tmforum knowledge

implementing GenAI

how to set up and maintain large language models

Author: Dean Ramsay, Principal Analyst
Editor: Ian Kemp, Managing Editor

Sponsored by: TCS TATA CONSULTANCY SERVICES

contents

03 setting the scene

05 chapter 1:
what is a GenAI large language model (LLM)?

07 chapter 2:
setting up telecoms LLMs

12 chapter 3:
maintenance of LLMs

15 chapter 4:
evolving standards for AI

17 additional feature from Tata Consultancy Services

23 meet the Research and Media team

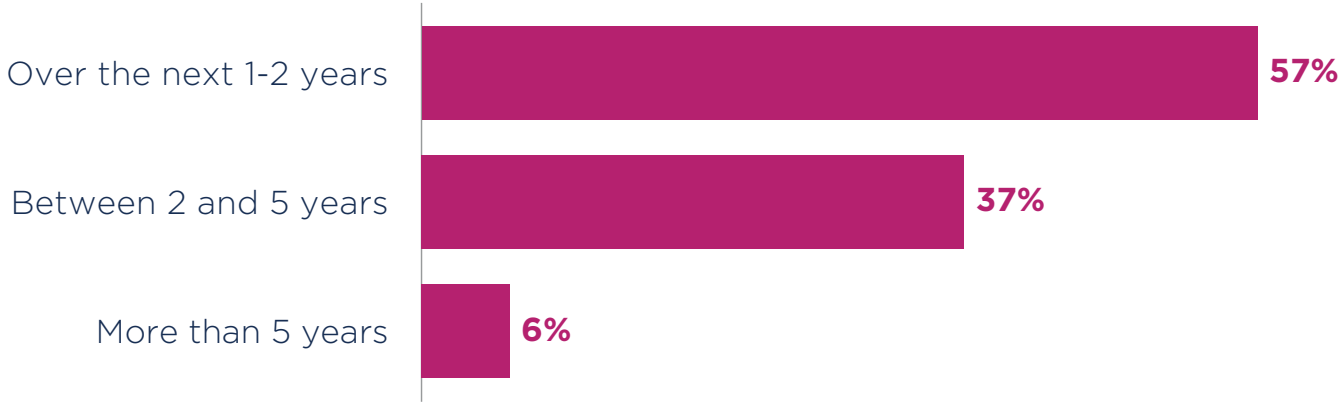
setting the scene

The creation and integration of generative AI (GenAI) models into telcos' business processes is a complex undertaking. As communications service providers (CSPs) move from proof-of-concept projects to live deployments it is becoming clear that constructing the right language model is key to success.

Some CSPs are setting out to develop large language models (LLMs) specifically for the telecoms industry as we have outlined in our benchmark report, Building an AI strategy: telcos put the foundations in place. The Global Telco AI Alliance, for example – whose founders are e&, Deutsche Telekom, Singtel, SK Telekom and Softbank – aims to develop multilingual LLMs for operator businesses globally.

Our recent survey – compiled for our benchmark report and its sister report Generative AI: operators take their first steps – shows that CSPs are expecting a significant impact on their businesses in the short term from implementing GenAI / LLMs.

When do you expect that GenAI / LLMs will have a significant impact on your business?



TM Forum, 2023

In this e-book we look at the typical processes that CSPs are adopting to set up large language models and how they intend to maintain them. For our research we interviewed several progressive CSPs to establish current best practice and their intended strategies for the short term.

Read our reports to find out more:

CHAPTER 1

what is a GenAI large language model?

GenAI works by processing huge volumes of data to find patterns and determine the best possible response to a question or situation, which it then generates as an output. By feeding the AI immense amounts of data it is able to develop an understanding of correlations and patterns within the data and produce its own content.

Whereas conventional approaches to machine learning require data scientists to develop artificial intelligence from scratch, GenAI involves the use of foundational models – deep-learning neural networks – as a starting point to develop models that power new applications more quickly and cost-effectively.

GenAI in the telecoms industry (and more broadly) uses a large language model (LLM) with potentially hundreds of billions of functions or parameters to enable the AI to produce rich and useful results. Parameters, used to build the models and its behavior, are configuration variables learned during a machine-learning process.

We are also seeing the emergence of small language models (SLMs) in the telecoms space. SLMs are still large – possibly several billions of parameters – but are compact enough to run on a limited storage and compute platform like a mobile device. They can be trained on very focused data sets from specific domains within CSPs' operations, so can be used for specific GenAI use cases such as engineering copilots. In this e-book we focus on LLMs as they are the predominant language model already in use.

