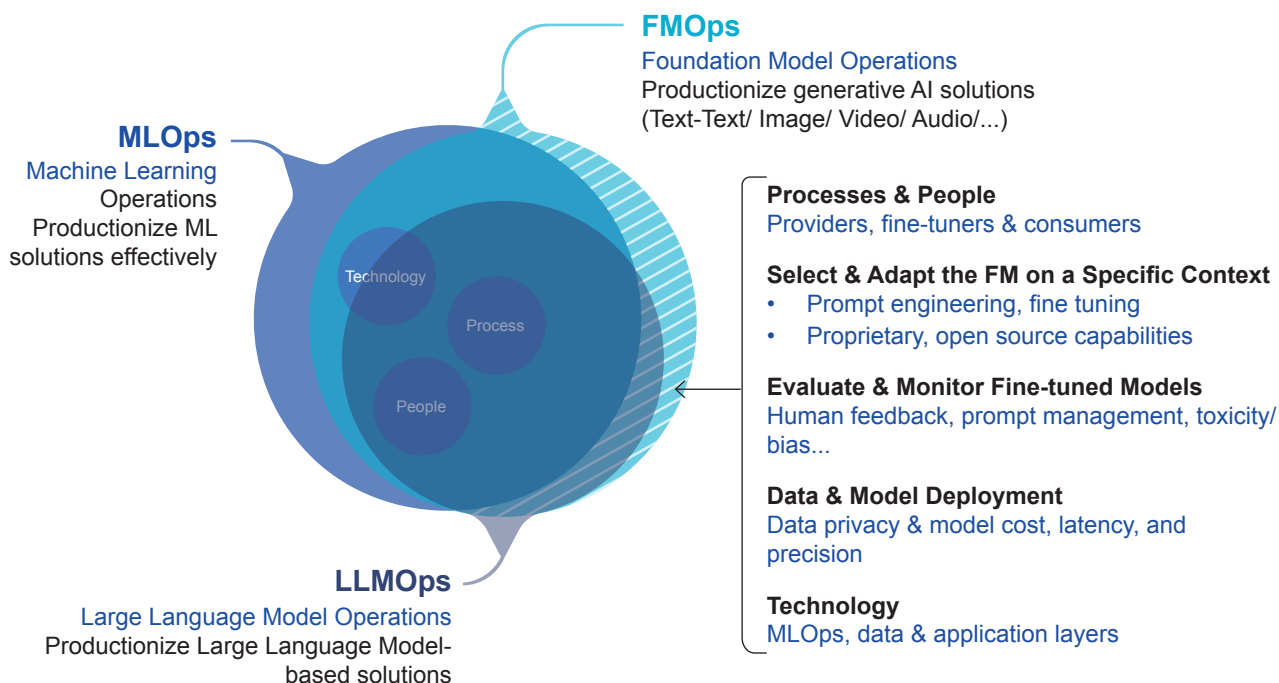
8

Enhanced Performance

FMOps directly contributes to improved model performance by ensuring high-quality and domain-relevant training data.

Source: Genpact

What does LLMOps Mean in the Context of FMOps and Generative AI?



MLOps vs LLMOps vs FMOps

Source: [AWS](#)

Operationalizing large language models is unlike traditional AI solutioning. As the vestiges of the MLOps paradigm serve as a reminder of the sophistication that surfaces at scale, GenAI, while mitigating a few of those challenges, brings a few of its own. This is where LLMOps (Large Language Model Operations) comes into the picture. LLMOps (Large Language Model Operations), a subset of FMOps (Foundation Model Operations) that includes Large Language and Visual Models, builds on the principles of MLOps (Machine Learning Operations) and helps enterprises deploy, monitor and retrain their LLMs seamlessly. For the purpose of this playbook, we will restrict our focus on LLMOps.

FM operations (FMOps) – This can productionize generative AI solutions, including any use case type.

LLM operations (LLMOps) – This is a subset of FMOps focusing on productionizing LLM-based solutions, such as text-to-text.

Differences between MLOps and LLMOps

Task	MLOps	LLMOps
Concept	Practices and methodologies for facilitating automated ML model operations and infrastructure monitoring and management.	Set of architectural best practices and methodologies to select, train, prompt engineer a Large Language Model application and infrastructure monitoring and management.
Data Management	New data needs to be sourced, wrangled, cleaned, and labeled.	Requires contextual data, which needs to be diverse and representative. When prompt engineering is required, sufficient examples need to be provided to control model bias and hallucination. Data privacy, security measures need to be integrated into the system design since data may leave the organization's environment when leveraging LLM's hosted outside the organization. Embeddings and vector database management metrics must be taken into consideration for production grade scalable solutions.
Experimentation	Improving ML performance with different model architectures and creating new features.	Improving the contextual capabilities of the selected LLM by applying a combination of techniques such as prompt tuning, few shot tuning, fine tuning of the LLM.
Evaluation	The model's performance is evaluated on a hold out validation set, metrics like accuracy and F-1 score are used to determine the models performance.	Model's performance is evaluated by using a broader set of metrics, like BLEU or ROUGE score. LLMs also need to be assessed on robustness, interpretability, and fairness. Therefore, a broader set of metrics are used like the BLEU score or ROUGE score. In some cases human feedback is used to evaluate the performance of an LLM.
Deployment	The deployment phase focuses on staging, split testing (A/B testing), and versioning (rollback).	Robust tools need to be used for the management of training data, the training process, and versioning of models. Implementing model drift systems are also important to keep track of misaligned inputs, and handling of adversarial attacks.
Monitoring	Tracking live metrics to monitor prediction quality.	Monitor the LLMs output for things like potential biases, and ethical issues to track the models performance.
Primary Cost Driver	Data collection & model training.	Inference and run costs in production.

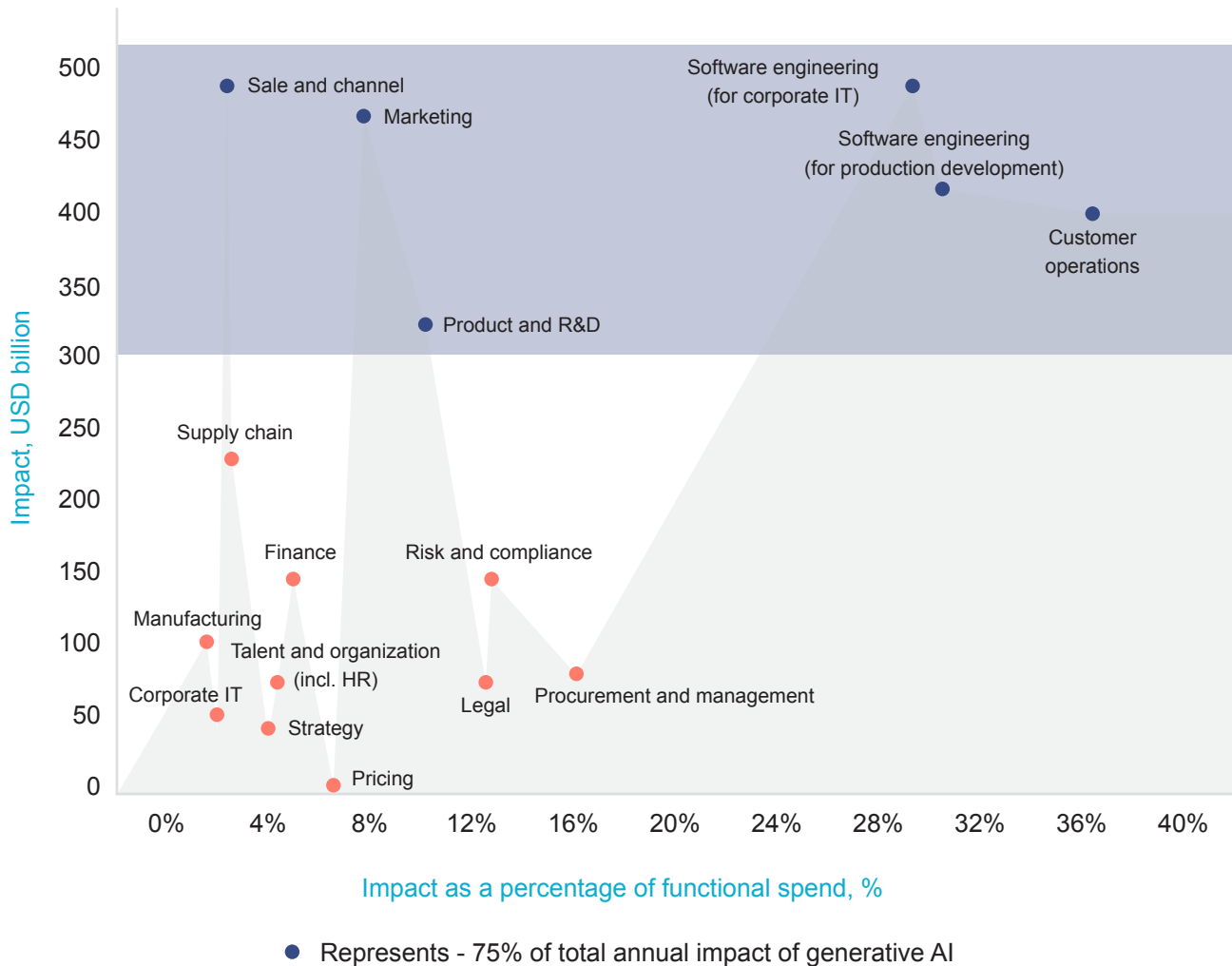
Source: Genpact, NASSCOM



02

Why Should Enterprises Embrace GenAI?

Why Should Enterprises Embrace GenAI?



Impact in billions vs Impact in functional speed

Source: NASSCOM

According to a survey by NASSCOM², GenAI could lead to a 20 to 30 percent improvement in productivity for select use cases in service lines such as application development and BPO. The productivity improvement at the organizational level is expected to be gradual, with a 10 to 15 percent gain in the first 12 to 18 months after deploying the technology at scale, with the potential to reach around 20 to 30 percent in 2 to 3 years. In service lines

that follow agile methodology, such as digital/cloud and D&A, the potential improvement is expected to be lower, at around 10 to 15 percent.

Enterprises are leveraging LLMs---open and closed-source---through readily available APIs. However, to effectively cater to their distinct requirements, enterprises must undertake the customization and fine-tuning of these models using their proprietary data.

2. [harnessing the power of generative ai – opportunities for technology services](#)

This tailored approach empowers the models to fulfill specific tasks, such as powering customer service chatbots or generating bespoke product designs, thereby optimizing operational efficiency and propelling competitive advantage.

The benefits of generative AI include faster product development, enhanced customer

experience and improved employee productivity, but the specifics depend on the use case. End users should be realistic about the value they are looking to achieve, especially when using a service as is, which has major limitations.

**\$2.6 trillion –
\$4.4 trillion**

Total economic³
benefit annually

~10 – 30 %

Technical Automation
potential in Knowledge
work

~30 – 40 %

Impact on collaboration
and decision making

0.1 – 0.6 %

Annual labour
productivity growth

3. [The economic potential of generative AI: The next productivity frontier](#)



03

How Should Enterprises get Started with GenAI?

Finding the Right GenAI Strategy

Embedding Strategies with Responsibility



If you have
a hammer,
everything
can start to
look like a nail.

