


Key LLM Considerations: Why Hybrid Multi-Cloud is a Natural Fit for Enterprise-grade LLMs

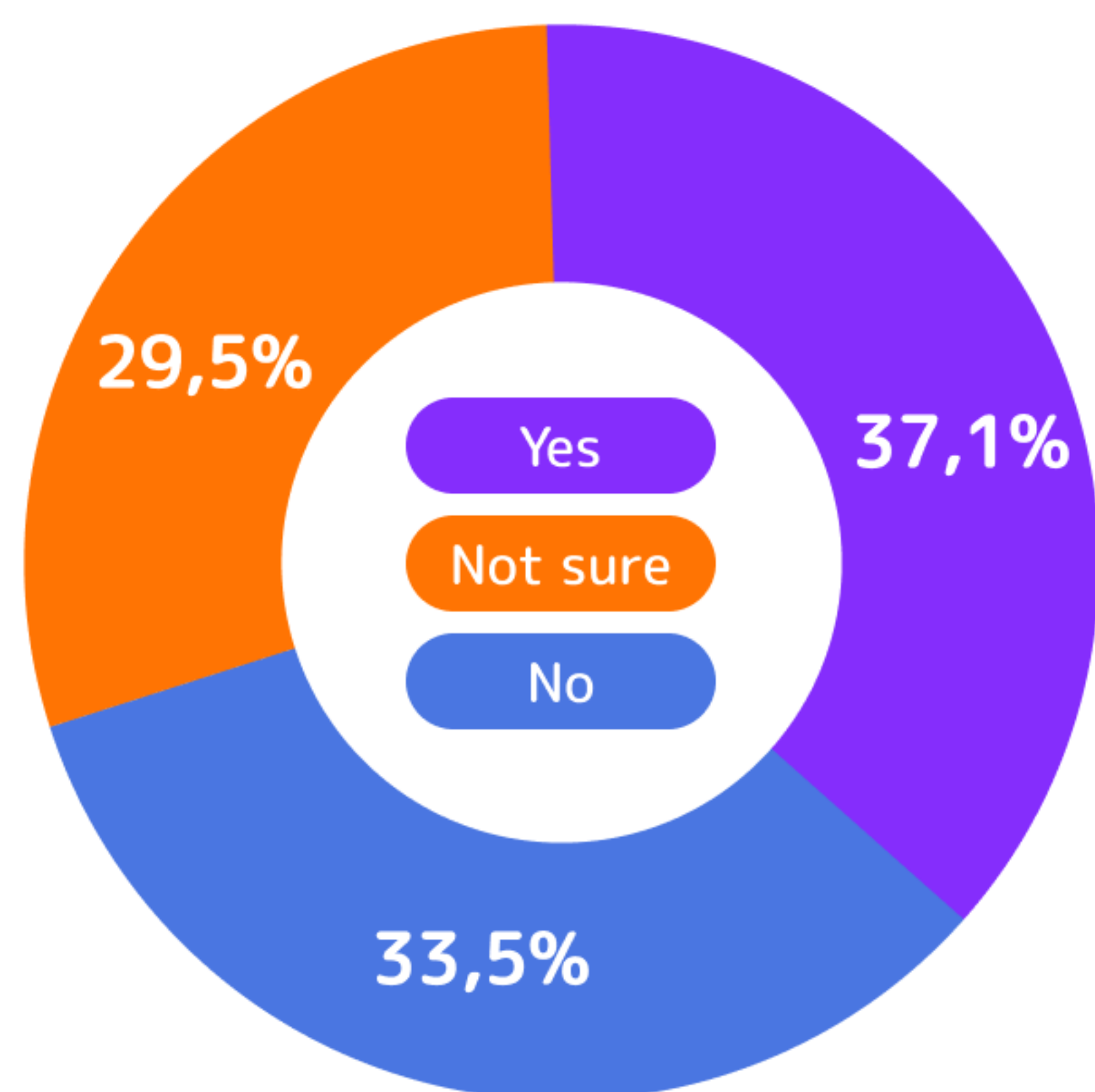
Contents

3	Introduction
4	What are LLMs?
5	LLM Adoption Within the Enterprise: Key Considerations
10	Enabling LLMs in Business: Why Hybrid Multi-Cloud is the Way Forward
12	Streamlining Hybrid Multi-Cloud Operations with the emma Platform



Introduction

2023 has unveiled a new market reality where Large Language Models (LLMs) are not only accessible but also essential for staying competitive. Generative AI and LLMs were already popular in tech circles, but OpenAI's GPT series became a pioneer, bringing the transformative powers of LLMs within the reach of all. Today, around [40%](#) of enterprises are in the midst of their LLM adoption journey.



