How to Deploy Generative AI Capabilities Behind the Firewall to Augment Your Workforce

Published 1 June 2023 - ID G00790749 - 11 min read

By Analyst(s): Nader Henein

Initiatives: Artificial Intelligence; Data and Analytics Programs and Practices; Evolve Technology and Process Capabilities to Support D&A; Future of Work; HR Technology Strategy, Transformation and Management

This is a CIOs guide to deploying generative AI by using microapps to demonstrate the value of the technology without exposing the business to new and unexplored risks. The example used demonstrates how to boost employee productivity and address the skills shortage by augmenting knowledge workers.

Overview

Key Findings

- Generative AI, specifically large language models, come with a unique set of risks when compared with other, more common AI implementations.
- Generative AI can and will change how organizations design jobs, resource tasks and allocate responsibilities across nearly all facets of the enterprise.

Recommendations

CIOs seeking to adapt and deploy generative AI capabilities within their organization should:

- Develop generative microapps to orchestrate the interaction with LLMs behind the firewall, rather than using conversational LLMs, to mitigate new and unexplored risks unique to the technology.
- Focus on building augments for their workforce with clear alignment to business targets rather than seeking new challenges that generative AI can address.

Strategic Planning Assumptions

By 2026, 50% of office workers in Fortune 100 companies will be AI-augmented in one form or another, either to boost productivity or raise the average quality of work.

By 2026, one in five of companies that rushed into replacing entry-level employees with generative AI bots will find that existing skills shortages are exacerbated by the lack of internally developed talent and leaders.

By 2028, the rate of unionization and industrial action among knowledge workers will increase by tenfold as a direct result of the acceleration in the use of generative bots for tasks that were previously managed by those workers.

By 2028, one in four of G20 countries will introduce laws to either shield or compensate skilled office workers having their work-product supplement generative models that would ultimately replace them.

Introduction

A wise man once said, “Beware of shiny new toys.” When the iPhone was first released, a lot of executives walked into their respective organizations and demanded to have their email set up on their shiny new phone. This caused a substantial amount of chaos because the iPhone was not ready for enterprise connectivity or long-established security requirements and would not be ready for another 18 months.

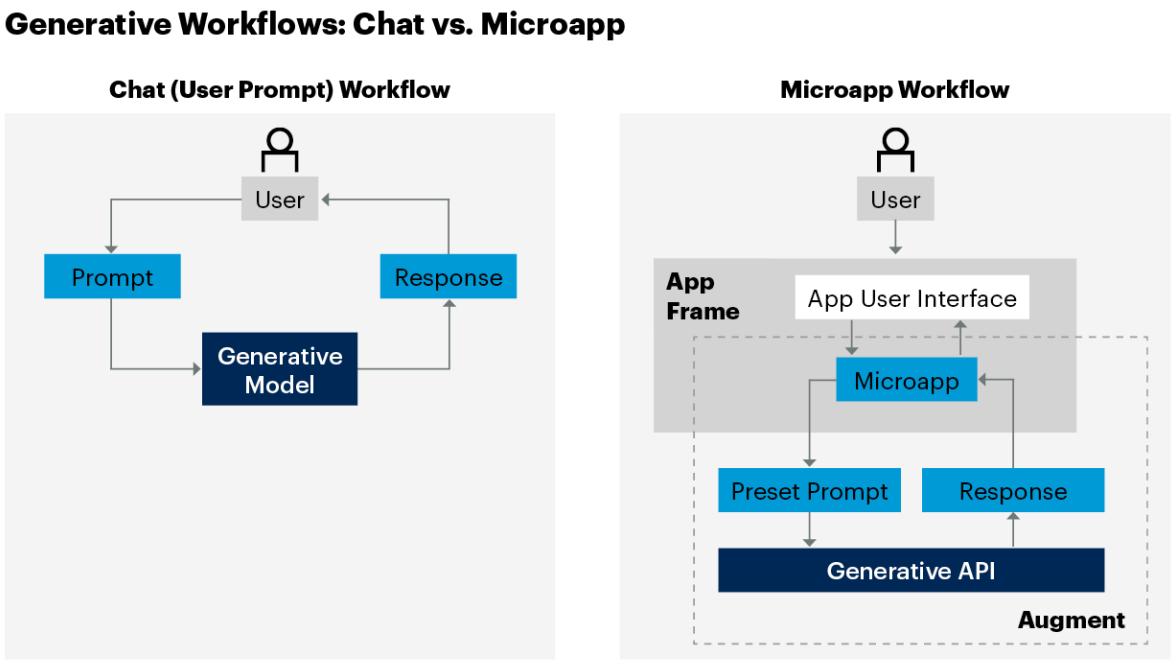
Generative AI, specifically large language models (LLMs), show substantial promise and hold even more potential than most realize. This note provides CIOs with a high-level roadmap to develop generative microapps as a vehicle to deploy LLMs behind the firewall and mitigate key risks unique to the technology.