~ Maslow's Hammer

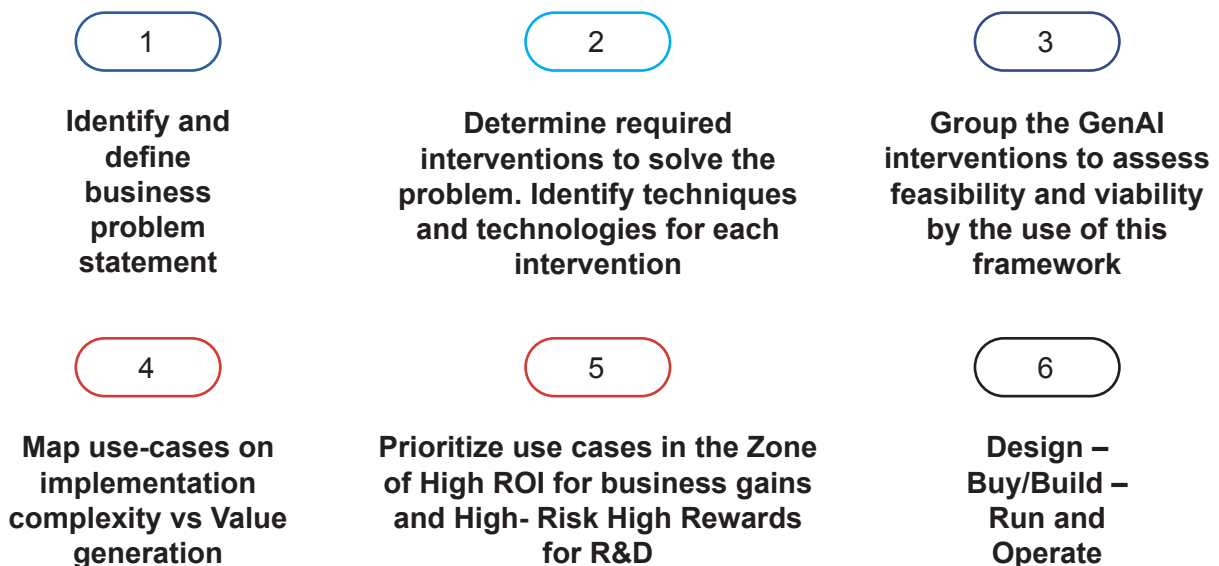
Every new wave of technology will have its fair share of advocates and adversaries. When the dust settles and the hysteria mellows, businesses typically end up in one of these camps: early adopters and laggards. But before that, the advocates will make the mistake of overselling and reinventing the wheel, whereas the adversaries make the mistake of over-regulating. For enterprises aspiring to be associated with GenAI, it is important not to fall into the trap of reinventing the wheel.

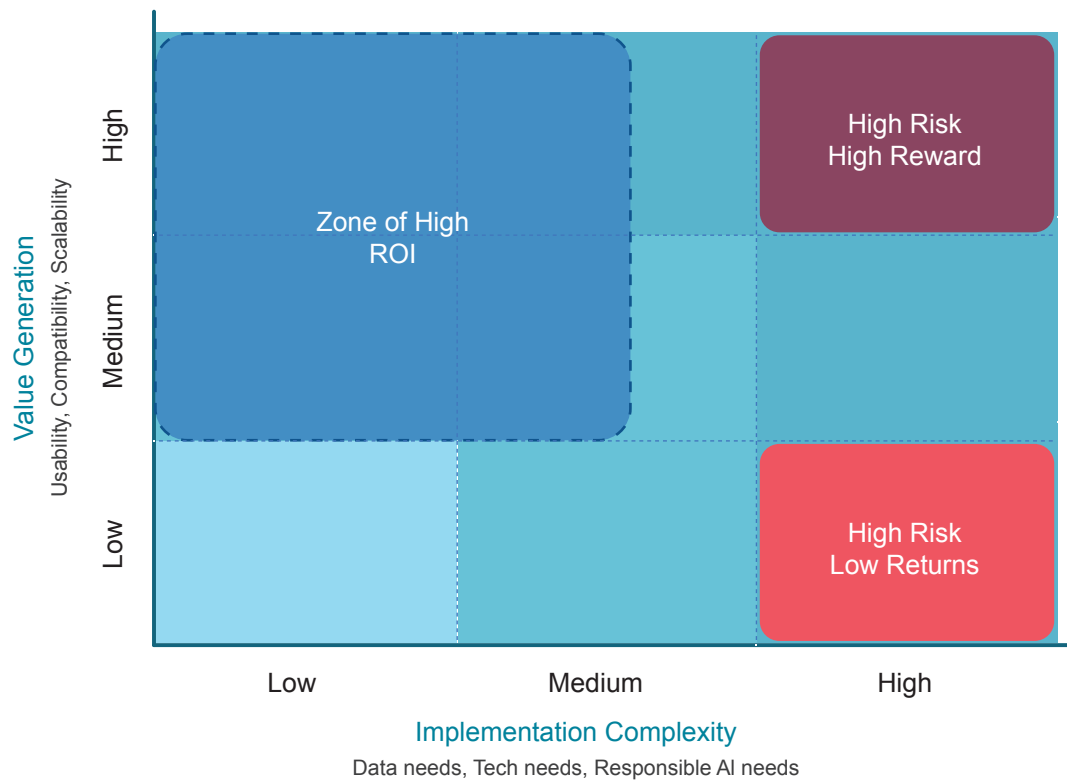
While GenAI lowers the barriers to entry for enterprises that are looking to leverage AI, it appends a new order of challenges that demands rethinking of AI implementation paradigms. From generic attributes such as the maturity of the organization to niche factors such as responsible AI frameworks, a thorough assessment of the in-house faculties and their effectiveness in the context of GenAI landscape is required.

Here is a 5-step plan to start a typical GenAI journey:

1. Identify use case viability:

The first step is to understand the business challenge to identify potential Generative AI intervention.





Use case selection framework

Source: Genpact

The use selection framework helps enterprises select the most feasible and profitable use cases with the right trade-offs.

The use selection framework helps categorize use case by mapping value generation of each use case to its implementation complexity. Here, the implementation complexity is a function of a variety of challenges that surface during the development stage of a GenAI solution. These challenges can vary range from fine-tuning costs to data availability, from accuracy requirements to governance requirements and more.

These complexities are mapped against the value generated by the use case. The value generation is a function of vendor competencies, maturity of the industry, multi-modal propensity, use-case interoperability and other such factors.

Mapping these functions will give rise to what can be called as 'Viability Zones'. These zones, as illustrated above, segregates the use cases into most important categories: high risk high reward zone and the zone of high ROI. Enterprises, as per the trajectory of their organizational maturity would choose to pick use case from one of these two zones. For instance, organizations which have invested heavily into R&D can afford to have a mix of high-risk-high-reward use cases and the use cases from the most viable High ROI zone.

2. Rightly assess for organizational maturity:

Have detailed discussions with CoEs to understand the skill and infrastructure requirements. This will help to arrive at a cost-benefit analysis.

3. Get your governance policies right very early into the game:

Conduct workshops information security team and the legal team to stay aligned with Responsible AI.

4. Identify the right partner ecosystem:

Ensure that the right partner ecosystem is in place to leverage during the development stage.

5. Educate the stakeholders:

Educate stakeholders about technology limitations so that guardrails are integrated based on joint decisions from development and business teams.

It is important to conduct a viability assessment of the use case at hand before latching on to one. Organizations should carefully evaluate the potential advantages and drawbacks of incorporating AI. Thorough scrutiny of service terms and privacy policies for each tool is essential, as these terms may differ significantly, even across different versions (e.g., web-based vs. programmatic interfaces, free vs. paid versions).

For instance, in the assessment of healthcare applications of GenAI, the introduction of novel image generation, while intriguing, [might not hold](#) as much relevance in the context of radiology, where the primary focus is on precise anatomical depiction rather than creative graphics. However, there could still be merit in generative applications geared toward improving existing medical images.

Finding the Right Gen AI Strategy

It is not necessary that every digital enterprise embrace GenAI. The onus is on the enterprises, especially the C-suite leadership, to pick an appropriate GenAI strategy. Low-maturity or late adopters might find it challenging to take up GenAI-powered use cases.

Any GenAI strategy that is a product of proactivity and diligence should, without a doubt, assess the organization's GenAI literacy and readiness thoroughly through a rigorous analytics maturity framework. These frameworks should assess data maturity, process maturity, technology maturity, talent maturity, and enterprise & leadership maturity in the context of GenAI. Such an assessment is critical to either bringing all the functions to a baseline maturity level or improving the maturity of an organization against the industry benchmark.

1. Democratize Ideas While Limiting Production

- To nurture in-house expertise, organizations should create Centers of Excellence (CoEs) that upskill employees into next-gen roles like prompt engineers, prompt compliant checkers, customer

protection officers, and other relevant roles.

- Setting up such CoEs is based on the principle of 'Democratize ideas while limiting production.'

CoEs help strike a balance between speed-to-market trade-offs with huge investment and the risk of confidential information leakage. Talent development and upskilling are crucial elements to succeed in the rapidly evolving Generative AI landscape.

2. Attain High-Quality Outputs

- Engaging subject matter experts (SMEs) throughout the project is crucial for attaining high-quality outputs.

Their input in defining requirements, specifying details, and validating results ensure the final product meets industry-specific needs.

3. Finding Profits in a Resource-Intensive Project

- Significant costs are encountered during the first stage of prompt tuning. This needs rigorous evaluation with multiple runs based on SME feedback.
- Factoring in these costs and sharing the information with stakeholders is vital to prevent unexpected surprises.
- While business leaders encourage employees to experiment with Generative AI, they also need to prevent employees from launching untested and unregulated AI projects.

A good practice is to log token counts and cost per call to Cloud APIs to provide more visibility to stakeholders.

4. Navigating through Limitations

- Hallucinations found in the outcomes might require further examination and fine-tuning.
- Additional labeling efforts could be needed to enhance results beyond prompt engineering.
- Token limitations might affect large-scale result inferences, and throttling errors could arise due to call constraints within a specified period.

It is important to educate stakeholders about these the limitations so that guardrails are integrated based on joint decisions from development and business teams.

5. Building Trust

- The chosen solutions should provide tangible benefits while adhering to ethical and regulatory standards.
- Scalability varies by task. It's crucial to convey these expectations to stakeholders for better expectation management and informed decision-making.

Knowledge-sharing and brainstorming sessions with clients have helped us to build more trust

in the Generative AI solutions, leading to faster operationalization.

6. How to Identify the Right Partner

- Enterprises would be better-off if they position themselves as a technology-agnostic player, so as to collaborate closely with industry leaders such as Google, Azure, AWS, Open AI, and NVIDIA for both specialized AI services.

This engagement is complemented by extensive experimentation with open-source architectures on platforms like Hugging Face and stability.ai, etc., to drive innovation and maintain a competitive edge while building Generative AI solutions.

7. Implementing Auditing Mechanisms

- Organizations must incorporate auditing mechanisms and robust data governance frameworks to protect customers from risks such as copyright infringement and proprietary data leakages while maintaining high standards of data privacy and security.

Robust Responsible AI practices enables right governance, guardrails, re-usable prototype delivery system, efficient operating rhythms and strong change management muscles in the enterprises. This also helps in prioritization of the use cases.