Foundation models

Foundation models are frameworks which organizations can build on top of to create content and tailored applications. The main application of OpenAI's Generative Pre-trained Transformer models (GPT-3 and GPT-4), for example, is the ChatGPT chatbot, but thousands of businesses across the world are now working on their own applications using these LLMs.

Many software vendors and systems integrators are now coming to market with their own telecoms-specific foundation models to ease the process of building an LLM from scratch. TM Forum's community of members is also very active in establishing blueprints and best practices for AI deployment and thinking about how the Open Digital Architecture can progress towards being AI-native (see chapter 4).

In the following two chapters we look at the typical action plan for setting up and maintaining an LLM within a CSP, drawing on best practice from outside the telecoms industry in enterprise IT, where AI deployments are slightly more mature.

Many software vendors and systems integrators are now coming to market with their own telecoms-specific foundation models to ease the process of building an LLM from scratch

CHAPTER 2

setting up telecoms LLMs

Setting up an LLM involves a series of complex and resource-intensive processes. Following is an overview of those processes outlined in seven steps.

Initial design of the model architecture

This first step, and one of the most important, is for CSPs to decide on the model architecture, including fundamental details of the type of transformer – neural networks that form the basic building blocks of all LLMs – the number of layers the model will need and the estimated number of parameters. They also need to start mapping out potential other "hyperparameters" – parameters which specify details of the AI learning process, such as the learning rate or choice of optimizer. An optimizer is an element of machine learning which defines how to tweak parameters during the learning process to get as close to zero loss function as possible.

This step requires very specialist knowledge of transformer models and how different configurations might interact. We are seeing many CSPs engaging with technology suppliers to help with this as a lack of in-house skills may hamper the CSP's ability to get this key step right.

