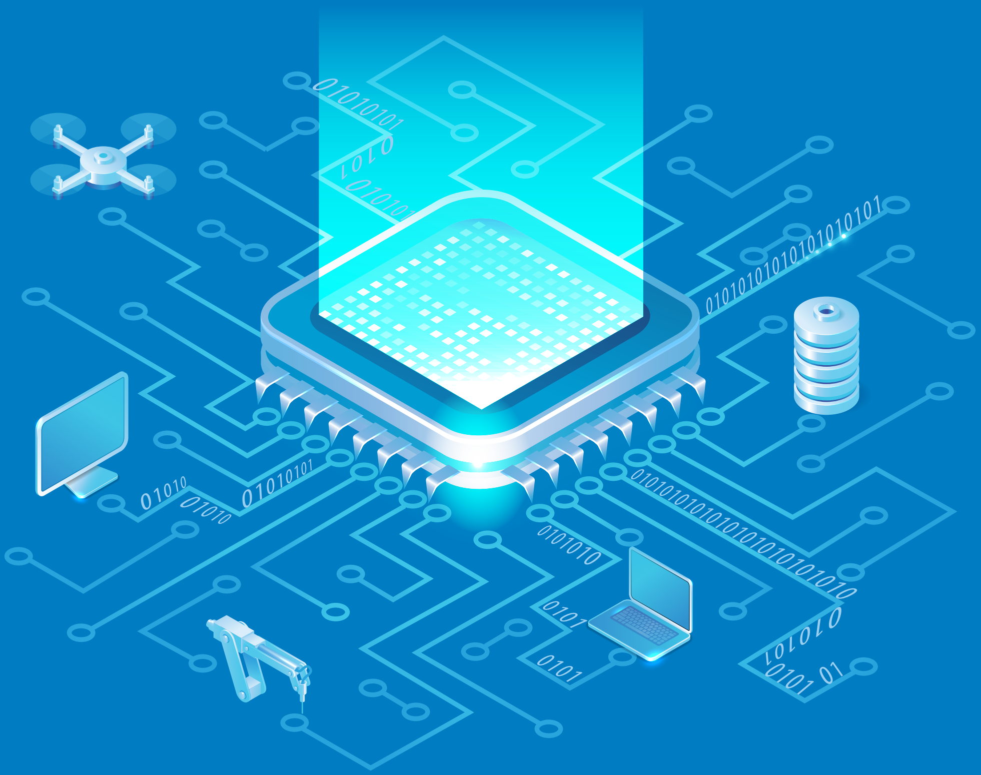
TECH NAVIGATOR: AI ENGINEERING EXCELLENCE

Infosys®
Navigate your next

Infosys® | Knowledge Institute

Contents

Develop and nurture top-tier coders



We are seeing that coders are in short supply and overworked, and they suffer from burnout and reduced productivity and efficiency. Fortunately, new tools are helping them cope with burnout and adding value to the organizations that deploy them.

Microsoft is an AI-first business investing heavily in this area, having recently committed $10 billion to research and development at OpenAI[1], solidifying its position as a leader in the AI industry.

Tools such as OpenAI's Codex, GitHub's Copilot and ChatGPT can be used to complete lines of code and to find bugs. They also create code in most programming languages from a natural language prompt. AI-based techniques assist in other stages of the software development lifecycle, including gathering requirements, design, deployment, and maintenance.

These AI tools, based on large language models, can perform a variety of tasks, but they demand heavy compute for inferencing. Infosys addresses this issue by building narrow transformers. Here, an appropriate smaller language model is used as the foundation model, and then fine-tuned with domain and task-specific data to augment a specialized task, such as code completion.

Full fine-tuning updates all parameters of the foundation model, demanding higher compute, and when there are multiple downstream tasks, full fine-tuning results in a new model for each task demanding more storage.

Infosys adopts parameter-efficient fine-tuning (PEFT) methods to build narrow transformers as these methods enable adaptation of foundation models to downstream tasks with fine-tuning only a small number of (extra) model parameters.

This results in reduced computational and storage costs, while achieving performance comparable to that of full fine-tuning. These narrow transformers are built and served at scale without external compute, thus ensuring data security.

In the DevSecOps[2] age, developers are not just responsible for writing code, but also for setting up the production environment, and for assembling, integrating, customizing, and maintaining applications once they are in production. This burden is now carried by another 2023 software tech trend – platform engineering[3].