In-House Adoption of Generative AI and LLMs

*Is your company considering building its own business specific Language Model, or adapting/tuning an existing one?

Let's also not forget the shadow AI — many employees are probably already using consumer-facing LLM applications under the hood. Just recently, employees at Samsung divulged proprietary code and data to ChatGPT in 3 separate incidents, prompting [Samsung](#) to place a ban on the use of LLMs in the workspace. However, outright bans are futile because organizations risk losing out on unprecedented convenience, speed, and productivity gains.

Earlier this year, tech leaders' [call for a moratorium](#) on large-scale AI developments miserably failed to gain traction. That is because organizations can no longer afford to remain on the sidelines when it comes to incorporating LLMs into their corporate environments and core business operations. So, the corporate focus has invariably shifted from resisting to enabling highly regulated, task-specific, and efficient use of enterprise-grade LLM applications. However, harnessing LLMs securely, effectively, and efficiently within the enterprise necessitates a comprehensive reevaluation of organizations' IT infrastructure and cloud strategies.

What are LLMs?

LLMs are a subset of generative AI specifically designed to generate text-based content. Basically, they are computer programs that utilize deep learning techniques and vast amounts of data to comprehend, analyze, and generate new content. Modern LLMs are highly sophisticated in that they can grasp the intricacies of how human beings communicate verbally and in writing and mimic the style to generate human-like text with remarkable accuracy.

Their versatility makes them suitable for a variety of business use cases involving human language tasks and services.



Human-Computer Interaction

- Chatbots for real-time customer support
- Customer care applications



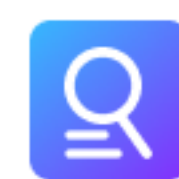
Language Generation

- Content summarization
- Content generation for marketing, programming, and creative endeavors



Information Extraction

- Knowledge mining to extract relevant information from data troves
- Content classification, metadata creation, and categorization
- Entity extraction to identify entities and the relationship between them



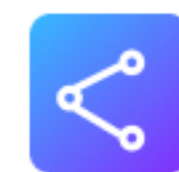
Search and Recommendations

- Enterprise search engines
- Semantic search for context-aware results
- AI-driven personal assistant
- Content recommendation engines



Content Generation and Enhancement

- Text generation for articles, reports, and stories
- Multilingual translation services
- Content personalization



Data Analysis and Insights

- Sentiment analysis based on customer interactions and feedback
- Data summarization for condensing large datasets
- Natural language querying for database retrievals
- Market research and trend analysis



Workflow automation

- Email automation
- Report generation
- Sales and marketing automation
- HR and recruitment automation



Compliance and Risk Management

- Regulatory compliance checks
- Risk assessment and analysis
- Compliance reporting.

LLM Adoption Within the Enterprise: Key Considerations

As businesses explore the vast potential of generative AI and LLMs across a growing spectrum of applications, they face an overwhelming abundance of choices and challenges. Crafting a robust LLM strategy, selecting suitable vendors, and reassessing their infrastructure decisions have become critical considerations. Challenges such as high computational needs, cost implications, and the intricacies of governance and compliance add further complexity to the LLM landscape.

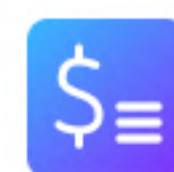
Cost Implications

The cost to train and run an LLM is extraordinarily high. An LLM like GPT-3 could cost north of [\\$4 million](#) to train and an additional [36 cents](#) per query answered. That's not even all. Below is a high-level cost breakdown of an LLM.