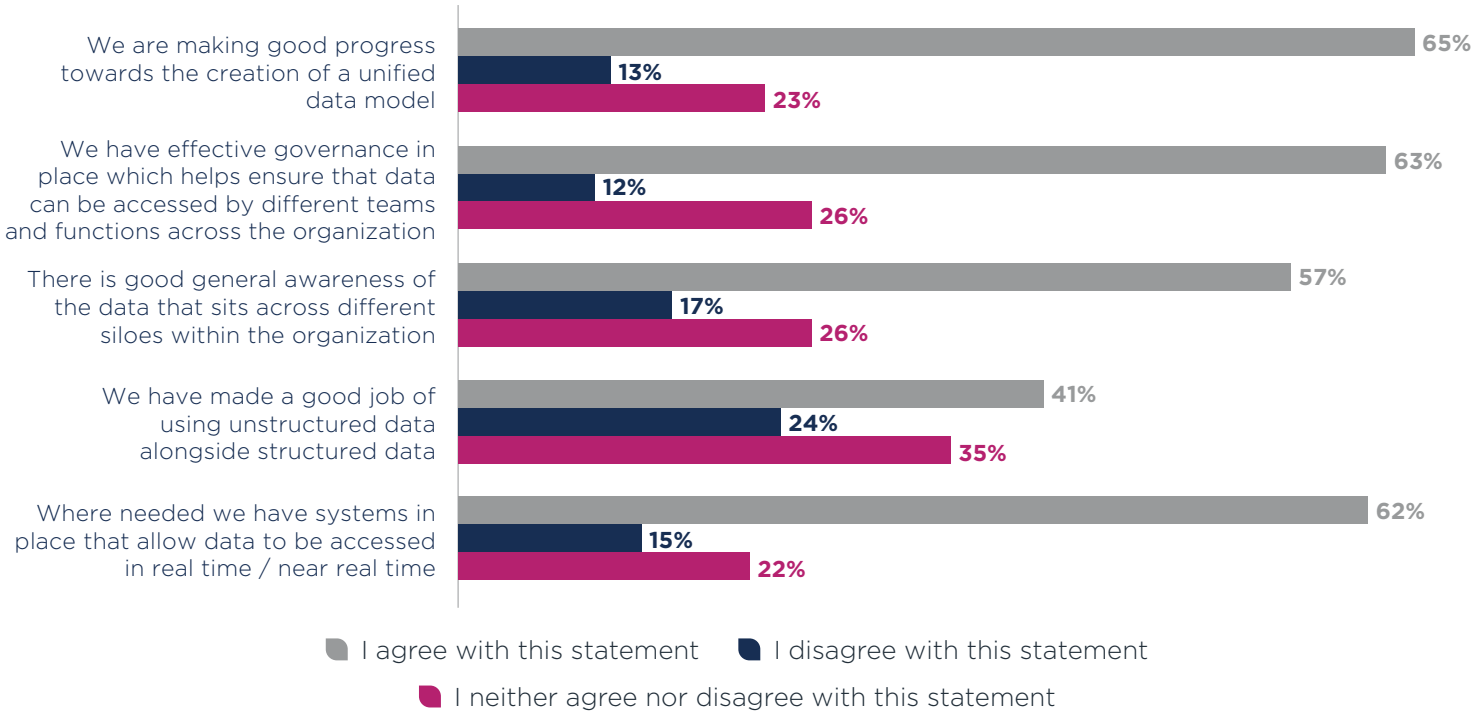
Data collection and formatting

The next step is to think about gathering data sets to train the model. This involves collecting data from a wide variety of sources to ensure the model can understand and generate content across different contexts and styles. This can be a real challenge for CSPs as the format of data used for decades serves the use of relational databases, which is not ideal for LLMs (see chapter 4).

Many CSPs are implementing a data cleanse and preprocessing phase here in order to ready the data for LLM use. This involves stages such as removing duplicate entries, filtering out low-quality text, anonymizing personal information and reformatting manually entered data such as engineering IDs.

TM Forum's Modern Data Architecture Group has been launched by member CSPs to support new AI-enabled business models and build a new standard in data architectures for the industry. Our survey for our earlier reports shows that there are mixed opinions among CSPs as to their progress in making their data available for AI, so there is still some work to do on this, although some 65% of respondents said they are making good progress (see chart).

How successful has your organization been in making data available for AI and machine learning?



TM Forum, 2023

Optimizing tokenization

Before the model can be trained, text data points need to be converted into a format that the LLM can understand, a process known as tokenization. Text is split into tokens – words, subwords or characters – that are then mapped to numerical IDs to process and generate language. Tokenization provides a structured way to break down text into manageable pieces that the model can easily process.

This may seem like a trivial step, but identifying the optimum tokenization strategy in the first instance will significantly improve the model's performance and its ability to understand different languages or specialized vocabulary once in production. Tokenization is all about establishing speed and performance down the line.

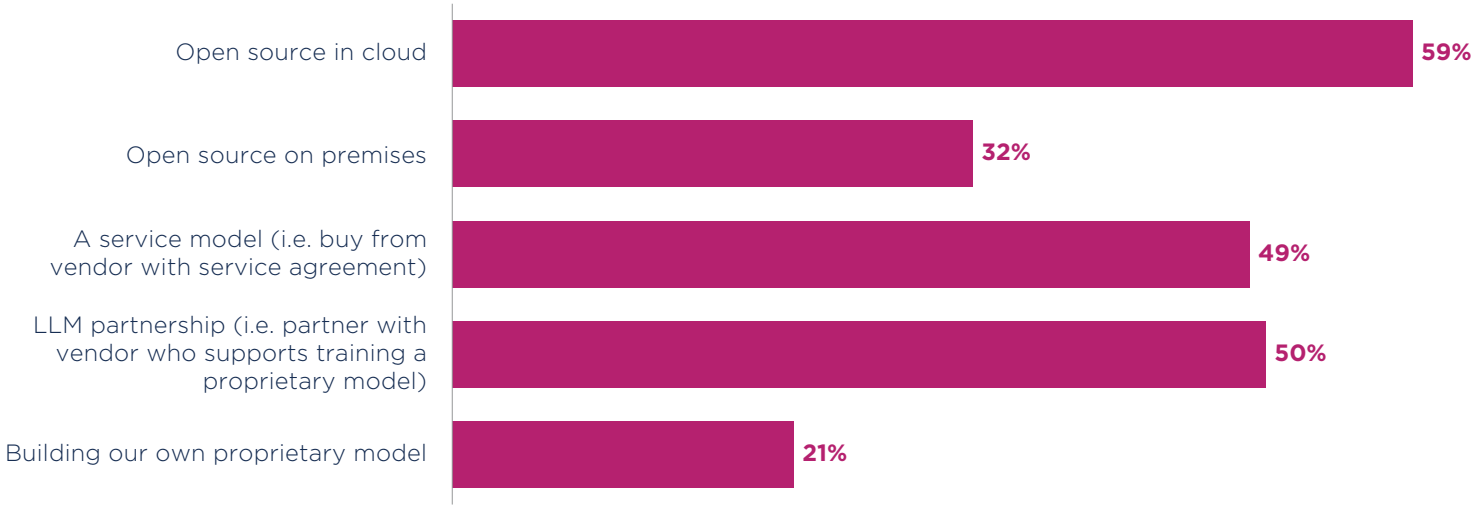
Training the LLM

Once the data has been prepared and tokenized, the actual training of the model can begin. This involves feeding large amounts of data into the model and actively adjusting model parameters based on the output and the expected result.