What are generative microapps?

They are applications that act as a proxy between the user and an LLM. The app has a preprogrammed set of prompts that address a focused (micro) number of tasks. The prompts are used to query the model and receive responses in a predefined format. This in turn makes it easier for the logic within the microapp to validate each response before passing them on back to the user. There is no conversational/chat interface (UI) available (see Figure 1).

Generative microapps can be stand-alone, but in most instances, they will be embedded as extensions to productivity platforms commonly used by knowledge workers. Microsoft's Copilot integrated into Office 365 tools such as Word or Excel are early examples of general-purpose microapps, albeit with some conversational capabilities. [1]

Figure 1: Generative Workflows: Chat vs. Microapp



Source: Gartner
790749_C

Gartner

How will users interact with microapps if there is no chat interface?

We will use the following example throughout the document to illustrate the process:

> **Example**
>
> **Training:** In our example, the LLM would be supplemented (through a knowledge graph) with Gartner research published within the past year as well as research drafts set for publication within the next 60 days. This will allow the LLM to answer questions relating to the recent and upcoming body of Gartner research.

Functionality: The microapp is integrated/embedded as an extension of the word processor used by the research author. As the author drafts their research, the microapp would read each section and use its prebuilt prompt library to ask the LLM for examples of supporting research, examples of contradicting research, as well as supporting statistical data (for detailed prompt examples see Note 1).

Responses would first be verified for accuracy by the microapp (further details on the verification process in the next section under Accuracy > Mitigation) and then provided in the form of suggestions or comments to the associated paragraph in the word processor.

What is an augment? When AI capabilities are deployed (through generative microapps or otherwise) to boost worker capabilities beyond what the average human could achieve, it is referred to as an "augment."

Example

This augment would boost the author's capabilities beyond what is humanly possible. No one person could be aware of every piece of published research but an LLM supplemented with enterprise data can provide that capability.

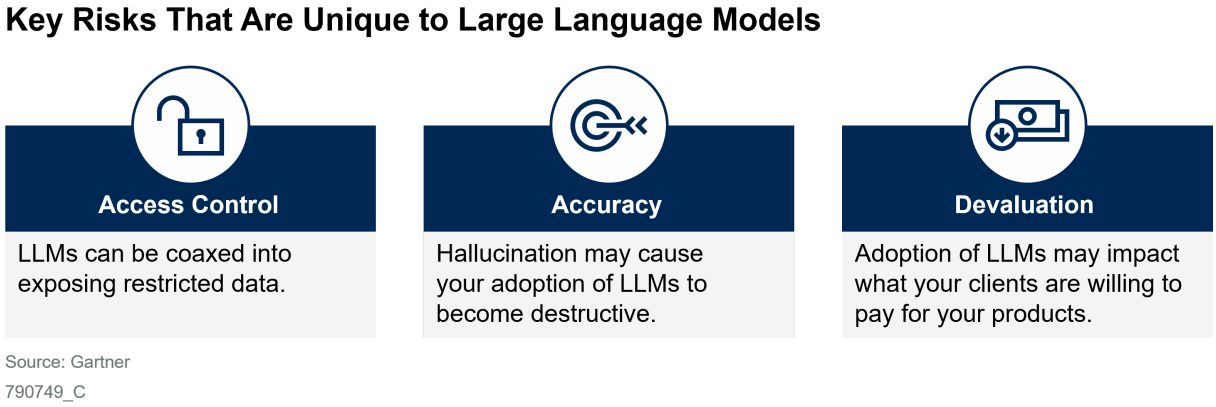
The LLM is not available to the author through a conversational/chat interface (such as ChatGPT or Bard), instead the exchange is governed (proxied) by the microapp integrated within the word processor (as an extension or an add-on). This may seem like we are limiting the models capabilities, but it is "by design" to focus the use case and limit hallucinations. By engineering the prompts in advance and placing them into a library within the microapp, organizations can mitigate critical risks that are unique to large language models.