Computational Costs

LLMs run complex algorithms for lengthy time periods, necessitating high upfront investment in GPUs or specialized AI accelerators as well as ongoing operational costs. The cost of scaling and optimizing infrastructure also adds up as LLM usage increases.



Data Collection and Data Preprocessing Costs

LLMs must be trained on high-quality datasets to generate meaningful and accurate responses. Organizations need to acquire or purchase these datasets and further clean, annotate, and tokenize them, all of which are complex and resource-intensive processes.



Skills Acquisition Costs

Developing and training LLMs demands expertise in data science, natural language processing (NLP), and machine learning (ML). Hiring, training, and retaining professionals with these in-demand and scarce skills adds to the overall expense.

With cloud, businesses can tap into virtually unlimited supply of compute power along with a range of pre-trained AI models and LLM-based services at a fraction of the cost of on-premise training and deployment.

However, pricing can vary significantly across different clouds, even for specific LLM tasks. For instance, one provider may offer lower pricing for training LLMs, while another may be more cost-effective for inference workloads. Legacy and single cloud organizations risk missing out on the cloud's cost optimizations for LLM workloads.

Fine-Tuning Strategies

Commercially available, pre-trained LLMs often do not meet specific business needs by default. That's because business applications require knowledge of proprietary business data as well as industry-specific terminologies and concepts. Developing a language model from scratch provides the highest level of confidentiality and customizability. However, this approach is feasible and necessary for only a small percentage of large enterprises and government agencies, such as the CIA. For most businesses, a more practical approach is to perform some form of fine-tuning to adapt open-source models to their business needs.



Prompt Engineering

Users refine their input prompts to pre-trained LLMs, providing specific data and context, to elicit contextually relevant and accurate outputs.



Full Fine-Tuning

Involves further training a pre-trained LLM on domain-specific or task-specific datasets, so it can generate contextually relevant and accurate responses for particular business or industry needs.



Parameter-Efficient Fine-tuning (PEFT)

Focuses on limited parameter adjustments for tasks that do not require extensive modification. PEFT offers resource efficiency at the cost of reduced customization.



Retrieval Augmented Generation (RAG)

Involves retrieving relevant information or context from a predefined knowledge base or database to generate accurate, up-to-date, and context-rich outputs.

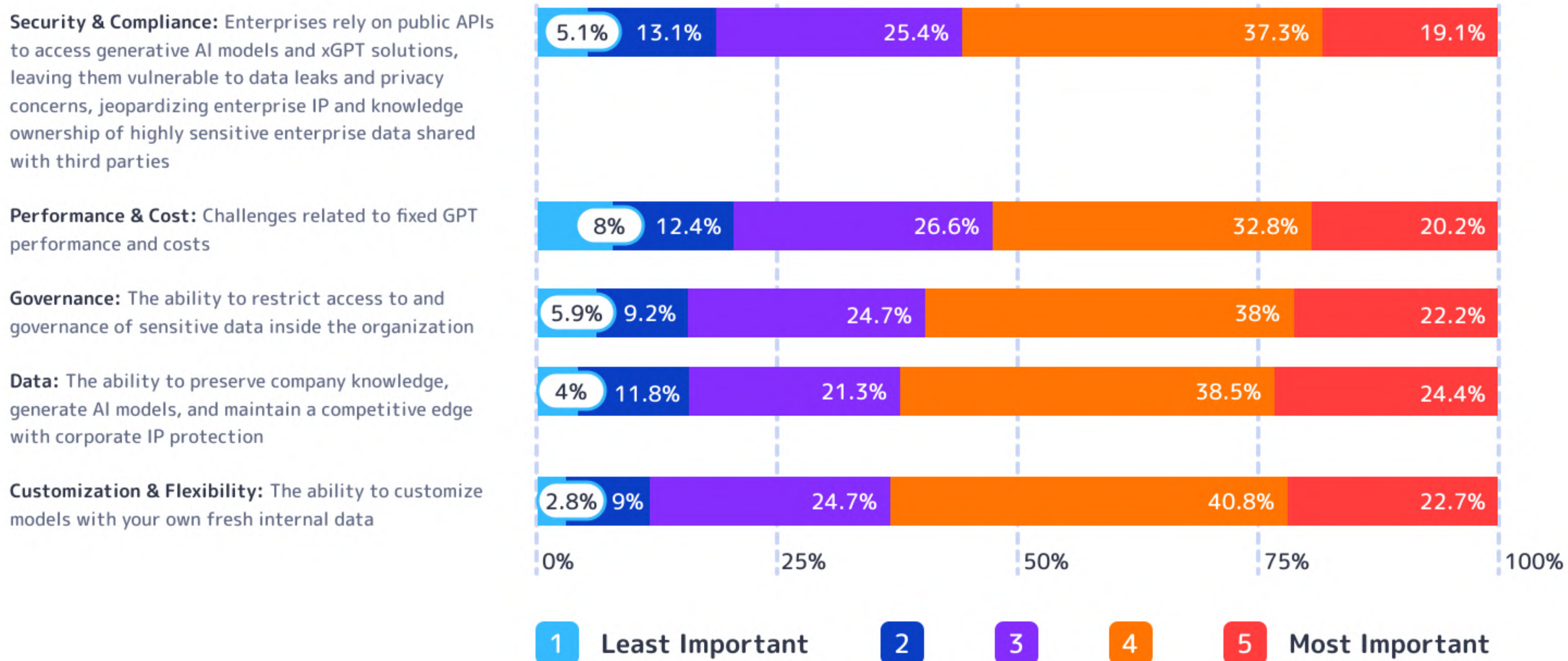
Fine-tuning also requires a deep understanding of Natural Language Processing (NLP) and Machine Learning (ML), as well as access to significant compute resources. Major Cloud Service Providers (CSP) offer dedicated platforms for fine-tuning models on their cloud infrastructure.