Training an LLM requires significant computational resources, often involving multiple graphics processing units (GPUs) or tensor processing units (TPUs) – designed specifically for machine learning workloads – for days or weeks. As such, many operators going through this process are realizing the monetary cost of training LLMs in the cloud, where they are billed for compute and storage on a usage basis.

Our survey shows that CSPs are generally not building their own proprietary models and that they are predominantly working in the cloud, so cost assessments in the architecting, tokenization and training phases are critical to the total cost of ownership calculation for telco LLMs (see chart on the next page).

Which large language model (LLM) approaches are you considering for your organization (choose all that apply)?



TM Forum, 2023

Appraisal, fine-tuning and specialization

Following the initial training, an LLM's performance is evaluated to understand the model's strengths and shortcomings, followed by a series of tuning and adjustment processes to improve the results. These might be adjustments to the model's architecture, training methodology or data quality. At this point the model may be trained using data sets specific to certain CSP operations, allowing it to adapt its knowledge to perform better on those tasks.

For example, prompt engineering is the process of refining LLMs with specific prompts and recommended outputs, and of refining input to various GenAI services to generate text or images. Retrieval augmented generation (RAG), meanwhile, provides a way to optimize the output of an LLM with targeted information without modifying the underlying model itself. The targeted information can be specific to a particular organization and industry, so the GenAI system can provide more contextually appropriate answers to prompts.

Deployment and user interaction

After the in-depth process of fine-tuning is complete and the LLM is deemed to be ready, it can be deployed and set up to allow users or other applications to interact with it. This involves setting up application programming interfaces (APIs), specifying compute resources for running the model, and making sure the model runs in the first instance without significant issues.

Post-deployment will likely involve re-training parts of the model in active use or more fine-tuning with new data. This phase focuses on updating the model to address emerging biases or inaccuracies, and scaling the deployment infrastructure as usage grows.

In the next section we look at a typical strategy being adopted by CSPs for maintaining live LLMs.

The post-deployment phase focuses on updating the model to address emerging biases or inaccuracies, and scaling the deployment infrastructure as usage grows

CHAPTER 3

maintenance of LLMs

Maintaining a telecoms operations LLM, post-deployment, involves several crucial considerations to ensure its efficiency and effective return on investment over time. These considerations can be broadly categorized into performance monitoring, content and data management, ethical considerations, security and scalability.

Performance monitoring and maintenance

Continuous monitoring of an LLM's performance is essential to identifying any degradation or issues in accuracy and response times. The first wave of GenAI applications are being introduced in customer-facing roles, and as such it is critical to ensure that the LLM consistently provides accurate and helpful responses. Performance metrics and thresholds should be established, regularly reviewed and used to fine-tune the model by a mixed-discipline team of AI experts, data scientists and subject matter experts from the CSP's own operations teams.

This phase may seem more trivial than the set-up process. But LLMs are living data models, which left to their own devices can deteriorate very quickly into inefficient or unusable models, so ongoing monitoring and maintenance is also an essential exercise.

Retraining with new data

To keep an LLM relevant it must be periodically updated and retrained with new data to reflect the changing technology options in specific CSP operations. For example, GenAI that is used in the network will need to be kept aligned not only with network technology advancements in the core, backhaul and radio networks, but also with new processes, automation strategies and DevOps

To keep an LLM relevant it must be periodically updated and re-trained with new data to reflect the changing technology options in specific CSP operations

methodologies. Establishing an ongoing process to feed this information into the LLM ensures that the model remains effective in understanding and responding to current topics and customer and employee inquiries.

In customer-facing roles, the LLM should be trained on data from the centralized product / service catalog, so that when producing results for a customer it is using the current portfolio of products, services, bundles and offers consistently across all channels.

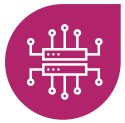
Ethics and bias mitigation

Ethical use of LLMs involves ensuring that the model does not start to produce biases or generate inappropriate or harmful content. This is particularly important in service industries with a large consumer base such as the telecoms industry, where the regulatory penalties for developing bias can be significant and very damaging to a company's reputation.

