

AUGUST 2024

# Buyer's Guide to Enterprise Generative AI Tools

This guide provides agencies who are seeking to procure an internal LLM tool or other enterprise generative AI tool with key issues and questions to consider during the buying process.

**GovAI**  
COALITION



## Introduction

As generative AI tools have become increasingly commercialized, many agencies seek to use enterprise generative AI tools, often internal large language model (LLM) tools, to help their employees work more efficiently. Generative AI tools have evolved since their public debut in 2022 with commercial vendors now offering enterprise-level LLMs tailored for use in enterprise business environments. Currently the marketplace consists of two categories of generative AI with enterprise-level licensing for broad-based, business productivity uses:

- **Stand-alone generative AI assistants.** These assistants use the same foundational LLM and chatbot interfaces available from a commercial provider's other paid-tier or free services, but with enhancements for enterprise use such as single-sign-on, auditing, and encryption. Stand-alone assistants do not have access to enterprise repositories of data. They are typically accessed via your internet browser. Common examples include ChatGPT, Claude, LexisNexis, and Gemini.
- **Integrated generative AI assistants.** These assistants are highly integrated into workplace productivity platforms and are designed to enhance these tools and utilize the organization's internal repositories of data including emails, spreadsheets, PDFs, etc. They are often accessed through an app on your computer. Common examples include M365 Copilot, Adobe AI Assistant, and Zoom AI Meeting Assistant.
- **Large Language Model as a Service.** These tools allow more advanced users to integrate a LLM directly into their applications. They are designed to be fine-tuned on a specific corpus of data and use case domain, which makes them more flexible, customizable, and labor-intensive than stand-alone and integrated generative AI assistants. LLM as a service is typically integrated into an existing project via an Application Programming Interface (API). Common examples include AI services provided by Azure Cloud, Amazon Web Services, and Google Cloud.

When considering the purchase of enterprise generative AI tools, government agencies should carefully evaluate their costs, capabilities, levels of data security, and access controls against the specific needs of their organization. In addition, both stand-alone and integrated generative AI solutions may come with an API to further extend capabilities and data access. Finally, agencies must also evaluate their capacity to adequately oversee and govern these tools for the benefit of their operations and the public good. This Buyer's Guide aims to walk government buyers through many of these considerations during the procurement process.

## What is an LLM?

A large language model (LLM) is a type of AI program that can recognize and generate text, among other tasks.<sup>1</sup> LLMs are trained on huge sets of data — hence the name "large." In simple terms, an LLM is a computer program that has been fed enough examples to be able to recognize and interpret human language or other types of complex data. They output responses by predicting the next word. There are many practical uses of LLMs, including text generation, document retrieval, and code generation.

Many LLM tools are enhanced by the use of Retrieval-Augmented Generation (RAG). RAG is the process of optimizing the output of an LLM by referencing an authoritative knowledge base outside of the LLM's training data.<sup>2</sup> The LLM uses the new knowledge from RAG and its training data to create better responses for users.

## Popular Products

The LLM market is flooded with new products every day. Below are some popular product options:

- Copilot (Microsoft)
- OpenAI (ChatGPT)
- Gemini (Google)
- Claude (Anthropic)

---

<sup>1</sup> <https://www.cloudflare.com/learning/ai/what-is-large-language-model/>

<sup>2</sup> <https://aws.amazon.com/what-is/retrieval-augmented-generation/>

## Potential Use Cases

Internal LLM tools can be applied in a variety of different ways. Below are some common applications:

- Document creation, editing, and summarization
- Information retrieval, especially for new staff\* (for finding information within a SharePoint site or other shared document space)
- Meeting summary and follow-up\* (for a live meeting)
- Email management\*
- Data analysis and visualization
- Task automation\* (from plain-text to automation without any additional steps)
- Software code production and quality checking
- Personalized tutoring in a subject

\*The use case typically requires the LLM to have access to your agency's internal environment or to be embedded into your document space (e.g., Copilot in Microsoft 365 or Gemini in Google Workspace).