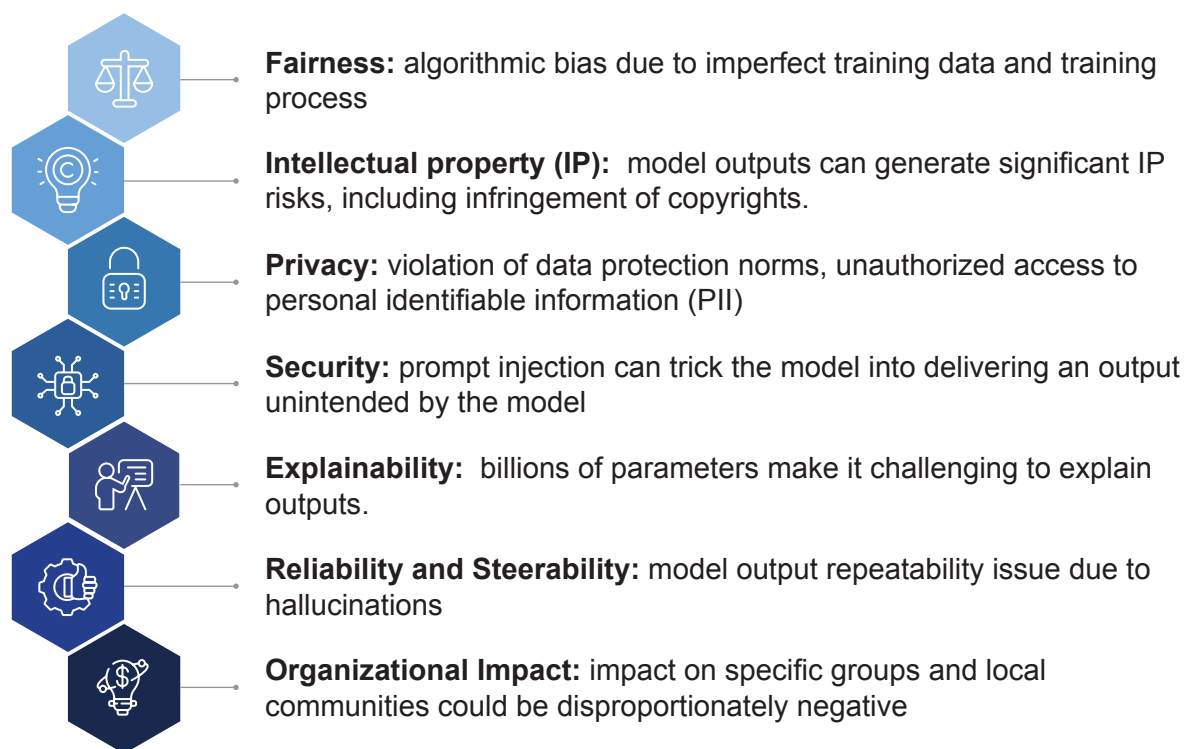
8. Adapt to Evolving Data Demands

- Outputs from Generative AI applications can include a mix of tables, code, and images generated from text. Enterprises must adapt by consolidating disparate data sources, talent, and technology stacks across various verticals to create reusable assets tailored to each industry's unique requirements.

Having custom frameworks for generating structured synthetic data helps build solutions for niches use cases, for example, for healthcare insurance companies, to ensure the availability of data for a lower environment without compromising confidentiality in production data.

Embedding Strategies with Responsibility

GenAI creates artifacts that can be inaccurate or biased, making human validation essential and potentially limiting the time it saves workers.



Primary challenges of building GenAI solutions

Source: Genpact

For instance, HR-leave policies are complex. More so, are the policies framed by the underwriters at an insurance company. It gets trickier if the insurance is related to a healthcare policy. Customer queries aren't always this straightforward. They can be ambiguously worded, complex, or require knowledge the model either doesn't have or can't easily parse. Complexities compound with few use cases while using GenAI. For instance, banks handle vast amounts of sensitive customer data, making data privacy and security their top priority. Generative AI relies on large datasets for training models, which can include personal identifiable information (PII) and financial details. Safeguarding this data from breaches and unauthorized access becomes crucial. Banks

must invest heavily in robust security measures and compliance frameworks to ensure data protection and maintain regulatory compliance, adding complexity and cost to the implementation of GenAI. The banking industry operates within a stringent regulatory framework to maintain transparency, fairness and protect customers' interests.

The shortage of skilled professionals in the market adds to the complexity of implementation. Additionally, banks need to foster a culture of AI adoption and create an environment that encourages employees to embrace the changes brought by generative AI. Many banks rely on legacy systems and complex IT infrastructures that may not be compatible with the requirements






of generative AI. Integrating AI algorithms and models into existing infrastructure can be a daunting task, often requiring significant investments in hardware, software, and data management systems.

Conventional chatbots don't have the intellectual make-up to stay in the context and think ahead of the customer. They can fumble over an adversarial question and jeopardize the whole conversation. Time lost in seconds translates to millions lost over a quarter. Large language

models have overcome this wrath of latency. However, LLMs can also be that eager employee who blurts out answers to incomplete questions in a board meeting.

According to a recent [study](#) by Microsoft, the emergent capabilities of LLMs surface with scale, and this contributes to "capability unpredictability," the idea that an LLM's capabilities cannot be fully anticipated. This makes embracing responsible AI practices more challenging.

Here are a few challenges of implementing RGAI:

API Opaqueness and Proprietary Nature of LLMs		LLMs are usually accessed through APIs. Most of the models are proprietary. It is impossible to access the weights, parameters, and provenance of the training data. Moreover, creators of LLMs like OpenAI can choose to deprecate or discontinue their models as per their requirements.
Stakeholder diversity		As the pre-trained nature of LLMs lowers the barrier to using and building on AI capabilities, the line between developers and end-users may be blurred.
Flawed perception		The natural language modality also contributes to a unique set of challenges.
Rapidly evolving ecosystem		Given the speed at which powerful new LLMs and their applications are being released, it is difficult for policymakers and third-party auditors to keep up and formulate governance mechanisms.
Mishandled Change Management		Disruptive technologies often induce organizational pressures to cut down the time between pilots and production. This chase can hinder establishing a robust governance mechanism for technologies in their infancy.

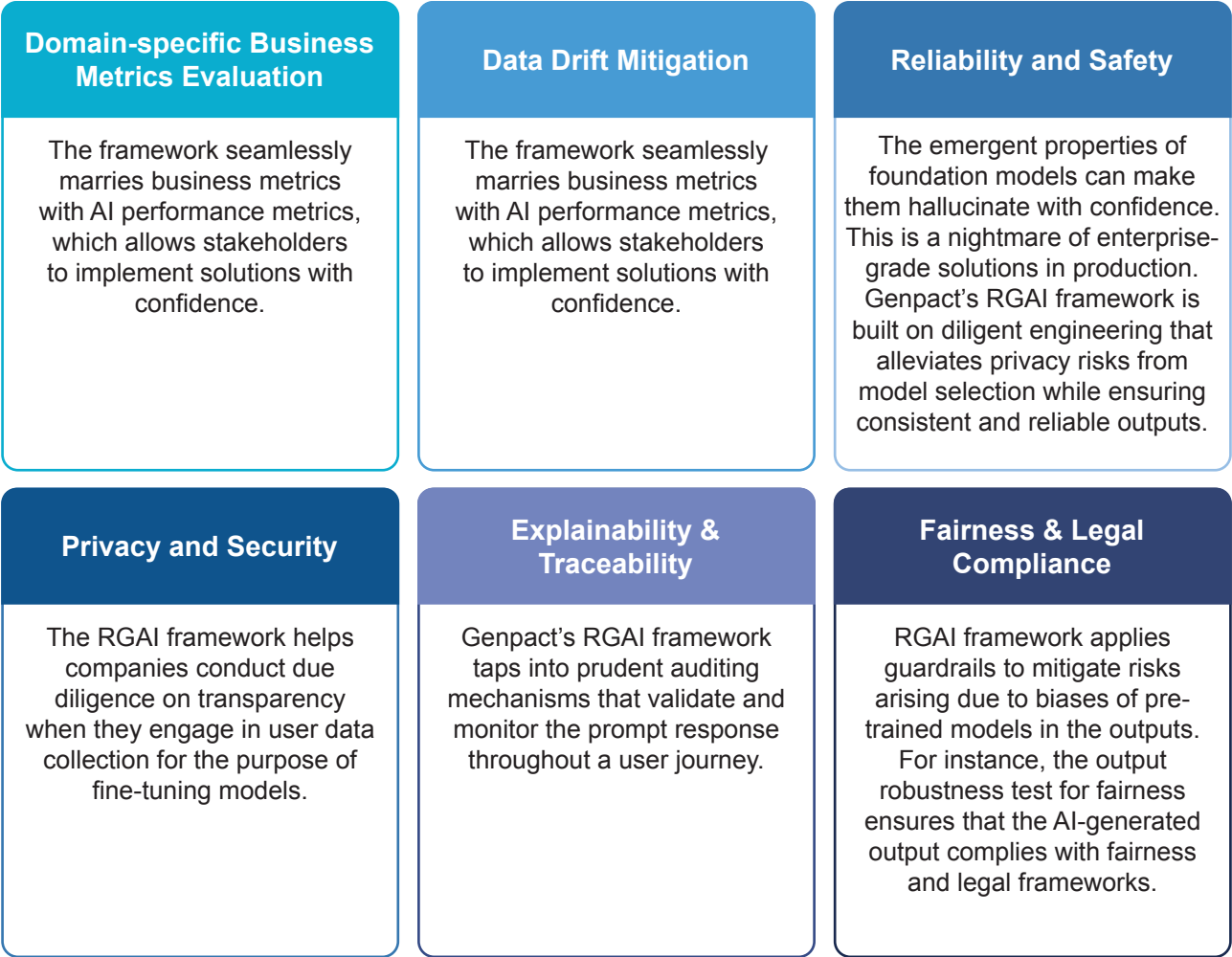
Asymmetric challenges of building GenAI solutions Source: Genpact

The capabilities of LLMs include question answering, dialogue generation, sentence completion, summarization, paraphrasing, elaboration, rewriting, classification, and more. Emergent capabilities of LLMs—like performing arithmetic or chain-of-thought reasoning—that are not present in smaller-scale models but emerge at scale. The thresholds of transparency can vary with industries and geographies. So, it is important for enterprises to establish an RGAI framework that can guide them through implementation.

Generative AI tools are built on a vast amount of visual or textual data, much of which is scraped from unknown sources on the internet. Due to the nature of data sets used in generative AI tools and potential biases in those data sets from unknown data sources, most RAI programs are unprepared to address the risks of new generative AI tools. Generative AI tools are qualitatively different from other AI tools. Rapid innovation and adoption of gen AI is unprecedented.

A study by MIT⁴ found that most RAI programs are unprepared to address the risks of new GenAI tools. However, early generative AI adopters at the enterprise level are creating their approach to responsible AI in parallel with building out use cases and MVPs.

Taking into consideration the evolving nature of the GenAI landscape, Genpact has formulated a Responsible Generative AI(RGAI) framework that is built on six key principles:



RGAI framework

Source: Genpact

There is no universal definition of responsible AI. That said, all the RAI endeavors commonly try to enforce fairness, transparency, and productivity. In the context of generative AI, RAI strategies should evolve too.

NASSCOM's guidelines define GenAI as a type of artificial intelligence technology that can create artefacts such as image, text, audio, video, and various forms of multi-modal content. The object of these guidelines is to promote and facilitate responsible development and use of GenAI solutions by different stake-holders⁵.