However, the degree of support, flexibility, and control for customizing AI models may vary across platforms. For instance, AWS SageMaker offers extensive flexibility and features, while Azure ML provides a visual interface for a code-free experience. Different teams within the same organization may prefer different or even a combination of fine-tuning strategies and platforms, creating a dilemma for organizations dedicated to a single CSP.

Security and Privacy Implications

Businesses need to feed LLM applications with valuable proprietary data to drive context-aware and accurate responses. For **over 50%** of organizations, preserving the ownership, confidentiality, and control of this highly sensitive enterprise data is a significant challenge.

Key challenges/blockers in adopting AI/LLMs/xGPT



Sensitive Data Exposure via Prompts

The corporate data fed to an LLM for prompt engineering can become a part of its training datasets, potentially becoming publicly accessible.



Extended Data Storage

Commercial and open-source LLM providers may have extended data storage policies, which can put corporate data at risk of a data breach.



Zero-Day Vulnerabilities

Hasty LLM deployments are prone to zero-day vulnerabilities that cybercriminals can leverage to breach the corporate security perimeter.



Regulatory Compliance

Data privacy regulations, like the GDPR and HIPAA, can restrict businesses from storing sensitive customer data on public cloud infrastructure or outside the customers' jurisdiction.

Organizations cannot risk sharing their sensitive, proprietary data with public LLMs or other third parties. Public cloud infrastructure can raise data privacy and governance concerns when using LLMs in highly regulated corporate environments. That is why organizations often need an on-premise or private cloud element in their cloud environments to meet their privacy and data residency needs.

Vendor Selection

Organizations have an array of options to choose from when it comes to LLM adoption. The market is brimming with commercial, ready-for-use and open-source, adaptable models. Organizations can pick versatile models, like ChatGPT and BERT, for their diverse needs or tailored solutions, like Jurassic-2 or Med-PaLM 2, for specific use cases and industries.

Below are key considerations for LLM selection



Model Capabilities

LLM capabilities, such as the model's performance, proficiency in natural language understanding and generation, its scalability, multilingual support, and the ability to perform complex tasks, must align well with business requirements.



Fine-Tuning Support

LLMs differ in the extent of flexibility and customization they allow. For instance, the immensely popular and powerful ChatGPT-4 does not support fine-tuning. Those who need heavy customizations prefer models like Cohere, even though they are less popular.



Compliance and Regulations

Companies in regulated industries can only choose from LLMs that are inherently compliant with applicable laws and regulations.