Continuous efforts to identify and mitigate biases in training data and model outputs are necessary, along with establishing guidelines for responsible AI use. Regulators are now putting guidelines in place to govern this aspect of GenAI, including in the UK and European Union.

Security and privacy

Protecting customer data and ensuring privacy compliance – such as with Europe's General Data Protection Regulation (GDPR) – is an important consideration for CSPs, particularly in settings where LLMs handle sensitive personal information. Regular security assessments, and implementing robust data anonymization and encryption techniques, are typical steps for CSPs building a strategy for customer trust and legal compliance.



Scalability and infrastructure management

As usage grows and the LLM is tasked with handling more complex queries or a larger volume of interactions, the underlying infrastructure must scale accordingly. This means planning for increased compute resource, continued optimization of LLMs for efficient operation, and ensuring that the infrastructure

Protecting customer data and ensuring privacy compliance is an important consideration for CSPs, particularly in settings where LLMs handle sensitive personal information

can cope with peak loads. This last point is especially relevant in telecoms, as the volume of data handled from metrics such as customer interactions or even network events is particularly high, but can also be extremely inconsistent.

Integration with orchestration

LLMs often need to interact with existing customer relationship management (CRM) systems, databases, other operational tools and a gamut of business and operational support systems (BSS/OSS). Ensuring seamless integration and the ability to pull in relevant information or execute tasks based on LLM interactions requires ongoing maintenance and updates to integration points.

The logical path many CSPs are taking as part of their automation strategy and increased dependence on service orchestration is to ensure the LLM is highly interoperable, primarily with the orchestrators rather than all the other legacy systems. In this way the orchestrator can behave as a gateway for enquiry between the operations and business systems and the LLM.

Cost management

Ensuring the operational costs associated with running an LLM do not run out of control is of paramount importance in this nascent phase of the use of GenAI. Costs are centered around computational power, storage / hosting, and maintenance expenses. As Vodafone CTO, Scott Petty, pointed out at a recent TM Forum press event, LLMs can become very expensive very quickly, so CSPs need to make a judgment call on the benefits they are deriving from the LLM versus the ongoing cost.

In summary, the monitoring and maintenance of a CSP LLM requires a multidimensional approach, simultaneously, to many metrics. Each of the areas outlined needs regular review and proactive management to ensure the LLM continues to provide value while aligning with industry standards and expectations. As such, we are seeing CSPs creating new teams and partnerships with technology suppliers specifically designed to get the most from LLMs and to form the basis of best practice in these early days of telco GenAI and LLMs.

In the next chapter we look at the importance of collaboration and standards and how TM Forum is working towards a fully AI-enabled Open Digital Architecture.

Ensuring the operational costs associated with running an LLM do not run out of control is of paramount importance in this nascent phase of the use of GenAI

CHAPTER 4

evolving standards for AI

In the current AI adoption boom it is important that the telecoms industry works together to develop standards for LLMs while the technology is still nascent. A prolonged period of experimentation and siloed development would lead to CSPs accruing technical debt in the same way they have with legacy IT.

TM Forum is currently working with members towards a fully AI-enabled Open Digital Architecture (ODA). In an interview for Inform following the Forum's annual Accelerate event, Andy Tiller, Executive Vice President, Products and Services, outlined some of those plans.

The Forum has been working on AI architecture, frameworks, governance and standards for about five years and has already embedded AI capabilities into ODA such as extensions to the e-TOM Business Process Framework and Open APIs for carrying intent, as well as reference architectures, maturity models and governance best practices. Now planning work is being carried out on how to fully AI-enable ODA.

Tiller outlines two key areas that are being addressed:

- The first is the data architecture. New uses of AI require data to be made available in real time and structured differently from the relational database tables used in traditional operational data stores and data lakes. This includes tapping into huge amounts of data in real time, which is facilitated by data streaming architectures, as well as the use of knowledge graphs.

  The challenge of how to create a data architecture which is fully AI-enabled and make that part of the ODA is a major new project, says Tiller. It includes how to share data across multiple parties within privacy constraints and regulation.

The second key area being addressed is how you implement the reference architectures for intent-based closed loop automation in a standardized way.

"I heard one of our members describe the way AI is being used as a lot of artisanal applications hand crafted and bolted on," says Tiller. "That's great for now because we need a lot of AI innovation and there are lots of very good use cases. But we need to start thinking about how we do it in a standardized way.