4. [Are Responsible AI Programs Ready for Generative AI? Experts Are Doubtful](#)

5. [Nasscom GenAI Guideline](#)

Here is a checklist that can help the leaders orchestrate an appropriate RGAI strategy:

1	Establish critical safety work streams	Evaluate your pre-deployment safety evaluation and adversarial testing to track the provenance of AI-generated outputs.
2	Foster RGAI awareness & human-centricity	In addition to incorporating transparency approaches, the organizations should reflect on the implications of the way they communicate generative AI with the stakeholders.
3	Develop a taxonomy of common objectives	Novel applications like generative AI require unique transparency goals. Amidst the diversity of stakeholders, it is important to define goals and create awareness across the organization.
4	Acquire appropriate XAI tool stack	It is important to incorporate the right libraries and frameworks that can assist in explainability. Organizations should look out for or maybe develop the SHAP and LIME equivalents of the generative AI era.
5	Embed reliability metrics	Formulate confidence score for the results generated and pass it through human evaluation.

A checklist of RGAI strategy

Source: Genpact

The background of the page is a solid dark blue. Overlaid on this are several concentric, wavy lines in a lighter blue and teal color. These lines originate from the left side and curve towards the right, creating a sense of depth and movement. The lines vary in thickness and spacing, giving the background a dynamic, organic feel.

Practitioner's Viewpoint

04

What Does it Take to Make a Winning GenAI Solution?

Finding the Right Model

Finding the Right Technique

Finding the Star Team

Finding the Right Tech Stack

LLMOps Current Landscape

LLMOps Implementation RACI Matrix.

Operationalizing LLMs with LLMOps

Finding the Right Metrics

Building the Guardrails: Policy Management

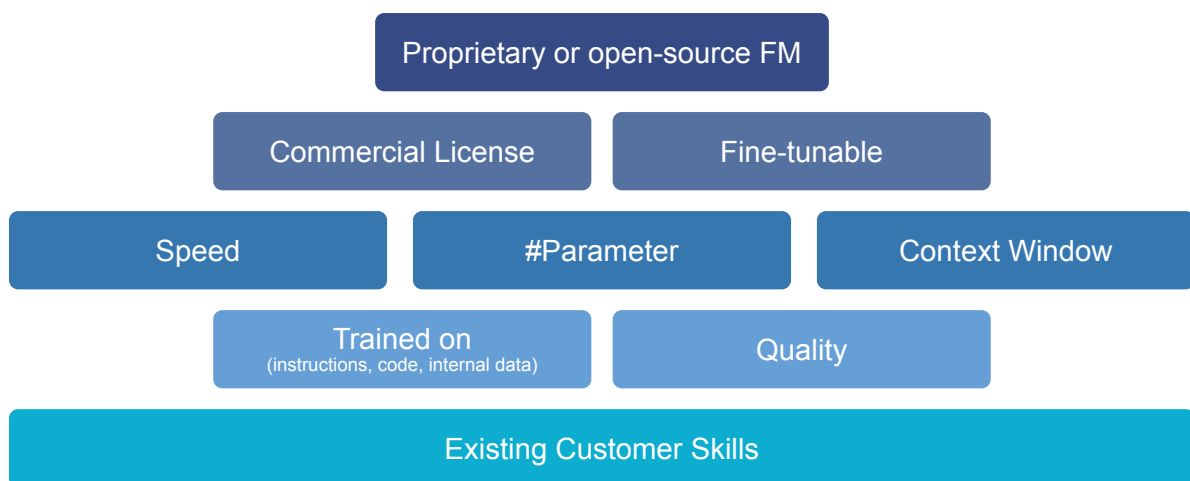


What does it take to make a winning GenAI solution?

In the early 90s, an analyst on Wall Street was expected to read 20,000 words every single day. The era of internet gets unleashed and by 2015, a typical analyst was devouring probably 200,000 words. In comparison, someone would only be able to go through ~100,000 words if they read at the average speed for 8 hours straight. Knowledge workers face similar information

challenges from spending three hours each day searching for information, flipping through 20 browser tabs, and reading five documents to find the one that matters. With GenAI, starting with the knowledge worker, everything is set to change. The immediate challenge for the enterprises is to find the right toolkit to implement GenAI.

Finding the Right Model



Attributes for selecting the right model

Source: Genpact

When selecting the ideal foundation models (FM), enterprises need to consider a variety of factors, including cost, latency, performance, privacy, and the training type (pre-trained or instruct-trained).

- **Proprietary vs. Open-Source:** Proprietary models often deliver superior performance and reliability but come at a cost. Whereas, open-source models are cost-free and offer flexibility, including fine-tuning.
- **Commercial License:** Verify licensing agreements to ensure alignment with commercial project goals.
- **Parameters:** More parameters typically mean a more powerful model but require greater computational resources.
- **Speed:** Larger models may process data slower, so consider latency in real-time applications.
- **Context Window:** Check for context window

size since larger context windows allow models to handle longer text sequences effectively.

- **Data Sources:** Investigate the model's training data sources and diversity to assess its suitability.
- **Data Quality:** Quality varies based on type, size, and training data, and is context-dependent.
- **Customization:** Assess whether fine-tuning is possible and its impact on customization and performance.
- **Domain Expertise:** Consider the team's expertise and familiarity with a particular model, as well as the availability of AI/ML experts

LLM selection Types of foundation models (open and closed source)

Feature	Open Source LLMs	Closed Source LLMs
Source Code Access	Source code is publicly available	Source code is proprietary
Customization	Flexible for customization	Limited customization options
Community Contribution	Active community involvement	Limited to internal development
Licensing and Costs	Permissive licenses, often free	Typically require licenses or fees

Types of Foundation Models

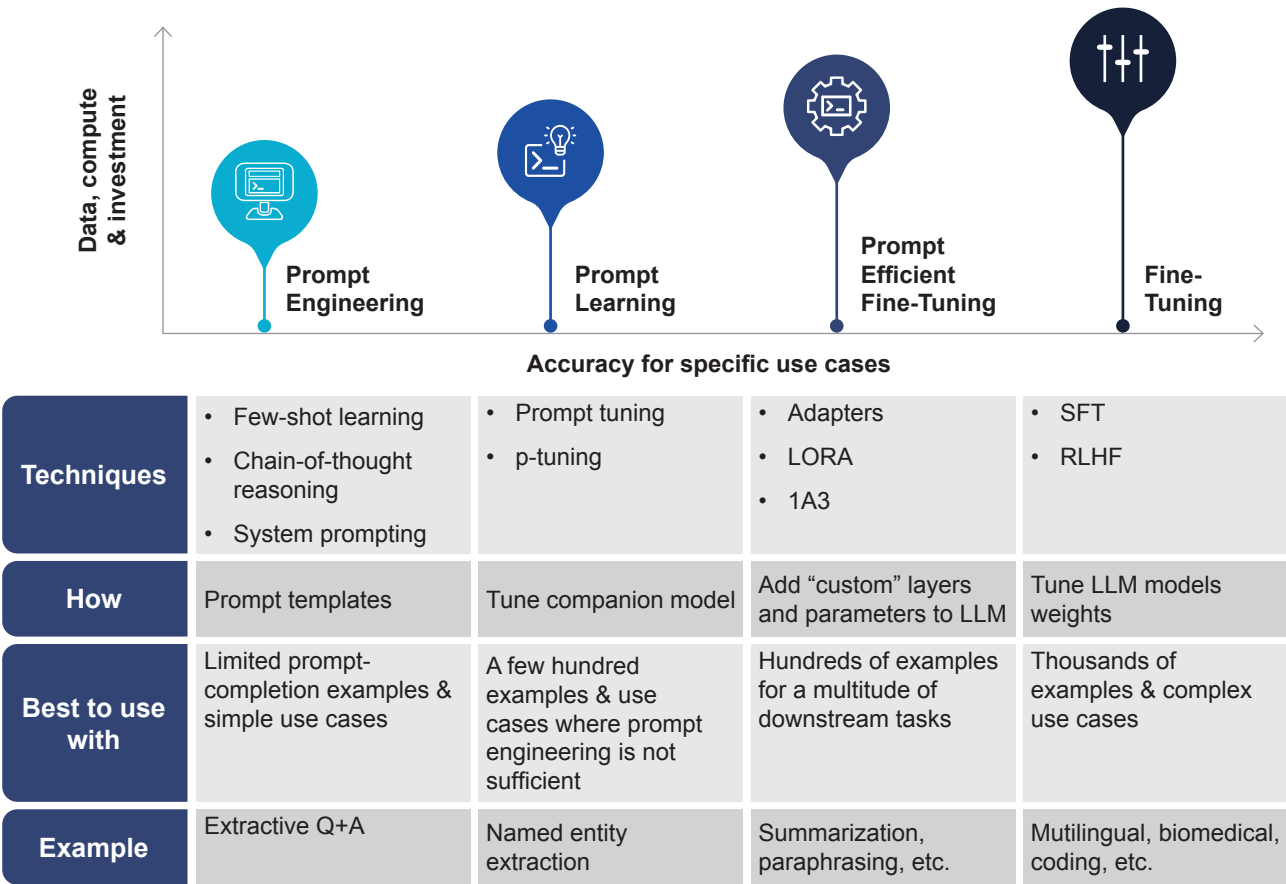
Source: Genpact

In summary, open-source LLMs offer transparency, flexibility, collaboration, and cost savings, while closed-source LLMs may deliver superior performance, IP protection, and technical support.

Models and their context window

Models	Maximum text length
gpt - 3.5 - turbo	4,096 tokens (~5 pages)
gpt - 4	8,192 tokens (~10 pages)
gpt - 4 - 32k	32,768 tokens (~40 pages)

Finding the Right Technique



Generative AI Techniques

Source: [NVIDIA](#)

It is important to understand that every technique comes with tradeoffs such as costs and accuracy. Choosing a technique is entirely upon the enterprise willingness to spend and the accuracy requirements of the use case under consideration.

- Prompt engineering involves supplying context, instruction, and examples to a model without altering its weights. The aim is to improve model output during inference, focusing on enhancing user experience.
- Fine-tuning LLMs refers to the process of adapting and optimizing pre-trained language models to better understand and process domain-specific data, ensuring enhanced performance in specialized fields.
- Prompt tuning combines prompt engineering and fine-tuning by using “soft prompts” generated with learnable parameters. These prompts can include additional words or AI-generated numbers in the model’s embedding layer. The goal is to adapt an AI foundation model to new tasks without the need for extensive retraining or weight updates, making it an efficient and cost-effective approach.
- Whereas RAG, short for Retrieval Augmented Generation, combines retrieval and generation to enhance LLMs, improving accuracy. It retrieves data from a text database to generate better responses. Used in legal summaries and chatbots, it shows promise but raises ethical concerns.

