UCL
SCHOOL OF
MANAGEMENT

ANALYTICS
LAB

Capgemini

THE HARD TRUTHS OF SCALING GEN AI

Authors
Paul Walton (Capgemini)
Chris Dawkins (Capgemini)
Ajay Mistry (Capgemini)
Jamie Rich (Capgemini)
Nensi Ahmetbeja (UCL student)
Kamen Ivanov (UCL student)
Tuhan Sapumanage (UCL student)
Kayla Gosal (UCL student)
Chen Luo (UCL student)
Idil Karaduman (UCL student)
Aydin Orhan (UCL student)
Daisy Yang (UCL student)
Jason Wu (UCL student)
Divyam Thapliyal (UCL student)
Ilham Wicaksono (UCL student)
Jiaying Li (UCL student)

INTRODUCTION

SCALING GEN AI IS ESSENTIAL FOR DELIVERING THE VALUE OF GEN AI, BUT MANY ORGANISATIONS ARE FINDING IT DIFFICULT.

This article analyses the challenges of scaling GenAI and how these can be resolved.

The article was written by Paul Walton, Chris Dawkins, Ajay Mistry, Jamie Rich (Capgemini) and UCL students: Nensi Ahmetbeja, Kamen Ivanov, Tuhan Sapumanage, Kayla Gosal, Chen Luo, Idil Karaduman, Aydin Orhan, Daisy Yang, Jason Wu, Divyam Thapliyal Ilham Wicaksono, Jiaying Li as part of a project between Capgemini and the UCL School of Management.

GenAI promises many changes to many organisational functions. But a combination of rapid technology change, the need for new processes and skills and many new risks make scaling difficult and uncertain. Scaling can only be successful if it addresses the underlying challenges head on and tackles change from four perspectives in parallel:

- **Organisational change**—GenAI is changing the nature of work and in the process, changing business processes, governance and skills in parallel for many different organisational functions and these changes are inter-connected;
- **Delivery processes**—GenAI is revolutionising software engineering and is enabling significant productivity improvements;
- **Architecture**—the new capabilities of GenAI and rapid technology change require an evolving approach to technology and data architecture;
- **Adaptiveness**—the future impact of GenAI is difficult to determine so the journey will need to be able to change as circumstances demand.

GEN AI REQUIRES WIDESPREAD CHANGE

Because of the opportunities it offers and the risks that it introduces, implementing GenAI at scale needs changes to many different parts of the operating model.

The underlying challenge with GenAI is that it is very different from conventional technology. Existing practices for technology-enabled change are based on a set of assumptions that do not always apply to AI—demonstrated by the world-wide scramble to regulate the technology (for example, the EU AI Act and the introduction of new approaches to assurance).

This means that GenAI at scale needs changes to all of the following:

- **The nature of work** – GenAI-enabled processes will change the nature of work, so the level of organisational change required is large, requiring new processes and skills;
- **Risk** – GenAI introduces new risks (e.g. legal/commercial, ethical, technical, automated pace and complexity, a new relationship between people & technology, sustainability) and the mitigation of these risks requires new capabilities;
- **Technology & architecture** – The GenAI technology landscape is changing fast and new capabilities (e.g. agentic systems) are being introduced. These all require fundamental changes to technology, data architecture and governance, with careful consideration about the right architecture and implementation approach;
- **Delivery** – GenAI is revolutionising software engineering and the assurance it requires. In addition, GenAI is becoming an increasingly important component in many different types of technology (e.g. in Internet tools, collaborative tools (e.g. Microsoft's Copilot tools), application platforms, cloud services and bespoke AI models). The introduction of GenAI at scale requires changes to the delivery approach for all of these.

MORE THAN TECHNOLOGY NEEDS TO SCALE

To scale GenAI, it isn't enough to scale the technology. All the different changes required need to be able to operate at scale. For example, an initial response to GenAI risks has been to constrain the possibilities for using GenAI and to police them using an expert team. But, given the shortage of skills available, this approach can quickly become a bottleneck and limit scaling.

The same applies to building the right skills, the right assurance model, understanding how to automate processes, how to build GenAI-enabled user interfaces and so forth. All of these need to be able to operate at scale with the right skills, processes and governance.

Figure 1: Widespread changes are required to the operating model

TOO MUCH UNCERTAINTY

This is why scaling GenAI is hard—the scaling approach needs to coordinate all of these different changes the operating model moving at different speeds. If the right mechanisms are not in place, then any attempt to scale GenAI encounters considerable uncertainty which blocks delivery.

The table below shows how the blockers affect the product lifecycle for GenAI-enabled products.

