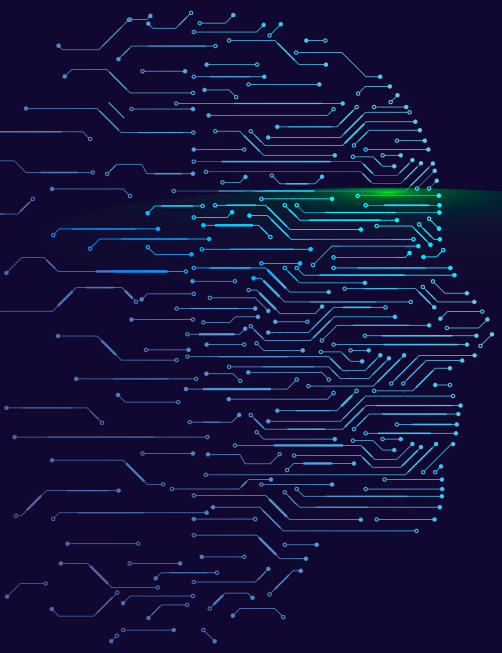
Infosys topaz

VIEW POINT

TOP 10 AI IMPERATIVES FOR 2024

Infosys®
Navigate your next



In the past year, generative AI technology has seen an exponential increase in growth and investments. It is taking a path into enterprises comparable to the one taken by mobile or digital transformation but at a more accelerated pace. Most enterprises have started working with generative AI. American companies plan to invest $5.6 billion over the next 12 months and European companies about $2.8 billion, according to the Infosys Generative AI Radar. They are all, however, concerned about data readiness and regulatory compliance, and acknowledge that very few of their initiatives have delivered tangible business value.

In the year 2024, we believe, most ongoing AI initiatives will be rolled out to production at scale and the number of AI initiatives will evolve from a handful of experiments to strategic AI programs that will help pave the way for AI-led business transformations. We have curated the top 10 learnings and takeaways from our own experience in scaling AI at Infosys that, we hope, will help guide our clients in their AI-powered transformations.

1 Models are perishable, but organizational data and knowledge are permanent

A year ago, GPT3/3.5 was the sole dominant large language model, but today, there are hundreds of options in both closed-access models (proprietary) and open-access models (open source) that are catching up in capability and innovation. There is an ongoing race for AI model development, and we anticipate further commoditization. That is why, we believe that the lifespan of models is short, and in the context of the enterprise, organizational data is crucial. This data can be used to fine-tune any model. In addition, the ability to choose a model that best fits a specific task or capability is crucial, and hence building abstraction into the architecture to then have the flexibility to choose the AI service provider and model will provide optionality and support ongoing innovations.

2 Hallucinations are features of LLMs and not really bugs

Most large language models exhibit a certain degree of hallucinations, either in the form of factual inaccuracies or outcomes inconsistent with the input and context. This is an important area of concern for most enterprises that use LLM-based AI products for customer interactions and critical business processes. The hallucinations are caused in most LLMs either due to the inherent quality of data or because of the way the model has been trained and inferenced. The extent of hallucinations can be minimized by removing misinformation and biases from the training data, by grounding and establishing knowledge boundaries using techniques such as retrieval augmentation, or by coupling it with a knowledge graph and minimizing training and inference-related hallucinations. Adoption of these techniques needs to be coupled with hallucination detection tests and benchmarks to continuously refine and minimize hallucinations.

3 Fine-tuned models for specialized tasks are very effective and can be developed at a fraction of the cost and time

While there is a race to build bigger models and we will soon hit the data ceiling for training these generalized models, within the enterprise context, the size of data and knowledge is relatively small. AI applications need to be contextualized and specialized. There is a need for both generalized pre-trained large models as well as smaller specialized models that are task-specific. A narrow transformer-based approach allows specialized fine-tuned models to be developed with minimal data. This narrow transformer approach involves using open-access models as a base, fine-tuning them with organizational data, coding, and building specialized private and secure models. We have found these models to be very useful for legacy applications, engineering, migration, modernization, and IT operations automation.

4 AI will manifest in several ways within the enterprise requiring a Poly AI architecture

Generative AI technology will manifest within enterprises in five different ways – Consumer AI assistants (such as ChatGPT and perplexity.ai), Specialized AI Assistants (such as Salesforce Sales GPT and Microsoft Copilot), AI products developed using closed access models (such as GPT-4), Specialized AI products built using fine-tuned open source models (like those based on Llama2) and industry-specific pre-trained models and products (BloombergGPT). Most enterprises will likely end up having a combination of these options. Considering the business and regulatory risks that can emerge, it is important to minimize lock-in to a particular provider or model within the architecture. Hence, developing a Poly AI architecture will provide the flexibility to use the best-fit option and cross-leverage AI capabilities across user journeys and business processes.