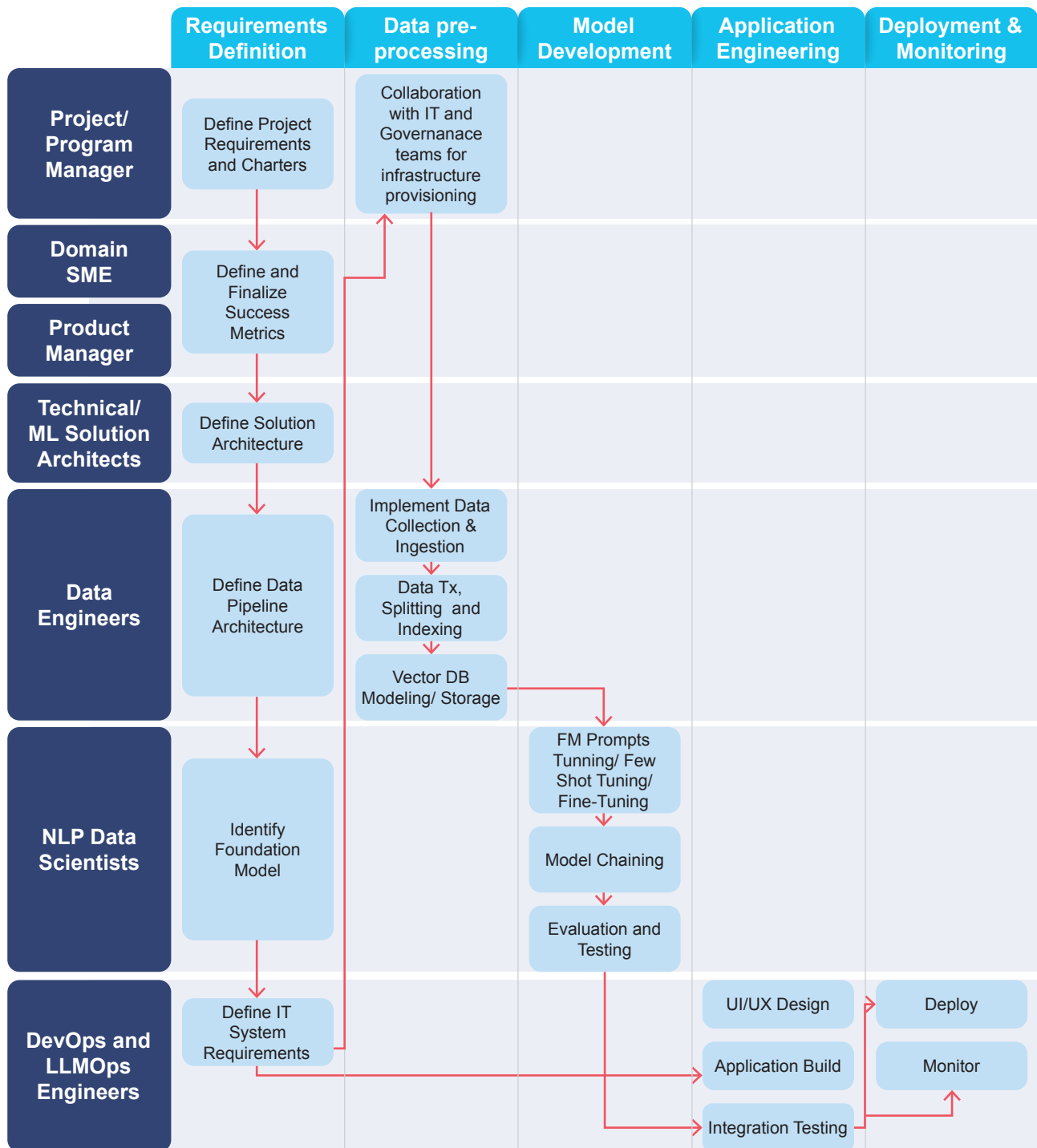
Cost Effectiveness	Risk Averse	Usefulness
<ul style="list-style-type: none"> • Custom solutions are tailor-made for critical downstream tasks in an enterprise. • Embedding, prompt tuning, and fine-tuning rake in their own costs. • Due to a lack of domain-based prompt engineering skills or extensive exposure or a combination of both, building Gen AI solutions can get expensive for enterprises. • Enterprises who are exposed to the varying cost cycles in production, can offer solutions that suit the use case while keeping it cost-effective. 	<ul style="list-style-type: none"> • Without the custom model, there are chances of enterprise data leakage to the global LLM model. • Publicly available applications like ChatGPT do not offer fact-checking, which a custom solution offers. This requires domain-based guard-railing efforts. • Custom solutions leverage cloud services. Hence the risk of leakage is mitigated via cloud security compliance. 	<ul style="list-style-type: none"> • The publicly available LLM interfaces will take your data as fresh data unlike in custom solutions where you can cache the results and combined techniques such as traditional embedding, prompt tuning, and fine-tuning. • The usage of these techniques depends on the use case. Selection of the technique requires exposure to solving real-world similar use cases. • Moreover. Reinforcement learning (RL) on enterprise data, considers an “informed user (SME)- a domain expert to perform tuning of the model outputs, unlike the RLHF of the ChatGPT or other such applications.

To harness GenAI for enterprises effectively, models must integrate proprietary data through techniques like fine-tuning or retrieval augmented generation (RAG)*. Often, this valuable data resides within documents spanning reports, presentations, and chat logs. The process of extracting this data into a usable format can be cumbersome. Take digital documents, for instance. While many documents use standardized file formats, each one contains infinite variations in formatting, content, and

layout. Developing pipelines to extract data from each of them would overwhelm data scientists and render a generalizable solution nearly impossible. This limitation has led document processing companies to specialize in specific document types or use cases, such as invoices, NDAs, or contracts.

* Check the Glossary to know more about the techniques

Finding the Star Team



Generative AI Roles and Responsibilities

Source: Genpact

Generative AI is poised to disrupt mundane tasks. However, setting up architectures to support gen AI would require additional skills. Employees should be equipped with knowledge of prompt engineering, prompt evaluation, prompt drift,

prompt injections, and guard railing mechanisms; and can work with chief policy officers, chief risk officers, and data ethnographers. A whole new gamut of job roles must be integrated with existing teams.

Building next generation enterprise grade AI solutions would require cross functional teams embedded with new roles. For instance, a domain expert, like a distinguished medical practitioner, must now embrace a foundational grasp of the technical bedrock that underpins these foundation models. This is no mere supplementary skill; it is the linchpin that unveils the transformative potential of these models, rendering them indispensable allies in your organizational toolkit.

Furthermore, as we navigate the expansive

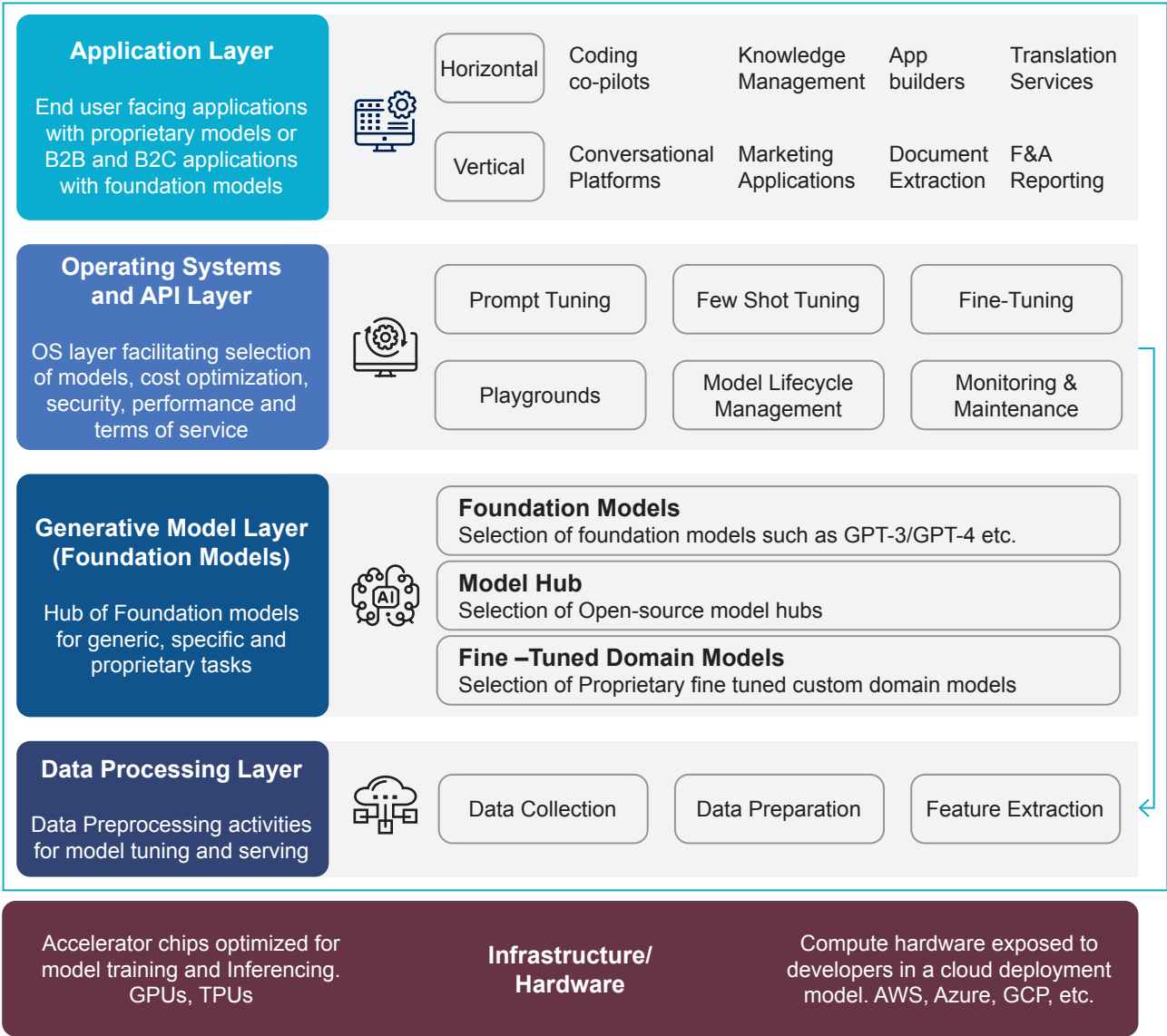
terrain of GenAI, a new vista of possibilities unfurls before us. Consider this akin to the emergence of novel disciplines within your organizational framework, embodied by linguistics experts, AI quality controllers, AI editors, and prompt engineers. In domains where GenAI illuminates the path ahead, it becomes imperative to methodically dissect existing roles into their elemental constituents. GenAI invigorates each task – orchestrating the symphony of full automation, harmonizing in augmentation, or preserving the sanctity of untouched tasks, all with the precision of a seasoned consultant.

1	Data Labelers and Editors	These individuals specialize in data annotation tasks, including labeling data pairs such as text and images or preparing unannotated free-text data. Their role is pivotal in supporting the advanced analytics team and optimizing data lake environments.
2	Fine-Tuners	These experts possess deep knowledge of Feature Models (FMs) and excel at the art of fine-tuning them. They expand the capabilities of the data science team, which is primarily focused on traditional machine learning.
3	Generative AI Developers	These professionals are highly skilled in the selection of Feature Models (FMs), the crafting of prompt sequences, and the refinement of input and output processes. They form a distinct team known as the Generative AI Application Team.
4	Prompt Engineers	These individuals are responsible for designing input and output prompts to customize the solution to specific contexts. They also play a pivotal role in testing and establishing the initial version of the prompt catalog. They are key members of the Generative AI Application Team.
5	Prompt Testers	This group conducts comprehensive testing of the Generative AI solution, encompassing both backend and frontend components. Their findings contribute to the improvement of the prompt catalog and evaluation dataset. They are essential within the Generative AI Application Team.
6	AppDev and DevOps	These professionals are tasked with the development of frontend components, such as websites, for the Generative AI application. They are integral to the functionality of the Generative AI Application Team.
7	RLHF Red Teams	These teams are run by humans who are domain experts. These experts flag the outputs of the LLMs and help adhere to the guardrails in place.

