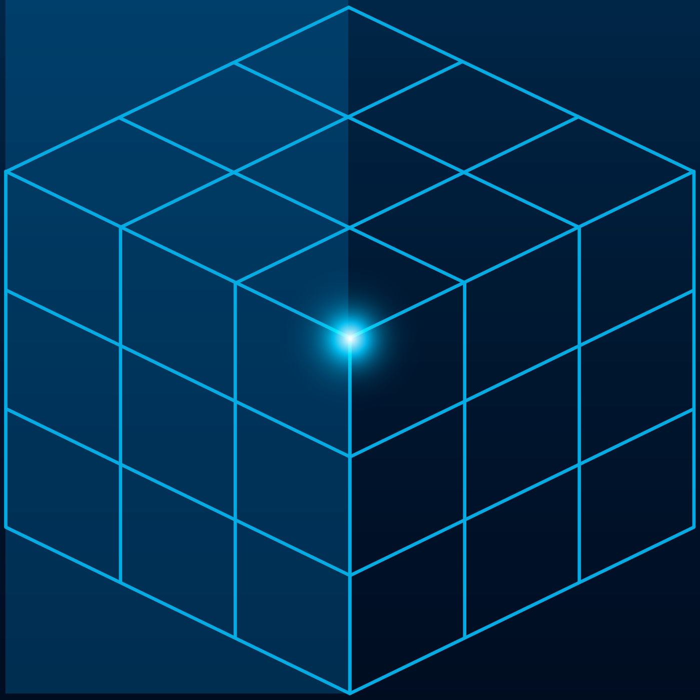
Infosys topaz

VIEW POINT

AI WILL MANIFEST IN SEVERAL WAYS WITHIN AN ENTERPRISE REQUIRING A POLY AI ARCHITECTURE



Infosys®
Navigate your next

Artificial intelligence (AI) is rapidly transforming the enterprise landscape and manifesting in multiple forms, necessitating a strategic approach to integration. This article explores the various manifestations of AI within organizations, the challenges they present, and the emergence of Poly AI architecture as a solution to maximize the potential of AI across the enterprise.

The Five Manifestations of AI in Enterprises

Infosys has identified five key categories of AI manifestation in enterprises as shown in Figure 1. Each manifestation offers different levels of business impact and varies in terms of time, cost, and complexity to implement.

1. **Consumer AI assistants:** Publicly available tools such as ChatGPT, Gemini, and Perplexity.ai that offer general-purpose AI capabilities
2. **Specialized AI assistants:** Enterprise-specific tools such as GitHub Copilot, AWS CodeWhisperer, and Atlassian Intelligence, tailored for specific business needs
3. **Custom AI apps using closed models and APIs:** Enterprises are developing applications such as Text2SQL for ERP, migration analyzers, and audit assistants using cloud-based AI models
4. **Custom AI apps using fine-tuned open models:** The growing adoption of open-source models enables the creation of specialized assistants for code, knowledge, and operations
5. **Industry-specific AI apps:** Highly specialized pre-trained models for specific industries, though currently limited in scope (for example, Alphacode 2 for competitive coding)