Platform engineering accelerates development velocity and reduces overload through self-service capabilities — standardized tools, components, and automated processes — offered through an internal development platform (IDP).

In fact, Gartner expects that by 2026, 80% of organizations will establish these platform teams as internal providers of reusable services, components, and tools for application delivery[4].

Beyond AI-driven coding, these platforms encourage consistency and efficiency in software development and provide relief from the management of delivery pipelines and low-level infrastructure.

As AI tools evolve, the software development sector is in for an operational and financial windfall. If firms can motivate their workforce to adopt this advanced intelligence and upskill in the art of AI prompt engineering, the future will be an AI-first, continuously learning and evolving organization – an AI-first Live Enterprise[5].

AI-first for better operations

In the first AI wave (H1), systems and methods were predominantly used to supervise and optimize operations. As we move into the AI-first era, these methods will pave the way to further improve operations and increase efficiency. We will see the onset of newer frontline ML methods such as reinforcement learning, backed by meta-learning-driven dynamic control.

Our paper on a self-driving cloud for greener business[6] discusses this trend, where companies analyze data using ML techniques to make informed decisions and take proactive measures. The dynamic control feature also means the system can learn and adapt to new scenarios on the fly, driving even more efficient operations.

AI-first operations improve customer and employee experience through generative AI. Using AI assistants, an AI-first organization can handle internal and external user queries at pace, improving productivity and innovation – and they also become companies that people want to work for. Data-driven organizations that care about post-sales customer engagement improve employee retention, finds our Digital Radar 2023 research[7].

These AI systems use machine translation to interact with users, driving localized contextual conversations that add further data points and improve efficiencies and insights. Many applications are possible, including voice assistants and recommendation systems. Developers and organizations can build applications and services that understand and interact with users in more natural and intuitive ways.

AI-first customer-centric operations systems deliver the following capabilities:

- Offer multilingual support.
- Improve efficiencies through task automation and FAQ, reducing reliance on human professionals.
- Customize responses based on the learnings from customer interactions.
- Increase UX and CX through operations efficiency for CRM and ERP integrations.
- Enhance operational scalability by reducing the necessity for additional customer interactions, such as during Samsung/iPhone launches in the retail and technology sectors.

This technology is also well suited to solving problems across industries such as telecommunications, utilities, and retail.

Google's Dialogflow is used by Optus, one of the largest telecoms providers in Australia, to power virtual agents in a support application. Because the technology comes with prebuilt agents, in-depth programming knowledge is not required, and time to production accelerates. For example, prebuilt agents answer requests such as "I need help paying my bills" or "I haven't received my order, where is it?" without requiring custom programming.

These capabilities can be extended to other industries too. Healthcare providers can extend healthbot instances to include novel scenarios and integrate them with other IT systems and data sources., for example. AI platforms allow the creation of operations with self-heal capabilities, anomaly detection, automated monitoring, and alerting. For example, predictive maintenance is a great use case in the utilities sector.

The AI organizations of the future will need to respond in close to real time to queries across platforms, including mobile, web, chatbots, smart devices, interactive voice response systems, and messaging apps.

AI-first operations, using advances in NLP and transfer learning, are the future of time-limited, data-driven conversations, extending from internal support to customer contact centers.

Data and MLOps to drive velocity

In H2, data engineering was key: in our 2021 paper Scaling AI: Data over Models[8], we estimated that between 25% and 60% of machine learning project costs at that time was spent on manual data labeling and validation.

Firms had to manage data lineage and build systems with active learning, in which a classifier examines unlabeled data and selects part of this data for further human labeling. For the process to operate effectively, machine learning systems needed to be efficient, scalable, and reliable. This landscape also required a central model repository and trustworthy AI practices.

Many firms are still working in H2, and technologies like Azure ML, AWS SageMaker, MLFlow, and products like DataRobot and Iguazio, are emerging as sources for model management, deployment, and managing training data. Meanwhile, our clients require online and offline feature storage for ML data management and monitoring.

We are now in the H3 era, and all eyes are on generative AI and the models that underpin these tools. Building these models is a complex process. It can take large firms several hundred days and thousands of CPUs and GPUs to build a new large language model.

Creators of these models now rely on MLOps to support scaling techniques, including data parallelism, pipeline parallelism, and tensor model parallelism. With data parallelism, tasks are run in parallel, with data divided into partitions, and the models run on separate subsets of the data, increasing model training speed. Model parallelism, as the name suggests, divides a massive complex model either vertically or horizontally, with different parts of the model running on the same data. In this way, Data + MLOps techniques increase operational efficiency for H3 technology providers.