Emerging new roles in the era of GenAI

Source: Genpact

Finding the Right Tech Stack



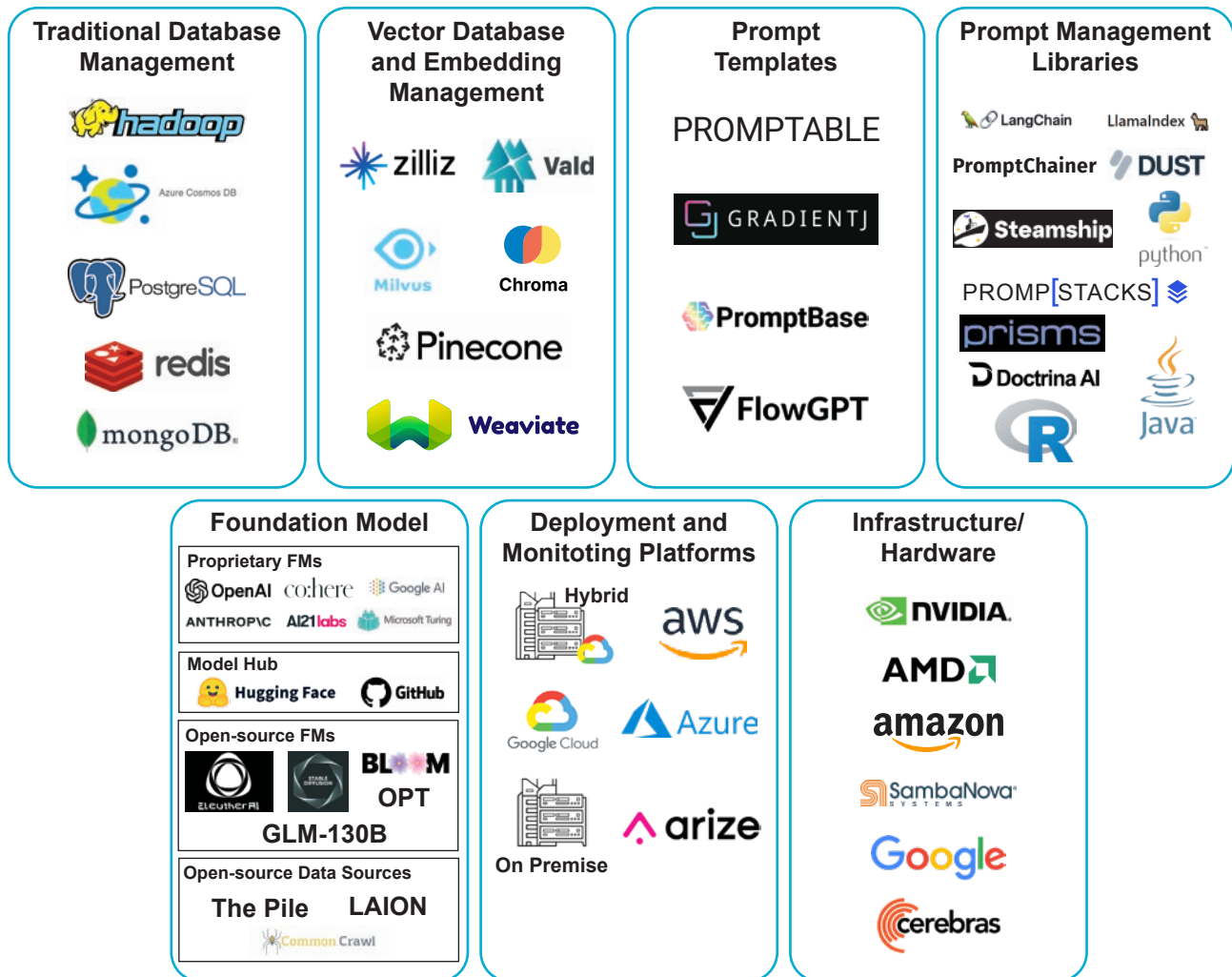
Generative AI Enterprise Stack

Source: Genpact

GenAI’s operational logistics, facilitation is provided by various tool vendors offering essential support in configuring, executing, and monitoring the end-to-end workflow. This underlines the importance of having a right tech stack that connects various layers; layers that

are specific to GenAI pipelines. Like any typical AI tech stack, GenAI application too, have an application layer, that communicates with the model hub, which in turn communicates with the data layer. Here’s a brief overview of a typical GenAI tech stack.

LLMOps Current Landscape



Source: NASSCOM, Genpact

Picking up the Hardware

Some enterprise may find it necessary to run their own AI models on GPUs. This could be due to the AI model being the core product or the need for fine-tuning tiny language models. In most cases, cloud solutions suffice. Major cloud providers like AWS, Azure, and GCP typically offer instances with dedicated GPUs, while specialized AI clouds provide alternative models like containers or batch jobs, which can reduce costs. Network bandwidth is crucial for training, and some models require dedicated

interconnects between nodes. The selection of GPUs depends on the specific application, taking into account factors like training versus inference, memory requirements, hardware support, latency needs, and the variability of demand. Software optimizations can substantially influence model performance. For instance, techniques such as using shorter floating-point representations, quantization, neural network pruning, and hardware-specific optimizations can improve efficiency.

GPT-3 has approximately 175 billion parameters, which for an input and output of 1,024 tokens, results in a computational cost of approximately 350 trillion floating point operations. (Source: [a16z](#))

Building Data Pipelines



Data layer is concerned with data gathering and data cleaning pipelines. The heterogeneity of data sources can make it cumbersome so this is a key step in the GenAI pipeline. However, it is to

be noted that GenAI data layer avoids typical NLP data pre-processing steps such as stemming, lemmatization and stop words removal.

Orchestrating the GenAI Engine



A foundation model (FM) by definition is considered to be good for diverse downstream tasks. Since most enterprises deal with niche use cases from multiple industries, a domain expert and prompt engineer would create prompts and additional data that would fine tune and help the FM to deliver desired output.

The next important step is choosing the right vector database. In the past, vector databases found applications in the search domain. Nevertheless, the emergence of ChatGPT has underscored the potential of vector databases to amplify the capabilities of large language models (LLMs).

Vector Databases are often used to store and manage vector representations of data, which may include embeddings of text, images, or other features. These vectors may represent context or information that the model needs to generate coherent and contextually relevant responses.

Once the cleaned data is stored in the vector database, the next step is to select the appropriate model. Depending on the cost-accuracy trade-offs, a suitable model is selected. For instance, an enterprise trying to build a tiny language model would prefer to fine-tune a open source model. Whereas, use cases which do not restrict data transfer can leverage the specialized cloud AI services.

Using the right technique, as discussed above, the instructions that dictate model behavior is decided and integrated with the backend.

The engineered instructions and the data stored in the vector database once deployed can be leveraged for querying, summarization and other tasks. The generated outputs are then evaluated for its accuracy and usability. This layer undergoes a iterative process of re-training to maintain the usability. This is done through RLHF or reinforcement learning through human feedback in combination with automated prompt tuning techniques.

User-Interfacing



The GenAI engines are then connected to the front end which are interfaces wrapped over APIs that communicate with the FMs to generate the desired output.

Traditional databases primarily retrieve rows where the queried value exactly matches the search term. In contrast, vector databases utilize similarity metrics to find the vector most relevant to the query. Vector indexing is done with the help of various algorithms.

Checklist for choosing between vector databases and vector libraries

Updatability

The database's ability to efficiently accommodate real-time updates and modifications to data is essential for maintaining LLMs adaptability.

Speed

Rapid data retrieval and processing is key for GenAI workloads to ensure swift access to vector representations of language and features, consequently improving model training and inference times.

Crash Recovery

Robust crash recovery mechanisms are crucial for maintaining the integrity of datasets and preventing data corruption.

Sharding

Sharding, involves horizontally partitioning of data across multiple servers to enhance both the scalability and performance of the LLMs.

Replication

Replication is essential in vector databases to ensure fault tolerance, and high availability while reducing the downtime in GenAI applications. Database replication improves reliability, scalability and performance

Execution

Efficient query execution capabilities determine how quickly LLMs can access and process vectors and features, directly affecting the overall performance and responsiveness of the applications.

Multi-tenancy

Multi-tenancy support is invaluable for maintaining data isolation and security, which become critical at scale as enterprise users get onboarded. The definition of scale quickly shifts from number of vectors to number of individual tenants.

For GPT-3.5-turbo using ~1,000 tokens per query, it costs ~\$0.002 per query, or ~500 queries per dollar (as of Apr 2023)
For GPT-4, again assuming ~1,000 tokens per query, it costs ~\$0.03 per query, or ~30 queries per dollar (as of Apr 2023)