Table 1: The blockers to scaling GenAI apply throughout the lifecycle

	Identifying the use case	Proving the concept and value	Scaling reliably and controllably
Product delivery	Aligning with AI strategy Understanding the impact of regulatory and compliance changes	Understanding sources of value, risks and costs Building a realistic roadmap Defining appropriate measures for value, risk and cost	Maturing the processes for developing, monitoring & supporting AI products and platforms Improving the productivity of delivery processes using GenAI Managing fast-moving technology changes Managing widespread changes in organisational context
Governance & controls	Understanding the ethical and governance limitations on the scope of the use case	Understanding the various risks	Implementing processes and automated governance at scale for different types of GenAI technology Managing risks effectively Catering for regulatory and compliance challenges and changes
People, skills & culture	Likelihood of the availability of the required skills for implementation	Proving the viability of the people element of the roadmap	Understanding the extent of culture change required Implementing culture changes Understanding the extent of skills and

			competency model changes required Implementing skills and talent changes Overcoming resistance (and the fear of redundancy)
Business Process design	Picking the right level to automate and how much process redesign is required (which depends on changing technology capability) Ensuring a detailed implementation of ethical principles, governance and risk management requirements	The need to redesign processes and ways of working using the capability of GenAI Proving how processes can be trustworthy and reliable	Integrating different redesigned processes Incorporating AI governance and risk management at scale
Data & org knowledge	Understanding the data and information required	Proving the viability of the data element of the roadmap	Quality and extent of required data Quality of data governance Accessible organisational knowledge Security and coding standards
Tech & architecture	Selecting appropriate technology for the use case	Proving the viability of the technology element of the roadmap Proving how the technology can be trustworthy especially when new ideas (like agentic AI) introduce extra complexities that need careful mitigation	Technical debt Selection & implementation of AI platforms and tools (to cover all use cases) Guardrails Model scalability Implementation of a new architecture

The difficulties are compounded because of the pervasiveness of GenAI technology. GenAI will be used in a range of technology types as shown in Table 2.

Table 2: GenAI is being incorporated in many types of technology

Type	Description
Public Internet tools	Standard publicly-available tools (like ChatGPT)
Collaboration tools	Standard enterprise collaboration tools (like the Copilot tools included in Microsoft365)
Low code	So-called no code/low code tools that enable people who aren't professional developers ("citizen developers") to develop apps (like the Microsoft Power Platform)
Chatbots	Chatbot tools that will increasingly use the language capabilities of GenAI for accessing organisational data (perhaps through Retrieval-Augmented Generation (RAG))
Enterprise applications	Enterprise applications (like SAP, Salesforce or ServiceNow amongst many others) that are increasingly using GenAI capabilities
Fine-tuning	Large Language Models (or smaller language models) trained on organisational data ("fine-tuning")
Bespoke	Bespoke models created for specific purposes like the uses in healthcare
Hybrid	Hybrid AI that combines a statistical AI approach (e.g. using LLMs) with more conventional logic-based technology
Agentic systems	Collaborating AI agents with more autonomy embedded in wider systems

As the technology matures, other forms of use will be developed but this list already shows the impact of the technology.

The scaling challenges affect these different types in very different ways. For example, there are no technological challenges of scaling the use of Copilot (because the scaling has been carried out by Microsoft) but its use does present challenges in mitigating ethical and legal risks.

WHAT ARE THE FUNDAMENTAL CHALLENGES IN SCALING GEN AI?

Figure 2 summarises the discussion above. It shows four types of challenge:

- **Use cases:** identifying value and the use cases to implement;
- **Architecture:** scaling technology and data architectures that incorporate evolving AI technology;
- **Organisational change:** creating an organisational context in which the ethical and reliable use of GenAI can thrive;
- **Enabling technology delivery with GenAI:** delivering technology using GenAI to improve productivity while implementing the new governance and assurance required to mitigate the risks.

The figure also references, in the centre, the core question: how can the delivery process reconcile the delivery of business value and user experience improvements with the ability to deliver? When will the various enablers (organisational, architectural or delivery related) be ready to deliver a particular use case at scale?