Integrations and Compatibility

LLMs must be compatible with the organization's business applications as well as the underlying infrastructure.

Cost Structure: LLMs can incur complex billings based on varied factors such as the number of queries and token size.

Currently, all major CSPs are vigorously competing for LLM dominance. They have adopted different approaches to navigate and maximize on the demand surge for LLMs. For instance, Microsoft is banking on its exclusive access to the market-leading OpenAI models, while AWS has adopted a multi-LLM approach via Amazon Bedrock, which allows businesses to leverage and combine capabilities from a diverse set of LLMs, including Anthropic's Claude 2, Cohere's models, and AI21 Labs' Jurassic-2.

Businesses can face a significant obstacle in their LLM adoption journey if their preferred LLM does not integrate well with their cloud provider. Single cloud organizations can find themselves in an LLM lock-out if their cloud providers favor particular LLMs.

Infrastructure Considerations

While the quality of data decides the accuracy of an LLM, its performance and efficiency depend on the underlying infrastructure. Organizations can choose to train or fine-tune and deploy LLMs within their on-premise datacenters, on private or public cloud infrastructure, or any hybrid combination of these. Below are some major infrastructure considerations for organizations' LLM adoption:



Scalability

Scaling IT infrastructure as LLMs grow in scale and complexity is hard and requires proactive planning.



Reliability

Downtime, even on cloud infrastructure, is inevitable to some extent. It can be costly for sensitive and mission-critical AI applications and must be accounted for.



Efficiency

Latency-sensitive LLM applications, such as medical-grade LLMs for timely diagnostic analysis, need to leverage geographical proximity to reduce latency, which can be challenging if the on-premise datacenter or the CSP does not operate in a specific region.



Control and Customizations

LLMs must be compatible with the organization's business applications as well as the underlying infrastructure.

Cost Structure: LLMs can incur complex billings based on varied factors such as the number of queries and token size.



Privacy and Compliance

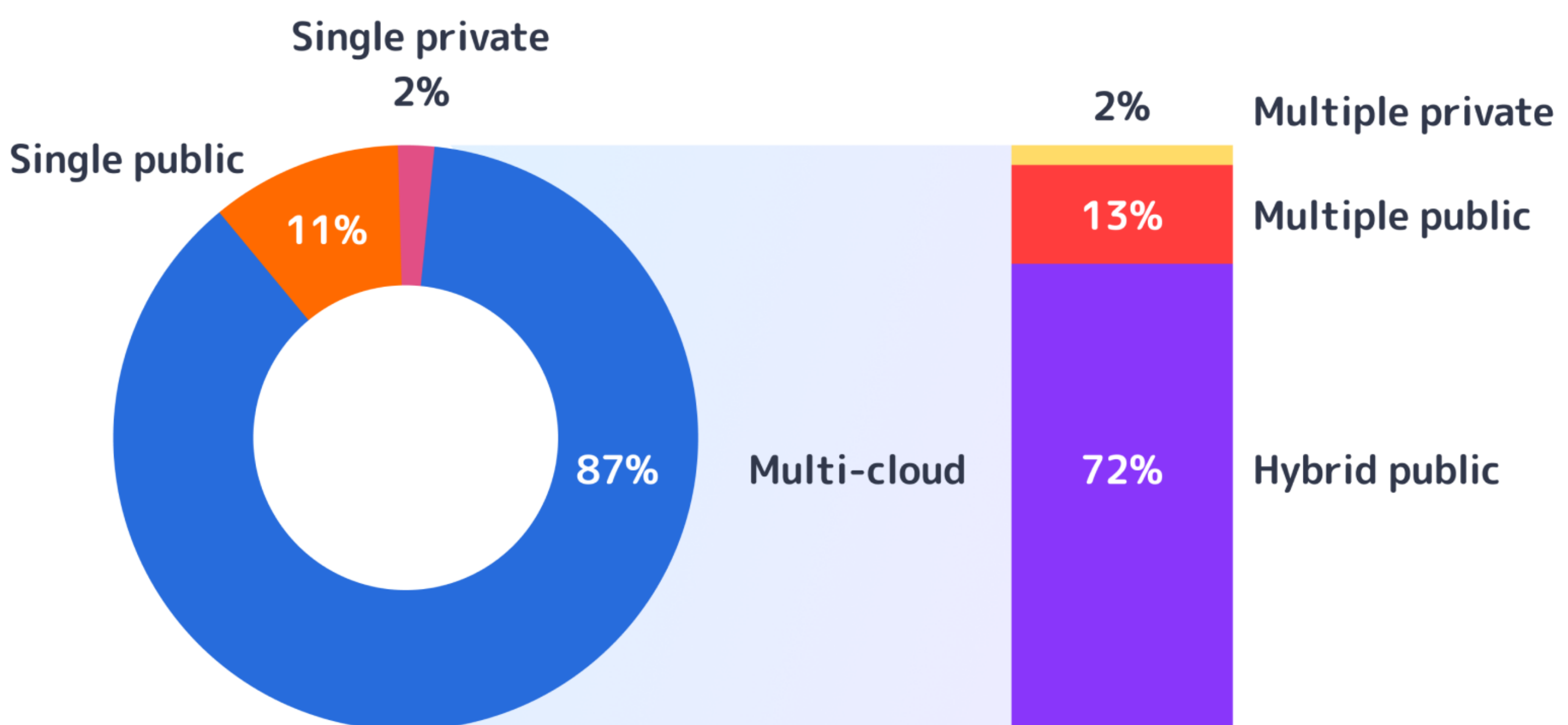
Regulatory compliance requirements or the sensitive nature of corporate data may mandate businesses to keep certain data and insights on private infrastructure.