"We want to create standardized ways of implementing intent-based automation using ODA Components, Canvas and Open APIs. The aim is to avoid future AI technical debt and the complexity and customization that's so expensive to maintain in legacy systems."

"I heard one of our members describe the way AI is being used as a lot of artisanal applications hand crafted and bolted on. That's great for now, but we need to start thinking about how we do it in a standardized way.

Andy Tiller,
Executive Vice President,
Products and Services,
TM Forum

ADDITIONAL FEATURE BY TCS

GenAI path to production

Enterprises are increasingly focussing on incorporating GenAI technology and exploring the feasibility of large language models (LLMs), since their emergence in late 2022. We are at the cusp of moving from the Proof of Technologies (PoTs) and Proof of Concepts (PoCs) into wide-and-deep production grade development and deployment. As with other technologies we've deployed in the past, it is essential to find important business objectives that can be met with GenAI. Enterprises need to outline a business case and enumerate the Return on Investment (RoI), manage risks, secure required resources, setup the right architectures, governance, and implementation plan. The path to production for GenAI solutions is not easy, and building an AI-mature enterprise is a long journey. While PoC architectures are relatively simple, production grade architectures require significant building blocks to be enabled and secured. Deploying GenAI solutions at scale also requires profound shifts in the roles of people, robust compliance and governance and ways of working, resulting from such deployments.

From TCS' experience of helping our enterprise customers deploy and scale GenAI use cases, we have developed a robust framework for ensuring success of GenAI programs. A value chain analysis is the first step towards preparing the pathway to production.

TATA CONSULTANCY SERVICES

The value chain analysis includes five critical steps:

1. **Identify ecosystem entities and boundaries:** A GenAI use case is seldom standalone and often requires the content and integration with core functional and application ecosystems. Boundary conditions are the limits or specifications that are applied to GenAI systems, such as LLMs, that can affect various aspects of the outputs of GenAI. Creating boundary conditions in GenAI is not a one-size-fits-all solution, but rather a very domain context dependent and iterative process, that involves a clear architecture strategy along with various tools and techniques.

2. **Frame interactions and value exchange:** It is important to design the model interactions or other systemic workflows and business transactions as well as the larger ecosystem which may include partners, customers, suppliers and more, with the most important aspect being the **human in the loop**. This also includes assessing the skills and role shifts that are required to adopt GenAI solutions at scale.

   Also, enterprises need to move beyond existing metrics to measure the success of AI implementations; this calls for the development of right performance indicators to measure the impact of the Gen AI technologies on their business.

3. **Identify hotspots in system and human area:** Hotspots are the contextual use case placement scenarios in a business value chain that will yield the maximum benefit from use of Gen AI technology. This will also provide an early view of the degree of impact, including the need to reimagine processes to be **AI native**, considering *Human vs AI* interplay at the core. AI native involves embedding reliable AI capabilities seamlessly within a system, processes, and all stages of the value chain, encompassing design, deployment, operation, and maintenance.

   This approach utilizes a data-driven and knowledge-based ecosystem creating AI functionalities and enhancing existing processes or systems with adaptive AI as it applies.

4. **Evaluate hotspots:** Evaluation of hotspots includes identifying the business impact to drive tangible business outcomes, KPI impacts and a clear ROI path post implementation. Technical complexity in terms of the ease of customizing, implementing, and integrating LLMs is also an important facet of evaluation. Critical questions like the depth of LLM customisation, prompt engineering, Retrieval Augmented Generation (RAG), fine-tuning and use of multi-modal techniques to increase contextual responses and limit undesired behaviours such as hallucinations and sandbagging, ease of implementing and integrating with current systems, data readiness for LLM training and so on are to be addressed. Relevance of LLMs in terms of the value and core capability to the use cases, support of LLMs to complement the allied areas, value (in terms of cost and complexity) than traditional AI methods are also factors to be decided.

5. **Risk assessment and classification based on AI infusion levels:** Analyzing all possibilities of the solution going wrong and assessment of organizational risk, use-case level risk, compliance and regulatory risk are very important.

Production scaling requires an **AI-First Enterprise Architecture** to be well defined. Strategy, architecture and design should encompass data or content readiness and integration, model management and scaling, policies and guardrails, design patterns for integration of LLMs, testing, operations and tools for observability, Finops ,etc. along with strategic governance , operational governance and change management. Setting up an **AI first North Star Architecture** will aid as a foundation for AI governance.