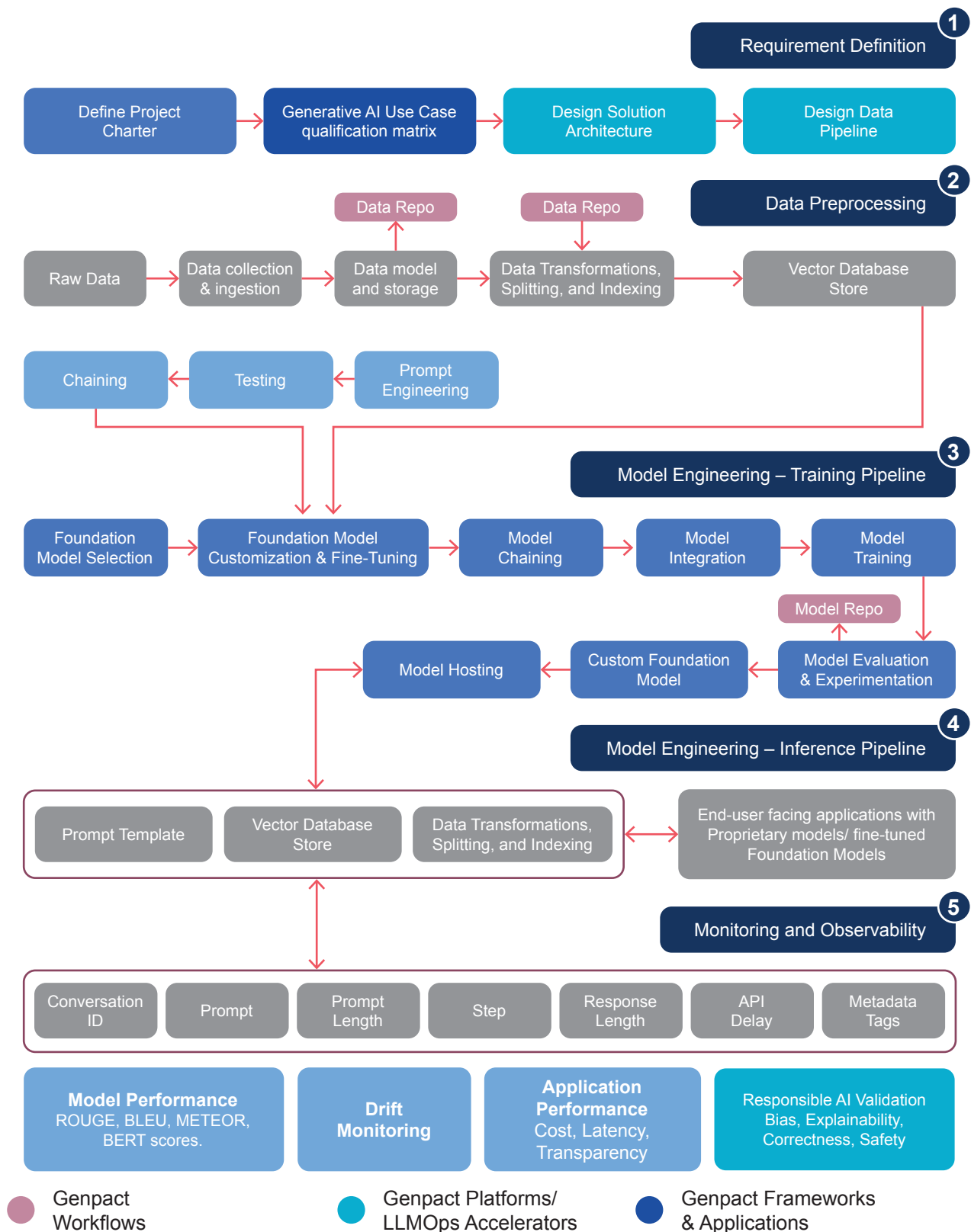
LLMOps Implementation RACI Matrix

Task	Sub-task	Stakeholder					
		Business Owner/ Leaders	Domain Subject Matter Experts	Technical Architects	Data Engineers	NLP Data Scientists	LLMOps Engineers
Requirement Definition	Define Business Requirements	R A	I	I	I	I	I
	Define Tech Stack	I	C	R A	C	C	C
Design & Architect	Design Solution Architecture	I	C	R A	C	C	C
	Design Data Pipeline	I	C	C	R A	C	C
Data Pre-processing	Data Collection, Ingestion, Storage	I	C	C	R A	C	C
	Data Transformation and Storage	I	C	C	R A	C	C
Model Engineering - Training Pipeline	Foundation Model Selection	I	C	C	C	R A	C
	Model Customization/ Fine-tuning/ Prompt Tuning	I	C	C	C	R A	C
	Model Chaining and Integration	I	C	C	C	R A	C
	Model Training, Evaluation and Experimentation	I	C	C	C	R A	C
Model Engineering - Inference Pipeline	Model Hosting/ Inference Pipeline Engineering	I	C	C	C	C	R A
	End User Application Hosting	I	C	C	C	C	R A
	Integration Testing, UAT	I	C	C	C	C	R A
Monitoring & Observability	Prompt Response Monitor Pipeline Engineering	I	C	C	C	C	R A
	Pipeline Integration	I	C	C	C	C	R A
	System Validation and Post Production Monitoring	I	C	C	C	C	R A

R Responsible
 A Accountable
 C Consulted
 I Informed

Source: Genpact

Operationalizing LLMs with FMOps



FMOps implementation framework

Source: Genpact

In traditional machine learning (ML), a combination of people, processes, and technology is key to productizing ML use cases. However, generative AI introduces new complexities.

Unlike classic ML and MLOps, FMOps and LLMOps differ in four main areas: personnel and processes, FM selection and adaptation, FM evaluation and monitoring, data privacy, model deployment, and technology requirements.

Generative AI developers integrate the chosen FM into solutions, collaborating with prompt engineers to automate converting end-user input into appropriate FM prompts. Prompt testers contribute entries to the prompt catalog for automatic or manual testing. GenAI developers establish prompt chaining and application mechanisms for the final output, where prompt chaining breaks complex tasks into smaller, more manageable sub-tasks, enhancing context-awareness. They also implement input and output monitoring, such as toxicity detection, and introduce a rating mechanism to enrich the evaluation prompt catalog with examples. To provide functionality to end-users, a frontend website is developed to interact with the backend. DevOps and cloud application developers (AppDevs) follow best practices

to implement input/output and rating features. Beyond basic functionality, the frontend and backend must support personal user accounts, data uploading, black-box fine-tuning initiation, and the use of personalized models instead of the base FM.

The FMOps pipelines come with a new application layer that serves as the workspace where GenAI developers, prompt engineers, testers, and AppDevs collaboratively build the backend and frontend components of generative AI applications. While GenAI end-users access the front end of these applications, typically through a web-based user interface, data annotators must be able to perform data preprocessing tasks without directly accessing the backend data storage infrastructure. Consequently, a secure web-based user interface (website) with integrated editing capabilities becomes essential for their interactions with the data.

LLMOps Implementation Challenges and Best Practices

Stage 1 - Requirement Definition

Sub - Stage	Challenges	Best Practices
Defining a Project Charter	Defining project goals, scope, and success metrics	Involve key stakeholders and subject matter experts to create a well-defined project charter with clear objectives and expectations.
Generative AI Use Case Selection	Identifying GenAI use case value generation and implementation complexity	Assess use case viability through a use case qualification matrix that considers all the salient points indication risk and ROI.
Solution Architecture	Designing a scalable and cost-effective architecture	Collaborate with SMEs and ML engineers to develop robust solution architecture
Design Data Pipeline	Determining data sources, integration methods, and data flow	Develop a comprehensive data strategy and identify data sources.

Stage 2 - Exploratory Data Analysis (EDA)

Sub - Stage	Challenges	Best Practices
Data Collection	Contextual Data to train the LLMs needs to be made available to reduce hallucinations and make the outcomes contextualized to the business at hand.	<ol style="list-style-type: none">1. Implement automated data collection scripts from enterprise systems and databases to retrieve contextualized data.2. Normalize data formats, and use data connectors that can handle various sources seamlessly.
Data Transformation	Cleaning noisy text data and handling missing data. This includes removing errors, correcting inconsistencies, and removing duplicate data.	Leverage imputation techniques to clean data and for utilize a combination human annotated and synthetic data to compensate for the missing data.
Data Storage	Conventional database storage systems are not adequate, optimized and fast to retrieve unstructured data stored in the form of embeddings	Leverage Vector databases to store such data. Employ scalable architecture to implement Vector Databases and optimize the Vector databse for fast retrieval.

Stage 3 - Data prep and prompt engineering

Sub - Stage	Challenges	Best Practices
Data Preparation	The data for training a LLM is required to be tokenized and normalized.	<p>Data Versioning: Maintain a record of data changes and monitor development through comprehensive data versioning practices.</p> <p>Data Encryption and Access Controls: Safeguard data with encryption and enforce access controls, such as role-based access, to ensure secure data handling.</p>
Data Engineering	Prompt engineering is the process of creating prompts that are used to generate text with the LLM. The prompts need to be carefully crafted to ensure that the LLM generates the desired output.	<p>Prompt Template: Develop reliable prompts to generate accurate queries from LLMs, facilitating effective communication. Make use of prompt templates for effective capture and utilization of Prompts.</p> <p>Few-Shot Prompting: Leverage few-shot learning to expedite model fine-tuning for specialized tasks without extensive training data, providing a versatile and efficient approach to utilizing large language models.</p>

Stage 4 - Model fine-tuning

Sub - Stage	Challenges	Best Practices
Model Training	Once the data is prepared, the LLM is trained. The Data scientist has to choose between prompt tuning, few shot tuning, fine tuning a LLM to create custom trained model.	Optimize model performance using established libraries and techniques for fine-tuning, enhancing the model's capabilities in specific domains with contextualized data specific to the business function and industry. Carefully assess between performance and cost to identify which stage of tuning is best suitable for the problem at hand.
Model Evaluation	Once the LLM is trained, it needs to be evaluated to see how well it performs. This is done by using a test set of data that was not used to train the LLM.	Establish comprehensive data and model monitoring pipelines, including alerts to identify model drift and potentially malicious user behavior. This proactive approach enhances model reliability and security.
Model Fine-tuning	If the LLM does not perform well, it can be fine-tuned. This involves adjusting the LLM's parameters to improve its performance.	Optimize model performance using established libraries and techniques for fine-tuning, enhancing the model's capabilities in specific domains.

Stage 5 - Model review and governance

Sub - Stage	Challenges	Best Practices
Model Review	A fine tuned LLM is required to be reviewed at regular intervals to ensure that it is safe and reliable. This includes checking for bias, safety, and security risks.	Develop robust data and model monitoring pipelines that raise alerts for detecting model drift and identifying potential malicious user behavior. Set thresholds for metrics to identify any toxicity, bias in the output.
Model Governance	A LLM must be tracked for performance, making changes to it as needed, and retiring it when it is no longer needed.	Define and set metrics that must be captured at each run of the model in production, to capture the prompt and response alongwith the latency and metadata. Set thresholds and trigger alerts to re-train the LLM for any deviations.

Stage 6 - Model inference and serving

Sub - Stage	Challenges	Best Practices
Model Inference	Defined architectural patterns need to be identified to deploy the fine tuned models in production. Alternatively, issues such as concurrent users, scalability and performance considerations have to be managed for off the shelf LLM usage as well.	Choose the appropriate deployment strategy based on considerations such as budget, security, and infrastructure requirements.
Model Serving	Identify the design pattern to make the LLM available to users.	Tailor pre-trained models for specific tasks, as this approach is cost-effective. It also applies to customizing other machine learning models like natural language processing (NLP) or deep learning models.

Stage 7 - Model monitoring with human feedback

Sub - Stage	Challenges	Best Practices
Model Monitoring	Setting up of policies, procedures and methodologies in conformity with various stakeholders in the organization to track LLM application performance, identifying any problems, and making changes as needed.	Create a playbook for involving IT, Infosec, Legal, Business Owners and ML Practitioners to arrive at a consensus of parameters that need to be monitored. Create a system with a combination of Human+Machine to capture machine and Human feedback to improve the performance of the LLM or trigger alerts as required.

Finding the Right Metrics

GenAI evaluation requires metrics* in addition to the metrics used for traditional ML solutioning.

1

For Discrete Outputs (e.g., sentiment analysis)

Employ accuracy metrics like precision, recall, and F1 score

2

For Unstructured Text Outputs (e.g., summarization)

Utilize similarity metrics such as ROUGE and cosine similarity. However, some cases lack a single correct answer, making model evaluation challenging without labeled test data.

Here are two approaches:

1

Human-in-the-Loop (HIL)

A team of prompt testers assesses model responses, either extensively or through sampling, depending on application importance.

2

LLM-powered Evaluation

An LLM assesses all model-generated responses, offering a faster, albeit potentially lower-quality, evaluation.

Evaluated prompts can be fed to top FMs, resulting in an evaluation result dataset with prompts, FM outputs, labels, and scores. For unlabeled evaluations, review outputs, providing scores and feedback. Aggregated results allow users to benchmark model quality. Post-evaluation, model selection depends on criteria discussed in the previous sections.

* Check the glossary for metrics

Building the Guardrails: Policy management

These are a few questions that would help the executives guide their teams. The next step would be to checks and balances that can allow the organization to make the most of the generative AI tools.

Here's a checklist for building LLM-based solutions with a focus on ethics and RAI:

- 1. Facilitate the creation of RAI focus groups:**
Establish a hybrid workforce consisting of ethical, legal, and public domain experts who can collaborate on enforcing RAI principles throughout the organization.
- 2. Make Solutions Responsible by Design:** Choose high-quality, diverse, and representative data that aligns with ethical values and principles. Design models that are transparent, explainable, and fair while avoiding bias and unintended consequences. This includes setting standards for data privacy, transparency, and fairness in decision-making processes.
- 3. Implement accountability mechanisms:** Enhancing accountability via methods, such as conducting audits or enabling users to challenge decisions and offer feedback can shape decision making.
- 4. Foster collaboration and transparency:** It is of important to implement safeguards that guarantee the privacy of individuals, as LLMs have the capacity to retain or leak sensitive data they have been exposed to. This is why offering transparency and collaborating with the user will lead to overall improvement in the model outcomes.
- 5. Build a Robust Monitoring Mechanism:** Having a monitoring mechanism will ensure that the model takes continuous real-world performance and user feedback into account. This mechanism will also help in analyzing how the deployed model would deviate from the desired outcome.
- 6. Embed RAI Governance Mechanisms:** Implement governance mechanisms to promote accountability and oversight, such as oversight committees, review processes, and protocols for responding to feedback and complaints.
- 7. Promote Explainability and Interpretability:** Collaborate closely with domain experts to determine the essential interpretability features for the model.