Analysis

How Do Generative Microapps Mitigate Key Risks Unique to LLMs?

There are many considerations that are common to AI-based initiatives such as governance, bias management and explainability. The following are three considerations that are unique to large language models (see Figure 2).

Figure 2: Key Risks That Are Unique to Large Language Models



Source: Gartner
790749_C

Gartner

Gartner

Access Control

Organizations have come to rely on access control. If an access rule is created, that rule is applied 100% of the time and if it fails, it simply denies access to all. But if you augment the LLM with different types of enterprise data and instruct it as to which data should be available to which users (access rule) through a chat interface, you cannot be sure that your access rules will be followed. Consider the following verbatim from OpenAI "GPT-4 is 82% less likely to respond to requests for disallowed content ... than GPT-3.5." [2]

Considerations:

- Would you invest in an identity and access management platform that you knew could be coaxed into exposing restricted data to unauthorized individuals?
- Would you be confident defending that decision to a regulator investigating the ensuing breach?

Mitigation:

- Generative microapps act as a proxy for the enterprise LLM, they do not allow the user to directly interact with the model though chat, as such they cannot be coerced into exposing restricted data.
- Generative microapps are developed with a specific set of tasks in mind. Different microapps are then deployed to augment users depending on their role in the same way as any enterprise application.

Accuracy

"Hallucinations" has become the term adopted by the generative AI community to describe how models will, from time to time, provide fictitious answers. The issue is not simply that the answers are wrong, it is that they are confident and convincing.

Society has developed an endemic automation bias, humans favor suggestions from automated decision-making systems blindly, often ignoring their own better judgment.

When that bias is paired with hallucinations, the outcome can go one of two ways:

- At best — employees using these platforms may distrust the answers provided and stop using the platform altogether.
- At worst — employees may use these hallucinations to make decisions in the real world that can come with catastrophic results.

Consideration:

Have you conducted projections on the impact of hallucinations in generative models for the use cases under consideration within your organization?

Mitigation:

Through rigorous prompt engineering, the preset prompts embedded in the microapps can limit hallucinations. Furthermore, the microapp can enforce that the answers provided are in a format that the app can validate before passing onto the user.

> **Example**
>
> Following our earlier example, the microapp integrated into the word processor would retrieve supporting references for the author in the form of a document title, a snippet from the document and the document URL (all comma separated). The microapp would look up the URL and confirm the accuracy of the title as well as the snippet's existence before it is passed onto the user. The user can then make a judgment call whether or not to use the references provided by either accepting or rejecting the suggestions provided by the microapp.

Devaluation

Are your clients willing to pay the same amount of money for your services if the work product is provided by one or more LLMs rather than a diverse group of trained and seasoned professionals?

This is not a question of being distrustful of technology, in many instances technology can deliver better and more consistent results than people. This is a question of the perceived value a consumer of a service associates with the price they are willing to pay for said service.

Considerations:

- Individuals are happy to buy mass produced, fast-fashion but they are not willing to pay the same amount as they would for tailored clothing.
- Does your usage of generative AI risk devaluing your business model and eroding your market cap beyond the potential increase in scale.