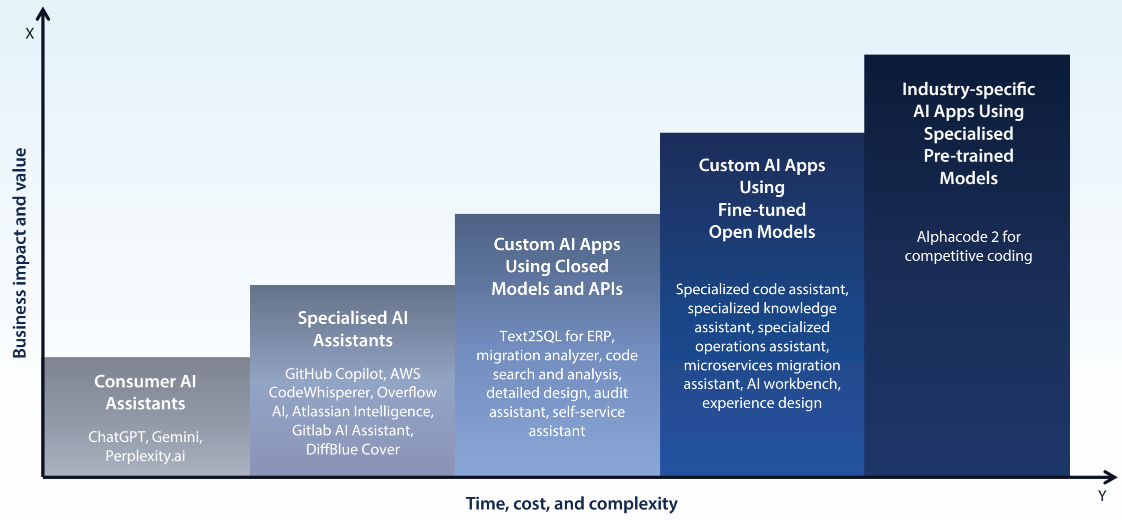


Figure 1: Five manifestations of AI within the enterprise

Source: Infosys research

Challenges in Enterprise AI Integration

As enterprises integrate various AI technologies, they encounter several challenges that can impede seamless adoption and utilization. The first key challenge is the use of multiple, disconnected AI solutions, which can introduce inefficiencies since each tool may require separate management, maintenance, and integration efforts. This disjointed approach can strain resources and reduce the overall effectiveness of AI initiatives. Addressing these challenges is essential for enterprises to fully leverage the benefits of AI and achieve a cohesive, efficient, and scalable AI ecosystem. The second major challenge is the fragmentation of experiences across different AI tools, leading to inconsistencies in user interactions and workflow disruptions. This fragmentation often results in varied and inconsistent processes, as disparate AI implementations may not align perfectly, causing inefficiencies and confusion within the organization. The most significant challenge is the lack of a unified data foundation, which can hinder the ability to derive comprehensive insights and make informed decisions. This aspect is covered exhaustively in Imperative 7.

Poly AI Architecture: A Solution

To address the challenges in enterprise AI integration, organizations are adopting a framework capable of managing the diversity and variability inherent in different AI platforms. This framework, referred to as the Poly AI architecture, provides a comprehensive solution to seamlessly integrate a wide array of AI technologies across the enterprise as shown in Figure 2.

