



# Telecom major successfully builds a one-stop LLM gateway

Optimizing Model Development Lifecycle of LLM use cases using an intuitive low code portal for citizen data scientists.



The client is a leading communications company known for its innovative work in Telecommunications, Information and technology, and consumer electronics. It is a key player in emerging technologies such as 5G networks and the Internet of Things (IoT) reshaping the future of connectivity and digital transformation.

Today, LLM technology offers substantial business benefits by automating tasks like language generation, data processing, and model inference, thereby enhancing efficiency and decision-making. However, its adoption is often hindered by technical complexity and high costs, posing challenges for integration into existing systems.

To gain a strategic competitive advantage using AI, the client recognized the necessity of building a centralized portal that would facilitate the seamless onboarding of LLM-based use cases such as Text Summarization, Report Summarization, Text Classification, etc., with minimal or no coding requirements. This platform aimed to cater to

diverse user personas, including project/business managers/owners, citizen data scientists, and data scientists.

The portal must function as a comprehensive solution for all LLM activities, including Completions, Prompt Engineering, Data Processing (such as text cleanup, deduplication, and PII masking), Finetuning, Evaluation, and Model Inferencing for both closed and open-source models.

Following intensive evaluation and intense competition through the RFP process in this new client environment, Brillio was selected for the LLM Gateway project for our thought leadership, expertise, and innovation in Generative AI. The client was impressed by our ability to showcase working demos and accelerators that are ahead of the curve in this emerging technology. Since then, we have been engaged in multiple conversations, solidifying our role and deepening our collaboration.

**An intuitive platform was crucial to simplify operations, reduce training overhead, and ensure seamless implementation, making LLM accessible and productive for organization.**


# Revolutionizing LLM Onboarding with Brillio's Comprehensive and User-Friendly Portal

To onboard any Generative AI use case, a structured approach must be followed. First, infrastructure provisioning and management are essential, involving the setup of API resources for closed-source models like GPT-3.5 and GPU computing resources for open-source models. It also includes configuring access for team members. Next, deep learning script development is crucial. It evaluates base and fine-tuned models using various techniques and runs inferences on these models, implementing deep learning optimizations for efficient model training. Following this, integrated testing with the full use case architecture ensures the system works as intended. Continuous Integration and

Continuous Deployment (CI/CD) practices are then applied to streamline the deployment of models and infrastructure. Finally, ongoing monitoring and management are required to maintain performance and address any issues that arise.

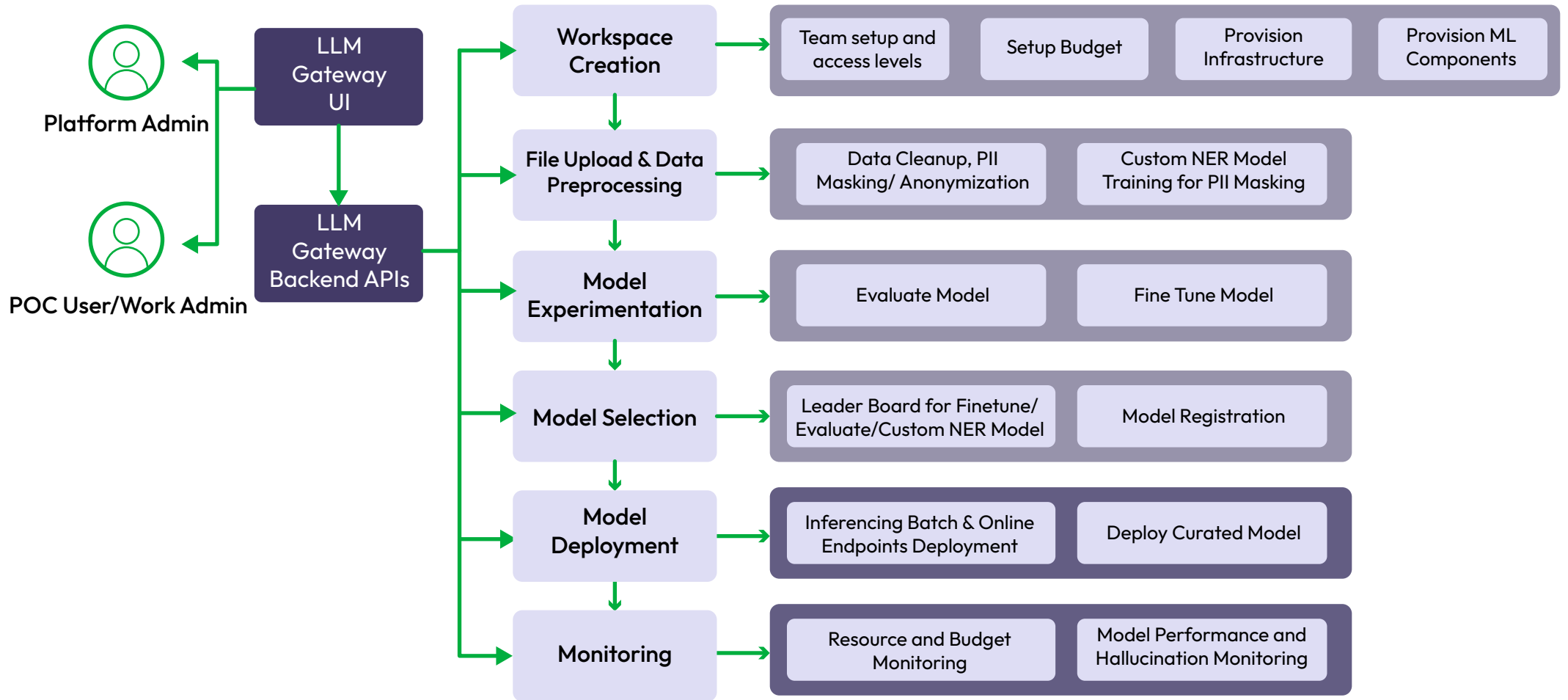
For onboarding every new use case, all the outlined steps must be repeated each time. This can become highly complex and costly when multiple teams are involved. Brillio's solution would automate the entire workflow, allowing teams to build and experiment with models directly from the UI, significantly reducing the time and cost associated with the process and setup.