It is now believed that companies like OpenAI and Google are harnessing generative AI methods to make their MLOps pipeline even more sophisticated, creating a meta-robot that can build even better robots.

For instance, Chat GPT's efficiency comes from chaining together several distinct models — starting with a regular large language model, creating a reward model with human feedback, and finally using reinforcement learning with human feedback (RLHF). This removes the operational burden from their MLOps teams as there is no need for the data engineering tasks used in H2.

In H3, the big question is whether MLOps will become obsolete for firms using ML technology. Will it fade into obscurity, or will it evolve to suit the needs of users of LLMs and generative models?

We believe that, even in organizations that buy out-of-the-box generative AI solutions, H3 models will be tailored to specific use cases, and will require MLOps to bring all components together to reduce operational complexity and increase velocity of AI products. Systems will be created that integrate several generative AI models, forging a fusion of models that's greater than the sum of its parts.

With this in mind, firms ought to implement maximum automation across the entire gamut of data engineering and model lifecycle management, ranging from training and inferencing to API abstraction and toolkit engineering.

This is vital due to the upcoming LLM landscape that includes content such as text, images, audio, etc., and multiple models that operate on disparate data.

To make this data mesh operational, we require more advanced AI factory operations that leverage MLOps. AI engineering life cycle management, part of our PolyAI suite of services (Figure 1), is an Infosys approach to MLOps that enables data scientists to use ML development tools of their choice and train and deploy their models at enterprise scale without having to deal with engineering complexity.

The approach also supports multiple versions of leading AI frameworks such as TensorFlow, PyTorch and others, and it also maintains the traceability of model artifacts while enabling versioning and sharing of artifacts among development teams.

Figure 1. PolyAI services portfolio at Infosys

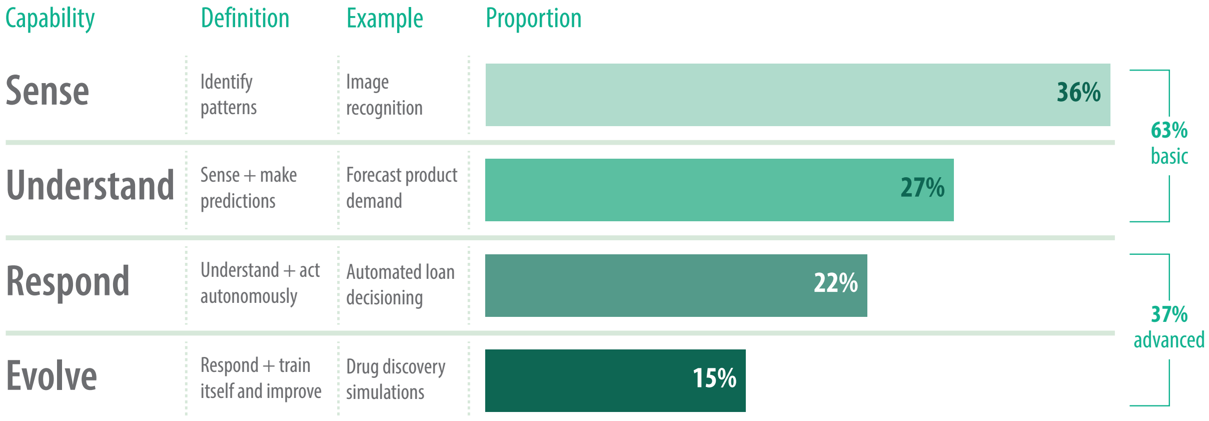
Managed models, datasets, pipelines, and end points	Frameworks (orchestration, technologies, etc.)
Open models	Open frameworks, including LangChain, Haystack, Ray Serve etc.
Closed models	Hyperscaler frameworks, including SageMaker, Azure OpenAI etc.
Hyperscaler models	Closed models (APIs)

Source: Infosys

AI systems that evolve

AI systems should also be built to evolve, or improve, with time. In Data + AI Radar[9], we introduced the **SURE** taxonomy. Here, an AI system moves from **S**ensing, to **U**nderstanding, to **R**esponding, and finally to **E**volving. Evolve, therefore, is the most advanced type of AI system, with models that are self-supervised and incorporate RLHF. At present, only 15% of firms (Figure 2), including giants such as Apple, Meta, OpenAI, and Netflix, can achieve these top-level capabilities.