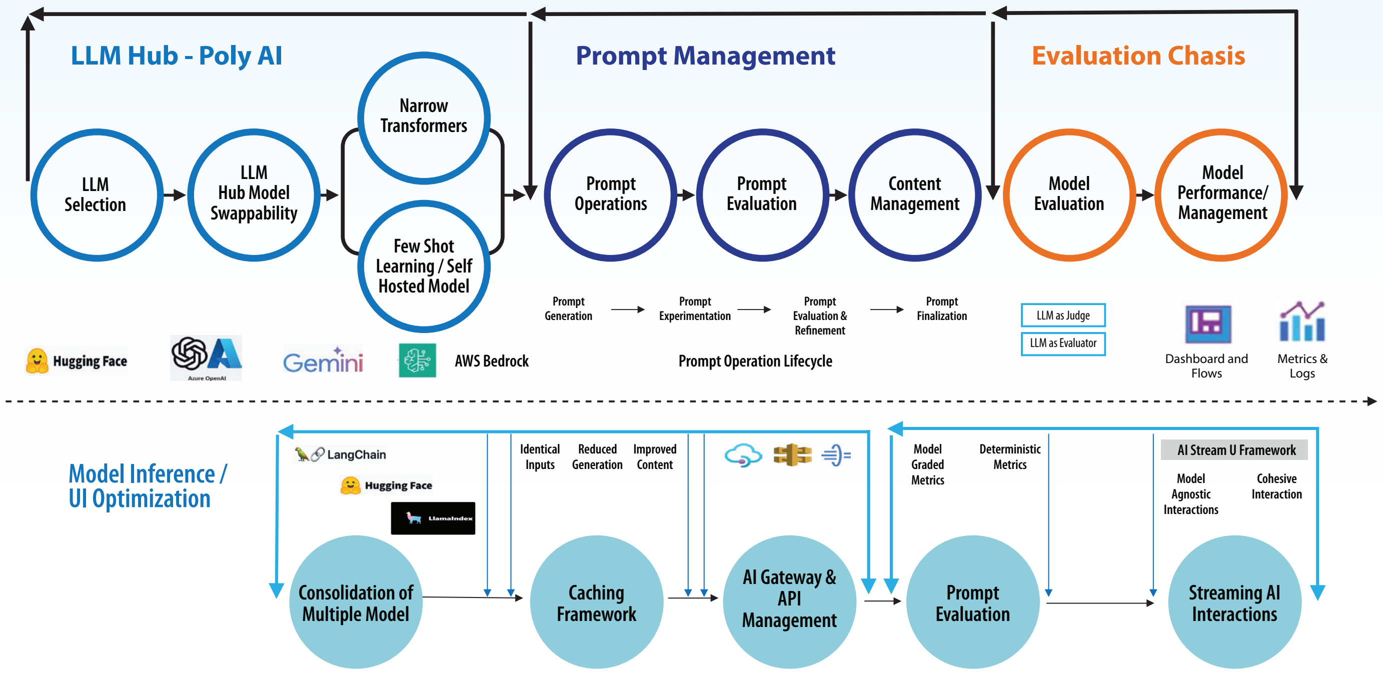


Figure 2: Poly AI architecture blueprint with key elements

LLM Hub

A large language model (LLM) hub is essential in Poly AI architecture for efficiently managing multiple AI models and ensuring seamless updates. It facilitates model swappability, which is crucial in the rapidly evolving AI landscape. The ease of swapping models depends on integration complexity, specific feature dependencies, and workflow impacts. For example, code generation models are easier to swap compared to embedding models, which require integrated automation workflows.

The LLM hub supports model versioning and switching, enabling updates or replacements without workflow disruptions. This is particularly valuable for adapting to new AI advancements or regulatory requirements. It also provides a unified interface for managing multiple AI tools, ensuring harmonious operations, simplifying compliance, implementing granular access controls, and maintaining audit trails—crucial for regulated environments.

Additionally, the LLM hub accelerates enterprise application development with customizable templates and pre-built components, improving efficiency and reducing time-to-market. It facilitates asynchronous collaboration, eliminating redundant efforts in data source connections or common actions. In the current AI landscape, an LLM hub is a necessity for leveraging AI effectively within the Poly AI architecture. It addresses model swappability complexities, ensures compliance and security, and streamlines the development and management of multiple AI applications. This helps organizations stay agile, innovative, and competitive.

Prompt Management

Effective prompt management is crucial in implementing Poly AI architecture, especially when migrating prompts across different AI models. To address the challenges of prompt adaptability and consistency, organizations should focus on two key strategies:

Implement structured prompt design: Implement methods to create well-structured prompts that ensure clarity and relevance in both input and output from LLMs. Frameworks such as Instructor guide users through structured thought processes, adapting to various tasks while maintaining contextual relevance. Similarly, Outlines helps in crafting prompts that are clear, focused, and effective in eliciting the responses desired from AI models. These frameworks standardize prompt structures, making them more consistent and adaptable across different models.

Optimize with management tools: Employ prompt management tools that enhance the creation, optimization, and versioning of prompts across various platforms. For instance, CoStar leverages advanced AI technologies to generate hyper-personalized responses based on user input, optimizing interactions and ensuring that prompts yield the best possible results.

By focusing on structured design and utilizing specialized management tools, organizations can create prompts that are not only effective but also maintain a high level of consistency and adaptability across different AI models. This approach ensures that prompts unlock the full potential of AI while remaining flexible enough to work across various implementations in the Poly AI architecture.