The following are some of the key aspects to address:

1. ***AI Experience Fabric***: Designing the user experience layer for AI including AI Playgrounds, Application UI, Chatbot, Multi-modal experience layer infusing intelligent user interfaces merged with AI for flexibility, usability and relevance for human -machine interaction.

2. ***Data and Integration requirements***: Data and Integration requirements vary across the various phases including build, test, deployment, and operations. During the build or development phase, data quality, data pipeline, vector DB designs, curation of transactional data, data lake, documents ,public data, semantic cache and UI/APIs form crucial foundations. Content and data readiness for development and prompt engineering are foundations of GenAI. During testing, data requirements are critical. Synthetic data, Q&A and customer experience testing are also required. Data services and CI/CD become features for integration during deployment and data operations and API monitoring are important aspects in the “run” or operations phase.

3. ***Model Management***: Model management includes hosting, versioning, tracking, sharing of models and prompt management. During the build phase, design and implementation of RAG, fine tuning (as required by the use case demands), multi-agent architectures and so on play a big role. Automated scoring of models during testing and model configuration are important in deployment phases. Once operationalized, model recalibration and optimization are continuous improvement processes.

4. ***Guardrails / Security***: During the build phase, grounding APIs, content filters, application security integration are key facets. During testing and assurance, red teaming, strong bias and hallucination testing, chunking strategies are required. Zero Trust Access + PII Scanners, Integration with IAM (Identity and Access Management) are foundational capabilities. Further, tracking, validation, evaluation, human feedback and continuous monitoring of safety, bias and other parameters are required. To prevent Indemnity & IP ownership conflicts, its utmost important to establish Usage Policies, Human in the Loop and Legal Frameworks.

5. ***Model Ops, Observability and Compliance***: This is a very important facet in production, where model monitoring, response monitoring and financial monitoring are the core steps. This requires various architectural enablement including but filters, compliance rules, behaviour control, observability tools and alerts, prompt and response logging and metrics management.

6. ***Non-Functional Capabilities***: Compute, storage, and network switching capabilities for the desired performance levels, reliability and other NFRs are core to production scaling and management. Capacity planning and quantization, NFR hook configurations, NFR testing, monitoring, reporting, and self-heal capabilities become important factors to be considered.

7. ***Cost and FinOps***: Is very core to scaling and implementation of GenAI. Model usage costs, training and inference costs, computation and consumption costs are all critical to monitor at granular levels. A clear governance strategy is key to ensuring optimal cost consumption. To have a clear strategy for cost containment and optimization, a hybrid strategy with proper assessment of Opensource + Frontier Models, Hybrid and Federated Strategy for Models, Platforms, Infrastructure, and cloud is critical.

Finally, ensure below factors and considerations are in place:

All GenAI use cases must be designed and instrumented with Responsible AI principles at the core. Accountability, transparency, fairness, reliability and safety, security and privacy and inclusiveness are key facets of Responsible AI.

Establishing governance both at a strategic level and operational level is important. The governance function must also ensure that the reliability and safety measures are defined to protect from failures through well designed processes for remediations management, continuous monitoring, feedback, and evaluation.

Regulatory, privacy, security and inclusiveness standards and compliance are vital for every enterprise and are equally applicable, if not more relevant when Gen AI technologies are deployed at scale for production.

Accountability ensures the right impact assessment, oversight of significant adverse impacts, fit for purpose AI, data governance and management and human oversight and control.

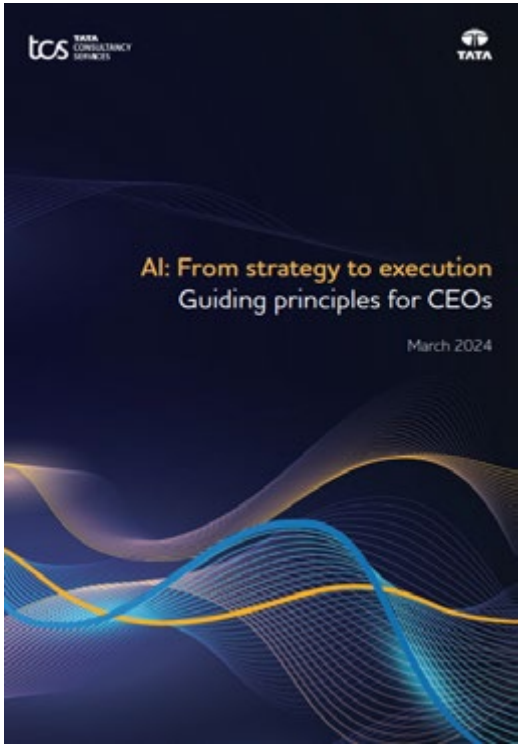
Transparency ensures system intelligibility for decision making communication to stakeholders and proper disclosure of AI interaction.