Mitigation:

Purpose-built microapps are developed to act as augments for your knowledge workers. This will be done to improve the average quality of the work and boost productivity, thereby helping mitigate skills shortages. But since the work is carried out by the same trained and seasoned professionals, the business model is shielded from devaluation risks. Continue to the next section to learn how to tune augments for optimal business outcomes.

How to Align Your Augments With Business Targets and Secure Buy-In?

Defining Augment Business Targets

It is critical to define a series of strategic business targets when developing augments. These targets should not be too prescriptive so as not to restrict innovation. First, to avoid the pitfalls associated with shiny new toys, when considering Generative AI, organizations should cluster business challenges into these four groups:

1. Generative AI has nothing to add here, it is overkill
2. Generative AI can help to boost supervised automation or efficiency
3. Generative AI is needed as there is no other choice

4. Generative AI may help, we just don't know and we're using this opportunity to investigate

Next, organizations should align their projects with one or more strategic business targets such as those outlined in the representative list below:

A. Improve service levels and response times

B. Raise the baseline standard of work from entry-level employees and facilitate recruitment

C. Accelerate the time-to-standard for entry-level employees

D. Reduce/eliminate/automate repetitive tasks

E. Create/support new client engagement models

F. Improve the quality output of a key/common task

G. Raise productivity of high-value employees and alleviate pressure due to skills shortage

> **Example**
>
> Following our earlier example, the microapp integrated into the word processor improves the quality of the research by providing supporting materials and enriching the readers experience, it also helps ensure the overall body of research is consistent. This augment aligns well with both points D and F, as it removes the need of manual reference checking and improves the quality of a key deliverable.

Leadership and Buy-In for Your First Augment

Finally, developing augments for employees cannot be done in a vacuum, it is by definition an interdisciplinary exercise. Leadership should come from human resources (CHRO) and data and analytics (CDAO) with involvement from security and operations to ensure the end product will address their respective concerns.

Securing Employee Buy-In

A critical aspect to consider when developing augments is improving buy-in from the teams where these capabilities will be deployed. Both teams and team managers will not immediately welcome this change, and any introduction of generative AI to the nature of work is likely to face resistance. This is where crowdsourcing in the ideation process is critical so that the teams have an active role in the selection, design and roll-out process of any augments.

Future outlook

General-purpose microapps will become commonplace within the applications we use day-to-day (such as word processors, email and conferencing). Organizations will develop specialized microapps, initially as augments for their high-value employees but it will become commoditized for all knowledge workers within a few years.

A new industry of development houses focused on developing specialized generative microapps, mainly staffed by prompt engineers will grow and thrive.

These augments are the first iteration of the software that will ultimately migrate to human neural implants once that technology is ready

Evidence

[1] Introducing Microsoft 365 Copilot — Your Copilot for Work, Microsoft.

[2] GPT4, OpenAI.

Note 1: Prompt Sequences Prebuilt Into the Microapp

Example

For the earlier example, the prompts would be engineered in advance for best results and embedded into the extension/add-on integrated within the word processor. Following the prior definition, the prompts may look something like this:

- **Persona definition:** Respond as an English language reviewer providing guidance to the document author.
- **Task definition:** You will provide the response in the following format: document titles and document links separated by commas.
- **Setting system-level boundaries:** You will only provide factual responses and you will follow the provided guidance only.

Recursion

- Can you provide references for published or upcoming research that supports this section?
- Can you provide references for statistical surveys that support the assertions made in this section?
- Can you provide references for published or upcoming research that contradicts the assertions made in this section?

Disclaimer: The organization (or organizations) profiled in this research is (or are) provided for illustrative purposes only, and does (or do) not constitute an exhaustive list of examples in this field nor an endorsement by Gartner of the organization or its offerings.

Recommended by the Author

Some documents may not be available as part of your current Gartner subscription.

How Large Language Models and Knowledge Graphs Can Transform Enterprise Search

Quick Answer: What Is GPT-4?

Innovation Insight for Generative AI