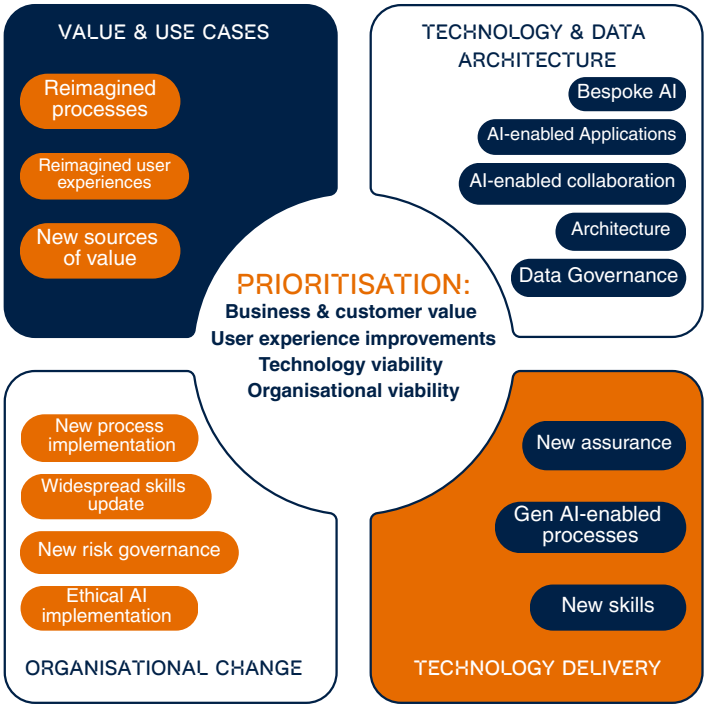

Figure 2: The fundamental challenges of scaling GenAI

Fortunately, for the five elements of Figure 2 there are well-established techniques available. These are shown in Figure 3,

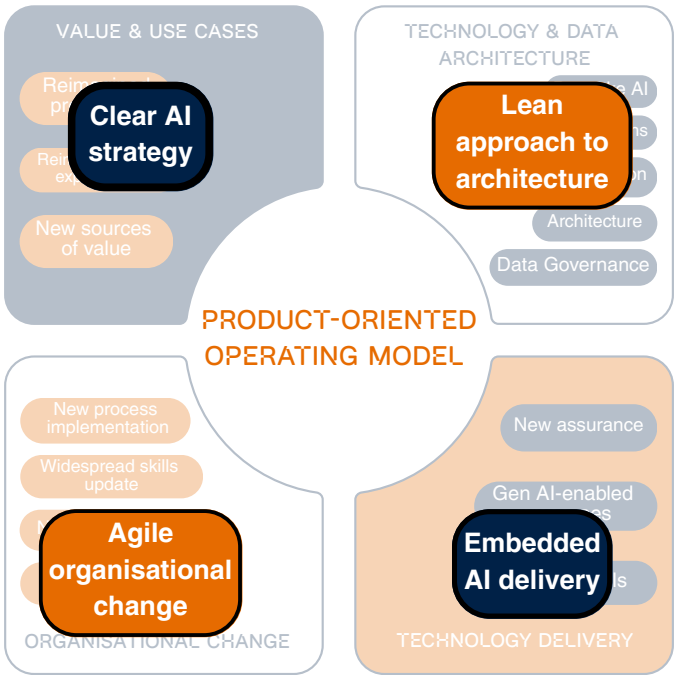

Figure 3: Scaling GenAI: the ingredients of success

AI strategy will clarify how AI can deliver value and the relevant use cases.
The techniques of lean architecture are designed to support the challenges of change of the type presented by AI.
Over the last couple of years, the techniques for embedded AI delivery have become clearer.
Agile organisational change has been developed over many years to support digital delivery more generally. Finally, a product-oriented operating model can pull all of the threads together in a form that provides understanding the a firm basis on which to make good quality decisions about roadmaps and priorities.

CONCLUSION

The journey to scaled GenAI can only be successful if it includes five ingredients:

- **AI strategy:** that identifies sources of value and use cases;
- **Lean architecture:** that can adapt to the evolving AI technology landscape;
- **Agile organisational change:** able to create the right context and culture for exploiting AI;
- **Embedded AI delivery:** delivery mechanisms that can both exploit the use of AI for delivery and embed the controls needed at scale;
- **Product-oriented operating model:** able to connect the different pieces into a roadmap that realises the value.

ABOUT US

CAPGEMINI

Capgemini is a global leader in consulting, digital transformation, technology and engineering services. The Group is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms.

THE UCL SCHOOL OF MANAGEMENT

The **UCL School of Management** is the business school of University College London, one of the world's leading universities, consistently ranked in the global top 20 for its academic excellence and research. The School offers innovative undergraduate, postgraduate, PhD and executive programmes in Management, Entrepreneurship, Business Analytics, Business Information Systems, and Finance, designed to prepare students for leadership roles in the

THE ANALYTICS LAB

The Analytics Lab is an enrichment module for UCL students where they are able to explore topical questions in the domain of analytics and digital economy via hands-on experience. Students are offered the opportunity to conduct research and work on projects with leading technology service and consulting companies, discuss and develop their own ideas and projects.

It aspires to help UCL students and alumni to be in the heart of fundamental changes and digital transformations in the business environment primarily but not limited to. Students enhance their practical abilities to manage analytics and digital operations effectively in view of rapidly developing technological advancements in the relevant domains.