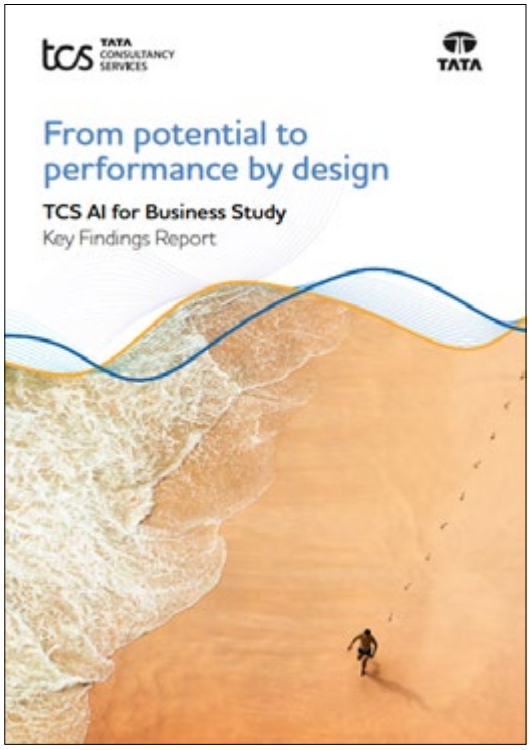
Fairness is about addressing quality of service, allocation of resources and opportunities, minimization of stereotyping, demeaning, and erasing outputs.

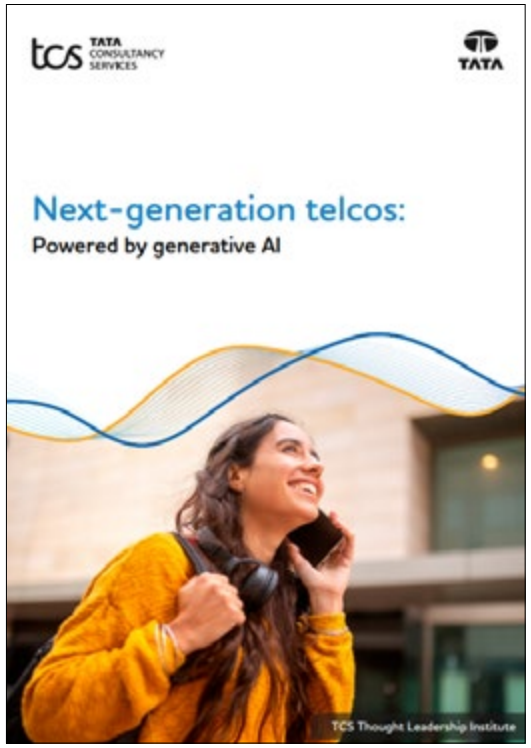
It is important to build with production and not POC in mind. Keeping control on costs (TCO vs. quantified benefits) from inception is important. It is vital to set up a proper FinOps model, understand costs and track them in the early stages.

Building repeatability into the process through LLMOps and reference architecture will enable both speed and certainty of implementation as well as provide end- to-end observability.

Read additional TCS reports to find out more:







Meet the Research & Media team

Author:
Dean Ramsay
Practice Lead
dramsay@tmforum.org

Chief Analyst:
Mark Newman
mnewman@tmforum.org

Senior Analyst:
Richard Webb
rwebb@tmforum.org

Head of Operations:
Ali Groves
agroves@tmforum.org

Commercial Manager:
Tim Edwards
tedwards@tmforum.org

Editor:
Ian Kemp
Managing Editor
ikemp@tmforum.org

Editor in Chief, Inform:
Joanne Taaffe
jtaaffe@tmforum.org

Global Account Director:
Carine Vandevelde
cvandevelde@tmforum.org

Digital Media Coordinator:
Maureen Adong
madong@tmforum.org

Marketing Manager:
Ritika Bhateja
rbhateja@tmforum.org

Published by:

TM Forum

US office
181 New Road
Suite 304
Parsippany, NJ 07054
USA
Phone: +1 862-227-1648

European office
TM Forum
Uncommon
34-37 Liverpool Street
London EC2M 7PP
UK
Phone: +44 207 748 6615

www.tmforum.org

ISBN: 978-1-955998-79-6

tmforum

To find out more about TM Forum's work on GenAI large language models please contact Andy Tiller