Key questions for the C-suite

- Q1. What are the limitations of the generative AI tools that I want to use?
- Q2. What are the ingredients of the model that sits behind these tools?
- Q3. How transparent has been the creator of these tools?
- Q4. What legal and ethical compliances would arise with potential downstream tasks?







06

Gen AI: 2023 & Beyond

GenAI Trends of the Future

GenAI Trends of the Future

Impact Lens	Key Drivers and Outlook			
 Change in economics	T&M: Higher productivity ★ ▲	FP: Productivity & discount ★		FP: Productivity gain retained ▲
	15-20% cost reduction at the same gross margin	15-25% direct cost reduction	50-60% passed as customer discount	20-30% of direct cost reduction to be retained by ITSP 20-25% of gains passed on to partners as cost of tech
 Scope compression	Substitution of existing revenue ★ ▲	Insourcing by enterprises ★ ▲		Erosion of existing revenue ★ ▲
	3-4% of existing revenues flow to partners (substitution of services)	2-5% of existing revenue is insourced by enterprises		2-3% loss with automation of workflows/services
 Growth in services/ offerings	New revenue opportunities ★	Higher-margin revenues ▲		
	15-25% of revenue from new/ reimagined offerings/ volume amplification of existing offerings	38-40% margins due to new offerings led by differentiated talent, proprietary assets		
 SG&A upside	Productivity improvements ▲			
	10-15% reduction in SG&A costs			

Generative AI Key Drivers & Outlook

Source: NASSCOM

GenAI has had a more promising beginning than Software as a Service (SaaS), generating over \$1 billion in revenue from startups alone—a milestone that took SaaS years to achieve. Innovative reasoning techniques, such as chain-of-thought, tree-of-thought, and reflection, are enhancing models' capacity to tackle complex reasoning tasks and will bridge the gap between user expectations and model capabilities. Today, developers are utilizing new frameworks and techniques to execute and troubleshoot intricate multi-chain sequences.

For instance, RAG or Retrieval-augmented generation introduces context about businesses or users, reducing inaccuracies and enhancing the authenticity and utility of generated content. GenAI is transitioning from zero-shot interactions to ask-and-adjust dynamics. Enterprises are moving beyond individual user workflows to address system-wide optimization.

Going forward, the landscape of GenAI will see more enterprise use cases performing crawling & indexing workspaces and creating

hyper-personalized digital companions in a human-centric way. There will be a surge in developer tools and application frameworks, such as LLMOps, serving the ecosystem with reusable components to build more advanced AI applications. We will also see AI-first infrastructure companies offering compute at reasonable costs along with on-demand availability and scalability.

Companies like NVIDIA are in the process of manufacturing AI hardware that can help users migrate their AI workloads from expensive data centers to their workstations under their

desks. Now, enterprises might see opportunities opening into platform agnostic GenAI solutioning with open source foundation models at their disposal for inexpensive experimentation and development of EnterpriseLLM.

GenAI will continue to nudge the business processes to a transformative zenith, catalyzing unprecedented efficiency in both back-end and front-end operations. Ultimately, it will bolster enterprise governance and fortify information security, safeguarding against fraud and enhancing regulatory compliance.



The generative AI era
is upon us, the iPhone
moment if you will.

~ Jensen Huang, CEO, NVIDIA

GenAI Trends of the Future

Dawn of AI-first Marketplaces

Generative AI will transform personalization, will attract more users, command higher prices, and enhance retention and engagement. The current price gap between digital and in-person services presents a pricing opportunity for AI apps. As GenAI capabilities improve, ushering in a new era of consumer services.



Rise of TinyLLMs

Open-source FMs such as Llama 2 from Meta and Technology Innovation Institute's Falcon are capable of offering organizations additional flexibility to balance cost against requirements, transparency that can be important in regulated industries, and clearer pathways to innovation, such as creating useful model variants efficient enough to run on local devices rather than in the cloud.



Rise of AgentOps

The existing agent system relies on natural language as a bridge between Large Language Models (LLMs) and external components like memory and tools. The autonomous agents use LLMs as their brains and then perform planning, task decomposition, reflection, and execution. AgentOps will also feed on the evolution of LLMOps, which again is a consolidation of best operationalizing practices through the Responsible GenAI (RGAI) lens.



Rise of Enterprise Co-pilots

Fine-tuning the FM on enterprise data will make the outputs accurate. However, this strategy will falter in the long run. The GenAI engines should have the flexibility to ingest new data. For example, from incoming data from enterprise software like Salesforce, JIRA etc. For generative AI to achieve its immense potential, it needs to be broadly accessible and easy to integrate into a range of services.



LLM Agnosticism

Hybrid strategy in LLM selection where open-source vs closed-source dichotomy boiling down task performance (example: summarization vs. Q&A or depending upon risk factors of the domain: healthcare vs retail). Enterprises will slowly move on to leveraging an ensemble of LLMs to generate most effective outcome.



Glossary

a

Alignment

AI alignment aims to ensure AI systems achieve intended goals and ethical principles, preventing misalignment where they perform different objectives.

Attention

A mechanism in the Transformer model enabling it to focus on different parts of input during encoding or decoding.

Agents

General AI tools like Siri; LLM-based examples include LangChain and AutoGPT.

b

Base Model

A pre-trained language model serving as a foundation for further fine-tuning.

BLEU (Bilingual Evaluation Understudy)

BLEU compares the

generated output to reference translations, quantifying the similarity between them. Higher BLEU scores indicate better performance.

c

Corpus

A collection of text documents for analysis or training the LLM.

Chain-of-Thought Prompting

Enhancing reasoning in LLMs by having them generate a series of intermediate steps for multi-step problems.

Context Window / Context Length

The number of tokens considered when predicting the next token.

CLM (Causal Language Modeling)

A pretraining task where a model reads text sequentially and predicts the next word.

Chain-of-Thought

Enhances LLM reasoning by breaking tasks into smaller

steps, with recent “tree-of-thought” allowing branching and back-tracking.

d

Decoder-only Models

Models that use only a decoder to generate output text based on autoregressive processes.

Diversity

Diversity metrics assess the variety and uniqueness of generated responses, analyzing factors such as n-gram diversity and semantic similarity between responses. Higher diversity scores indicate more diverse and distinct outputs.

e

Embedding

A vector representation of a word or token in a continuous space.

Emergent Abilities

Abilities that appear in larger models but not in smaller

ones, which can then be used for developing solutions in diverse downstream tasks.

Encoder-only Models

Models that use only an encoder to process input text and create representations for downstream NLP tasks.

Encoder-Decoder Models

Models with both encoder and decoder components for sequence-to-sequence tasks like translation.



Few-shot Prompting/ Learning

A technique where more prompts are provided to enable in-context learning and guide the model well towards a desired output.

Fine-tuning

Adapting a pre-trained model to a specific task by training on task-specific data.

Frequency Penalty

Reduce the chance of repeating a token proportionally based on how often it has appeared in the text so far. This decreases the likelihood of repeating the exact same text in a response.



Grounding

Grounding associates words and phrases with real-world entities and concepts, enhancing AI's consistency, accuracy, and informativeness.



Hallucination

In AI, hallucination is when an AI confidently responds with information not supported by its training data.



Instruction Tuning

Fine-tuning LLMs on formatted instances in the form of natural language instructions.



Low-rank Factorization

Low-rank factorization decomposes weight matrices of a neural network into smaller matrices with lower rank to reduce the number of parameters and computations.

Low-Rank Adaptation (LoRA)

A memory-efficient fine-tuning method that accelerates model tuning, reducing resource consumption.



Modality

High-level data categories like text, numbers, images, video, and audio.

MLM (Masked Language Modeling)

A pretraining task where a model predicts original text from a corrupted version with masked tokens.

Max Response

Set a limit on the number of tokens per model response. The API supports a maximum of 32768 tokens shared between the prompt (including system message, examples, message history, and user query) and the model's response. One token is roughly 4 characters for typical English text. response. One token is roughly 4 characters for typical English text.

Multi-modal

Unlike most models that work with text or images, multi-modal models handle various types of mixed media inputs, such as text and images, enabling versatile processing.

O

One-shot Prompting

Provide the model with single prompt example as a foundation to answer the prompt query.

P

Prompt Engineering

Designing effective natural language prompts to guide large language models in generating relevant task-specific content.

Pre-training

Training a language model on vast amounts of unlabeled text data to learn language structure and patterns.

Prompt Injection

A technique used in platforms or products powered by LLM technology, like chatbots or GitHub Copilot, where inputs can be manipulated to override the prompt or system instructions. This can reveal internal information or disrupt alignment training, known as “jailbreaking.”

PEFT (Parameter-Efficient Fine-Tuning)

Fine-tuning methods that adapt pre-trained models using a small number of additional parameters.

Pruning

Pruning involves removing unnecessary connections or nodes from a neural network to reduce its size while maintaining accuracy.

Plugins / Tools

LLM agents access APIs for added capabilities, reducing risk of errors.

Presence Penalty

Reduce the chance of repeating any token that has appeared in the text at all so far. This increases the likelihood of introducing new topics in a response.

Perplexity

Perplexity measures an LLMs’ ability to predict a given text sample. A lower perplexity value signifies superior performance.

Q

Quantized Low-Rank Adaptation (QLoRA)

A quantized version of LoRA to reduce memory demands.

Quantization

Quantization reduces the precision of model weights and activations from floating-point numbers to fixed-point numbers with fewer bits to save memory and computational resources.

R

RLHF (Reinforcement Learning from Human Feedback)

Training models by using human feedback to create a reward model for reinforcement learning.

Retrieval Augmented Generation (RAG)

Enhances prompts with web-searched or internal data for better context and content quality.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE encompasses a collection of metrics for evaluating summary quality by comparing generated summaries to reference summaries. It calculates precision, recall, and F1-score, offering insights into the language model’s summary generation capabilities.

S

Self-attention

A key component in Transformers, where input is transformed into queries, keys, and values used to compute attention scores.

Self-supervised Learning

Training a model using unlabeled data by predicting some aspect of the input.



Token

A basic unit of text, representing a word or subword.

Temperature

A hyperparameter controlling the randomness of a model's output, with 0 making it deterministic.

Transformer

A neural network architecture popular in NLP, consisting of components like input embeddings, positional encoding, encoder, decoder, and output layer.

Top P

Similar to temperature, this controls randomness but uses a different method. Lowering Top P will narrow the model's token selection to likelier tokens. Increasing Top P will let the model choose from tokens with both high and low likelihood. Try adjusting temperature or Top P but not both.



Vector Space Model

Represents text as vectors in a high-dimensional space.

Vision Transformer (ViT)

A model for image recognition using Transformer architecture.



Zero-shot Prompting

Prompting a model to respond to tasks that the model is unfamiliar with.

“

This compendium helps the readers especially C-suite executives to navigate through the Generative AI turbulence that has overwhelmed the IT industry in the past year or so. This report consolidates the expert views and industry-wide surveys founded on ground realities. Large language models are no longer the pet projects of esoteric research labs. They have become a household name amongst the enterprises. This study serves as a companion guide to building ethical and feasible Generative AI solutions at scale. It offers a blueprint for building tool stacks, frameworks and governance mechanisms that will bring the enterprise grade pilot projects to production.

”



Global AI/ML Leader,
Genpact

Contact Information

ANKIT BOSE
Head of NASSCOM AI



<https://www.linkedin.com/in/ankit-bose-78b46912/>

nasscom ai



genpact