Both on-premise and cloud deployments have their own strengths and challenges. Public cloud can provide scalability, efficiency, and cost benefits, while on-premise and private cloud provide better privacy, control, and customizations. As such, the infrastructure requirements may differ per use case, so most organizations will need a combination of private and public environments.

Enabling LLMs in Business: Why Hybrid Multi-Cloud is the Way Forward

The unique requirements of LLM data storage and workloads combined with the business dilemma of giving access to sensitive and valuable proprietary data require businesses to rethink their existing infrastructure environments. A hybrid cloud infrastructure is critical to any organization that must provide data access to third-party tools and platforms. In addition, multi-cloud is also a given for organizations with diverse infrastructure needs.

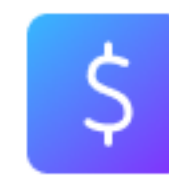
At present, [72%](#) of organizations have a hybrid cloud set-up, while 69% rely on multiple public clouds (multi-cloud). A hybrid multi-cloud is essentially a combination of both hybrid and multi-cloud environments. These architectures operate a combination of on-premises and multiple public cloud environments, effectively being both hybrid and multi-cloud simultaneously. With multiple deployment options, organizations can pick and choose the best suited environment for different LLM needs and workloads.





Data Privacy and Compliance

Regulatory requirements like the GDPR mandate businesses to store users' data locally. With a hybrid multi-cloud, organizations can keep their private and sensitive data within a secure and private environment while still leveraging public cloud infrastructure and tooling to host their LLM workloads closest to where the users are.



Cost Optimization

Private infrastructure requires upfront capital investments while public cloud incurs ongoing expenditure. Despite smaller initial costs, the expense of long-running, resource-intensive LLMs can become exponentially high over time. Large-scale enterprises that have already invested heavily in private cloud deployments can deploy long-running LLMs privately while bursting workloads to the public cloud during demand hikes.

Similarly, smaller organizations with limited funds and in-house infrastructure can get instant access to the cloud's power and scalability at optimal cost. Adding a "multi" factor to hybrid cloud allows all organizations, big or small, to leverage the lowest price provider for each LLM task to optimize their overall LLM spend.



High Availability and Fault Tolerance

Running LLM projects across hybrid multi-cloud provides additional reliability — critical workloads can be configured to automatically switch to another infrastructure environment if one of the constituent environments faces downtime. To ensure business continuity, organizations can distribute workloads across different regions as well as across different providers within the same region.



Performance Benefits

Organizations can leverage the proximity of on-premise infrastructure and analyze queries locally for latency-sensitive LLM applications, like those needed for critical decision-making. The data is processed closest to the edge and transmitted over intranet connectivity, ensuring low latency and optimal bandwidth utilization.