5 Strategic AI value map analysis identifies high business impact areas rather than siloed use cases

Every enterprise, today, has collated hundreds of AI use cases that could be implemented using AI and generative AI technology. These cases could potentially lead to productivity benefits, increased efficiency, and/or better user experience. However, most enterprises are struggling to articulate the business value and identify high-impact areas. In our experience, we have found that doing a strategic value chain analysis of the business process using the 'five AI evolution capabilities' adapted from Google DeepMind Levels of AGI paper, can help address this challenge[1]. The five AI evolution capabilities are - AI as a general-purpose tool, AI as an assistant to do specific tasks, AI as a collaborator to help accomplish a goal or task, AI-first experiences, and processes driven by AI with humans reviewing, refining, setting guardrails, and autonomous AI agents driving the process or operations autonomously with human overrides.

6 To drive ambidexterity, the left and right sides of the enterprise brain need to work together

Most AI use cases under development are for specific tasks such as coding, operations, learning, and summarization. While these use cases are narrow in scope, they can create medium to high impact.

To realize the capabilities driven by a strategic AI value map, an architecture-first approach is required. This approach enables the agility to evolve with emerging patterns, flexibility to use the right-fit model, velocity to run hundreds of AI projects at scale, Poly AI to support multiple AI evolution models, and responsible data management to enable compliance by default. To deliver maximum business value, the architecture needs to natively enable orchestration across organizational data, knowledge, systems, and models (ML, DL, LLM) with context, in a responsible manner.

7 Without robust and responsible data management, AI initiatives will fail to deliver value

To scale AI initiatives, most enterprises face challenges in ensuring data security, privacy, and data usability for training and fine-tuning models. To address these issues all enterprise data and content covering transactional, historical/analytical, syndicated, and training, both user-generated and machine-generated, needs to be brought under the responsible data management framework. This framework helps organize and fingerprint data and establish vertically integrated control functions to manage permissions and rights based on business needs. This will ensure compliance and governance across data at rest, in motion, and for consumption. Data readiness assessments will quickly help identify gaps and provide a foundation for developing an action plan to fix them.

8 AI-led transformation requires a bimodal approach to innovating with speed and scale

As generative AI becomes a general-purpose technology, new models, design patterns, and features will emerge. To incorporate these into the enterprise, a bimodal approach is advised. Establishing an AI foundry to experiment and incubate new technologies and develop new patterns and use cases will help the enterprise innovate at speed. The AI-factory-like approach will help bring in extreme automation and productization of learnings from the AI foundry. This approach will help balance and manage the risks associated with AI evolution while scaling its adoption within the enterprise.

[1] https://arxiv.org/pdf/2311.02462.pdf

9 AI regulations are continuously evolving and are fragmented today; being responsible by design cannot be an afterthought

There is a race to develop and enact regulations for governing AI, and this time the regulations are evolving alongside the technology. The coverage of regulations and their impact varies based on societal expectations and anxieties related to the impact of AI. For global enterprises, this means keeping track of emerging AI regulations across countries where they operate to ensure compliance. Existing processes, policies, guidelines, and tooling will need to be enhanced to cover model assurance, model security, bias, fairness, explainability, reproducibility, training data privacy, safety and alignment, IP/contractual risks, sustainability impact, and AI regulatory compliance. The tool landscape required to implement this is also fragmented. Additional efforts need to be factored in to integrate all these elements into engineering processes. In the absence of this integration, most AI developments may never reach the production scale.

10 AI will amplify the potential of all humans

AI has the potential to not only displace certain tasks and operations but it can also amplify several existing tasks. This enhancement can help us be less busy and more productive and allow us to dedicate more time to activities that add value. At Infosys, we are rolling out an AI Assistant to all our team members - developers, support engineers, consultants, sales personnel, learning consultants, and managers - so that they can use these tools to be more productive and efficient. For an AI era, the existing talent also must be transformed for individuals to become AI-aware, AI Builders, and AI Masters. AI-aware personnel should know how to use AI tools to collaborate and co-create. AI Builders should use AI capabilities as a utility, product service, or API to build AI-embedded or AI reimagined solutions and AI Masters will build models and novel techniques that operate at scale and lower costs to drive value to the business.

We have curated the top 10 learnings and takeaways from our own experience into Infosys Topaz, our AI-first set of services, solutions, and platforms using generative AI technologies. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale, and build connected ecosystems.

For more information, contact askus@infosys.com

Infosys®
Navigate your next