Evaluation Chassis

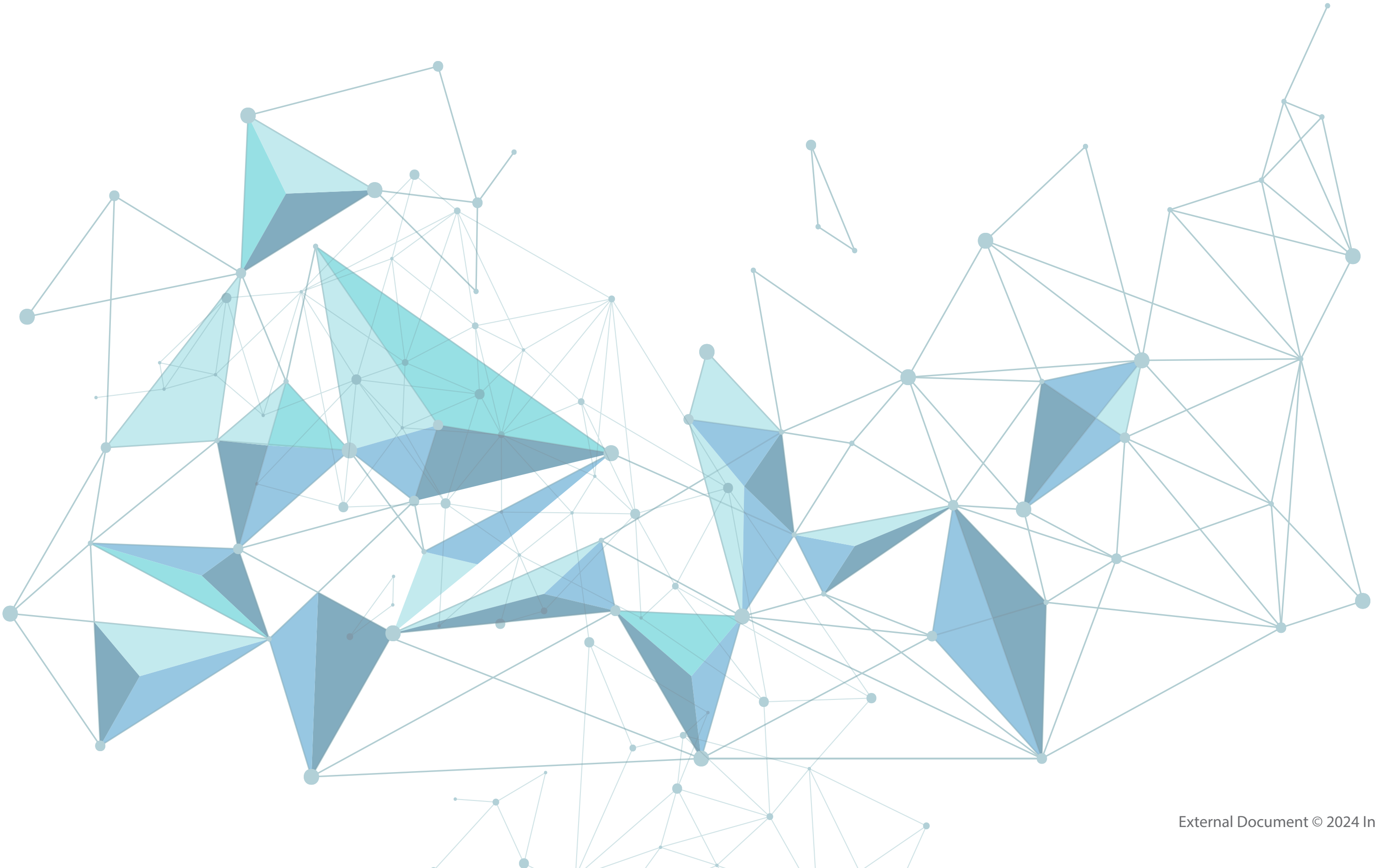
The evaluation chassis stands as a cornerstone of the Poly AI architecture, meticulously designed to assess AI model performance and maintain consistent quality across diverse models, particularly during model swaps. This robust framework leverages automated evaluation datasets tailored to specific domains, significantly reducing the need for manual verification by enabling precise performance measurements both before and after model migration.