Figure 2. Only 15% of firms achieve evolutionary AI design capabilities



Capability	Definition	Example	Proportion
Sense	Identify patterns	Image recognition	36%
Understand	Sense + make predictions	Forecast product demand	27%
Respond	Understand + act autonomously	Automated loan decisioning	22%
Evolve	Respond + train itself and improve	Drug discovery simulations	15%

Source: Data + AI Radar, Infosys

So, what is an "Evolve" system, one that can respond, train itself, and improve?

For enterprises to leverage foundation models used in generative AI, they need to do three things: acquire up-to-date knowledge; perform advanced reasoning; and use actuation to make them more useful, such as automating business workflows.

1. If the system doesn't have enough knowledge, this knowledge will have to be continually updated by the system from the outside in. This could be enterprise domain knowledge or data gleaned from searching the internet. OpenAI stopped training GPT-4 in September 2021, so ChatGPT, which is based on that large language model, has no knowledge of events after that date. To hedge against poor outcomes, OpenAI uses evolutionary design principles: ChatGPT 4 has a modular plugin architecture so that other applications can plug into it and provide additional services, including up-to-date knowledge or insights.

2. If the question requires advanced reasoning capabilities, then the evolutionary architecture can use chain-of-thought (CoT) prompting – a series of intermediate reasoning steps – that increases the ability of LLMs to perform complex reasoning.

3. Plugins for actuations are also available. If you want to book a flight, Expedia can plug in to OpenAI's architecture, for example; for e-commerce, there's an Amazon API.

These systems must be trustworthy. For this, we need external control. An external control system — sort of like a master control plane such as LangChain[10] — feeds the question or prompt to the foundation model and uses APIs to orchestrate behind-the-scenes plugins, gets the answer, and then feeds the response back to the user.

> Companies need advanced AI if they are to achieve the loftiest ambitions of AI and stand out from competitors.
>
> **Sunil Senan**
> Senior vice-president and head of data and analytics, Infosys

Further, evolutionary systems like OpenAI also use human preference data (such as asking for thumbs up/thumbs down prompts) to continuously revise itself and offer answers that get better over time. This is an example of RLHF[11], which drives improvement in the model over time.

This is one kind of evolutionary design, using a modular architecture to iron out deficiencies. Another approach, used by Microsoft's Bing and DeepMind's Sparrow, is to continuously update knowledge by retrieving metadata. Bing does searches daily — retrieving information and then sending it to the LLM for training or fine tuning.

References

1. Microsoft invests $10 billion in ChatGPT maker OpenAI, Dina Bass, January 23, 2023, Bloomberg.
2. Infosys Tech Compass, 2022, Infosys Knowledge Institute.
3. Platform engineering, Vishwanath Taware et al, 2023, Infosys.
4. What is platform engineering?, Lori Perri, October 5, 2022, Gartner.
5. The Live Enterprise: Create a continuously evolving and learning organization, Jeff Kavanaugh and Rafee Tarafdar, 2020, Infosys.
6. A self-driving sustainable cloud for greener business, Rajeshwari Ganesan, Professor Ravishankar K. Iyer, and Harry Keir Hughes, March 2, 2023, Infosys Knowledge Institute.
7. Infosys Digital Radar 2023: The next digital frontier, Harry Keir Hughes, 2023, Infosys Knowledge Institute.
8. Scaling AI: Data over models, Rajeshwari Ganesan, Sivan Veera, and Harry Keir Hughes, May 3, 2021, Infosys Knowledge Institute.
9. Data + AI Radar 2022: Making AI real, Chad Watt, 2022, Infosys Knowledge Institute.
10. Welcome to LangChain, Harrison Chase, 2023, Python.
11. Illustrating reinforcement learning from human feedback (RLHF), Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla, Dec. 9, 2022, Hugging Face.

Authors

Rajeshwari Ganesan
Distinguished technologist, Infosys

Rajeev Nayar
CTO of data and AI, Infosys

Kamalkumar Rathinasamy
Distinguished technologist, Infosys

Rafee Tarafdar
CTO, Infosys

Kate Bevan
Infosys Knowledge Institute

Harry Keir Hughes
Infosys Knowledge Institute

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision-making on critical business and technology issues.

To view our research, visit Infosys Knowledge Institute at infosys.com/IKI or email us at iki@infosys.com.

Infosys®
Navigate your next

For more information, contact askus@infosys.com