Brillio's solution driven by our vision to deliver value at scale promised to substantially diminish the time and costs associated with the process and setup.



The new process flow with the solution in place simplified LLM use case onboarding significantly. Brillio's LLM onboarding portal features a **RAG system** for direct data interpretation, a **Prompt Playground** for testing prompts across various LLMs, and a **Model Leaderboard** for identifying the best performing models. It simplifies the inferencing process and provides a unified system for budget tracking. This comprehensive toolset makes the portal an efficient, user-friendly, and cost-effective solution for leveraging LLMs. First, workspace provisioning involves requesting an LLM Gateway workspace and budget, followed by the admin setting up the infrastructure and assigning budgets, teams, and access levels. Data onboarding is streamlined through data science-backed methods, allowing users to upload data, trigger processing jobs, and configure cleanups like PII removal and deduplication. Model experimentation and selection using deep learning scripts, enables users to configure and trigger workflows for finetuning models, evaluate their performance, and compare results to select and register the best model. Model inferencing is similarly simplified, with deployment options for batch or real-time endpoints, and comprehensive endpoint monitoring and management. Finally, the solution includes features for cost tracking and team management to ensure efficient resource allocation and project oversight.

# LLM Use Case Onboarding Flow





# Streamlined LLM Onboarding and Experimentation for All Users

The solution uses a React-based web app as UI which is integrated with the Azure ML backend using .Net and Python APIs. The Azure ML layer hosts all the templated data processing, evaluation, and finetuning workflows. All these layers and artifacts are developed and provisioned with the help Azure Repos and Pipelines. LLM use cases can be onboarded from the portal within a few hours, given the necessary finance and budgeting approvals.

This process is facilitated by automated infrastructure provisioning and tracking, along with comprehensive cost and budget management. Users of any technical proficiency can experiment with LLM functionalities,

including finetuning activities, with minimal to no coding knowledge required. Most activities are templated and can be directly triggered from the UI by configuring parameters or using default settings.

Additionally, the solution simplifies the finetuning and evaluation of base models, helping users easily identify the best model candidate for their use case based on quantitative performance benchmarks. The platform can be further extended to utilize the client's on-prem systems in the future, which can support multiple new use cases. It can also act as an internal marketplace for sharing models, datasets, and features within the organization.

Increased user adoption through an intuitive Low code/No code LLM onboarding portal for rapid use case onboarding.



## ABOUT BRILLIO

Brillio is one of the fastest growing digital technology service providers and the partner of choice for many Fortune 1000 companies seeking to turn disruption into a competitive advantage through innovative digital adoption. We help clients harness the transformative potential of the four superpowers of technology: cloud computing, Internet of Things (IoT), artificial intelligence (AI) and mobility. Born digital in 2014, we apply our expertise in customer experience solutions, data analytics and AI, digital infrastructure and security, and platform and product engineering to help clients quickly innovate for growth, create digital products, build service platforms, and drive smarter, data-driven performance. With 17 locations across the U.S., the UK, Romania, Canada, Mexico, and India, our growing global workforce of nearly 6,000 Brillians blends the latest technology and design thinking with digital fluency to solve complex business problems and drive competitive differentiation for our clients. Brillio has been certified by Great Place to Work since 2021.



<https://www.brillio.com/>

Contact Us: [info@brillio.com](mailto:info@brillio.com)