Additionally, organizations can distribute customer-facing LLM applications across various cloud regions, including those that their primary cloud provider does not cover, allowing better application performance and speed for end-users.



Best-of-Breed LLMs and LLM Platforms

Regulatory requirements like the GDPR mandate businesses to store users' data locally. With a hybrid multi-cloud, organizations can keep their private and sensitive data within a secure and private environment while still leveraging public cloud infrastructure and tooling to host their LLM workloads closest to where the users are.



Platform Independence

Public cloud providers face immense computational and competitive burdens due to the massive LLM adoption and experimentation. As the LLM landscape evolves, fundamental changes in CSP offerings, support, and pricing may occur. Organizations with a hybrid, multi-cloud strategy can respond and adapt to such changes better and don't have to make compromises due to vendor lock-in.

Streamlining Hybrid Multi-Cloud Operations with the emma Platform

Despite being a perfect fit for most organizations, especially for LLM initiatives, organizations are often on the fence due to its operational and management complexities. Each environment can have different requirements, specifications and configurations, and getting CloudOps up to speed with each one of them is not feasible for most. Besides, managing and monitoring several CSP contracts, pricing models, and SLAs can be daunting.

As such, organizations need an end-to-end, comprehensive multi-cloud management platform, like the emma platform, to abstract and manage the complexities of merging disparate private and public setups. Here's how the emma (enterprise multi-cloud management application) platform enables businesses to access LLMs in a democratic, vendor-agnostic, and flexible environment:



A Comprehensive Management Dashboard

The emma platform integrates infrastructure management, cost management, networking, and governance capabilities across on-premise, private, and public cloud environments into a single, powerful platform. Organizations can centrally monitor their LLM performance, resource utilization, security measures, and compliance seamlessly across all cloud environments.



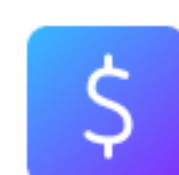
Seamless Integration

The emma platform is cloud agnostic, which means it can seamlessly integrate with all major hyperscalers, like AWS, GCP, and MS Azure, and regional providers. Organizations can utilize best-of-breed LLMs and ML platforms from any provider to stay ahead of the curve without being locked into a single platform.



Centralized Visibility and Control

The emma platform provides end-to-end visibility across diverse cloud architectures. Businesses can monitor performance, optimize resource utilization, enforce access policies, and ensure data governance and regulatory compliance across all on-premise and cloud environments.



Cost Optimization

The emma platform provides real-time visibility into resource utilization and cloud spend across all the disparate environments of a hybrid multi-cloud deployment. It uses ML algorithms to identify cost-saving opportunities and provides data-driven insights, enabling informed decision-making for cloud cost optimization.



Data Consistency and Service Interoperability

The emma platform uses abstractions to ensure consistent data, processes, and application instances across on-premise environments and multiple clouds. This ensures service compatibility and seamless data flow and communication between services deployed across different environments. This way, organizations can embrace cloud agnosticism and achieve seamless interoperability within a hybrid multi-cloud set-up.



About emma

At emma, we believe that cloud resources should be as accessible as electricity or the internet. That's why we created the emma platform — the world's first end-to-end, no-code cloud management platform that enables organizations to unlock all the benefits of multi-cloud (on-premises, private, public, and edge) without the usual complexities and security risks associated with multi-cloud operations. Discover the emma platform's unique features:

- 1** A unified dashboard for monitoring performance, resources, security, and compliance across all clouds.
- 2** A truly cloud agnostic platform for managing infrastructure and applications no matter where they are hosted.
- 3** No-code approach to enable infrastructure provisioning and configurations in just a few clicks.
- 4** Global networking backbone for high-performance connectivity to cloud services in 50+ regions and 150+ cloud locations.
- 5** An all-in-one, end-to-end cloud management solution, providing comprehensive cloud management, cost management, network management, and governance capabilities.

With the emma platform, businesses can maximize their cloud environments, drive innovation faster, and gain a decisive edge in a rapidly evolving business landscape, regardless of how they approach their multi-cloud strategy.