A key innovation within this chassis is the integration of LLMs as judges, often referred to as LLM-as-a-judge or LLM-as-an-evaluator. This approach revolutionizes the evaluation process by facilitating rapid assessments prior to human involvement. LLMs acting as judges can quickly analyze model outputs, considering factors such as accuracy, relevance, and coherence, providing a preliminary evaluation that closely aligns with human judgment.

The dual approach of automated datasets and LLM-based evaluation not only streamlines the model selection and migration process but also acts as a safeguard against performance degradation. By leveraging the ability of LLMs to understand context and nuance, the evaluation chassis can identify subtle issues that might be missed by traditional metrics alone. This comprehensive evaluation strategy ensures a high level of reliability and operational efficiency in AI deployments, allowing for swift iterations and improvements.

Caching Framework

To optimize performance and resource utilization, a caching framework should be incorporated into the Poly AI architecture. This offers several benefits, including cost savings by eliminating the need to recompute responses for identical inputs, reduced generation latency, and improved content safety by serving previously vetted responses. The central theme of this is caching human preference data, where the system can store and reuse responses that align with user preferences, eliminating the need to reach out to the LLM for each request. This approach has several advantages such as efficiency, cost savings, reduced latency, and improved content safety.



AI Gateway and API Management

An AI gateway and robust API management are crucial for maintaining system stability and ensuring quality of service within the Poly AI architecture. By leveraging services such as Azure API management (APIM), AWS API Gateway, or GCP Gateway, Poly AI architectures can efficiently manage and throttle requests, ensuring system robustness under fluctuating demands. These tools provide structured approaches to rate limiting, authentication, and authorization, essential for preventing overload and securing the system. Centralized logging and monitoring of API usage patterns facilitate proactive issue identification and performance optimization, similar to MLflow's tracking component for experimental data.

Streamlining AI Interactions

The final aspect of the Poly AI blueprint tackles the fragmentation of experiences across various AI tools. We need to examine the frameworks we can incorporate in Poly AI to reduce the friction in user experience. A notable example is Vercel AI's StreamUI framework, which offers an effective solution to the fragmentation issue. By enabling the streaming of UI components directly from language models, StreamUI facilitates:

- **Model-agnostic interactions:** The underlying language model becomes transparent to the user, allowing seamless engagement with the application regardless of the specific LLM or ML model in use.
- **Cohesive interaction model:** By providing a unified interface for diverse AI functionalities, StreamUI ensures a consistent user experience across different AI tools and capabilities.

By adopting this strategy, enterprises can reduce the cognitive load on users and allow them to focus on their tasks without navigating the complexities of underlying AI models. The fluid and cohesive interaction model enhances user experience by promoting more natural engagement with AI-driven features. Furthermore, as new AI models or capabilities are introduced, they can be seamlessly integrated into the existing interface without disrupting user workflows.

Conclusion

The integration of AI within enterprises is a multi-faceted endeavor that necessitates a strategic and cohesive approach to maximize its potential. Poly AI architecture provides a robust framework to address the inherent challenges of deploying diverse AI technologies. By leveraging components such as LLM hub, prompt management tools, evaluation chassis, caching frameworks, and AI gateways, organizations can streamline AI interactions, reduce fragmentation, and ensure consistent performance across various AI models. As AI continues to evolve, adopting a Poly AI architecture will be crucial for enterprises to remain agile, innovative, and competitive. This will ultimately enable them to harness the full power of AI to drive business success.

Infosys®
Navigate your next

For more information, contact askus@infosys.com

Infosys.com | NYSE: INFY

Stay Connected