## Potential Benefits

Internal LLM tools provide a range of potential benefits for agencies, including:

- **Reduced work friction** – quickly find associated links, transcribe meetings, summarize documents, and answer questions. A common example is transcribing and summarizing notes from a meeting. Then, using the tool to synthesize the notes across 100 meetings to recommend key insights, outstanding questions, and next areas to explore.
- **Shallower learning curves** for new projects, staff, and skills. A common example is understanding on what page in which document relevant pieces of policy, guidelines, code, or procedure are located. Experienced staff may have an intuitive sense for where to go across the organization for information, and internal LLMs can help newer staff do the same.



## Risks

Like all AI systems, internal LLM tools present several risks, including:

- **Information leaks** – depending on the administrative settings, personally identifiable information, performance reports, and financial documents can be seen without needing the link to documents. This concern is greatest with integrated LLMs like Copilot in M365. Imagine a performance report with share settings open to everyone in the organization. Since Copilot inherits document access from the user, the link to the document is no longer necessary for the user to view the document since the AI has inherited access to the document.
- **Inaccurate information** – the tool might index outdated documents or hallucinate information for documents that don't exist. LLMs can be very convincing when stating incorrect information. It is important to fact-check the statements by either independently searching through the document or asking the LLM to cite its sources.
- **Overreliance** – excessive use may condition staff to think less critically or “blame the AI” for ideas or decisions that are ultimately the agency's responsibility. One example is staff using information given by an AI without fact-checking the information first.
- **Environmental impact** – the use of LLM tools requires an enormous amount of energy to power the underlying foundation models (large multipurpose models like ChatGPT). One potential way to reduce energy usage is to fine-tune existing models or to use an existing model instead of training your own foundation model.



## Metrics

When considering the purchase of an internal LLM tool, it is important to evaluate the system's performance across a range of metrics. Since many internal LLM tools are a bundle of multiple services, it may be necessary to utilize different metrics for different functions.

The best way to evaluate the tool is to compare it to a human doing the task without the LLM.

Examples below:

- **Question and answer:** accuracy of statements in the response, relevance of the response to the question.
- **Document summarization:** quality of summary, relevance of summary to the document as determined by the user.
- **Meeting summarization:** accuracy of assigning next steps or action items, correctly recording decisions reached in the meeting.
- **Document retrieval:** relevance and recency of the documents recommended (i.e., providing the documents that an experienced staff-member familiar with the document space would have provided).
- **Memo or policy drafting:** time required to correct or edit the document, adherence to agency's writing standards.
- **Translation:** cultural sensitivity of translation, reading level of translation, ROUGE score.
- **Coding:** accuracy of code generated, ability to find and resolve bugs, speed, and efficiency.

## Pricing

Pricing schemes for internal LLMs will differ by provider, but generally are either “per user” or “consumption-based”.

- A **per user** price model is a flat fee per user of the system, per month. This is common for systems that staff might use for internal work functions, such as Microsoft 365 Copilot (\$30/user per month) and ChatGPT Teams (\$25/user per month).
  - Per user price models may simplify cost calculations but can lead to unused licenses costing your agency unnecessary funds. If you choose an LLM tool with this price model, be mindful of usage rates in your enterprise.
  - Consider requiring staff to undergo a training on how to use the LLM tool before they are given a license.
- A **consumption-based** price model charges by usage of the system. This price model is also known as a “token-based” scheme. For example, you might be charged for each time you prompt the LLM tool with a question, or charged by the number of tokens used in each prompt (for reference, 1 token is approximately 4 English characters, or 0.75 words).<sup>3</sup> This is common for public-facing systems that anyone can use at any time, such as external facing chatbots ([see Denver’s assistant “Sunny” at <https://denvergov.org/home>](#); this guide does not claim the pricing model for Sunny and is only referring to it as an example of an external facing chatbot).
  - Consumption costs can be difficult to predict but can possibly better avoid unnecessary charges.

Sometimes pricing schemes may be a mix of per user and token-based models. Because of the token system, many LLM tools (e.g., chatbots) also have a token or message limit. For instance, even with per user pricing, you may be limited to the number of tokens that can be used per day or per hour.

---

<sup>3</sup> <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>



## Contractual Terms

When purchasing an LLM tool, pay careful attention to the contractual terms of agreement.

Check for:

- **Support** offered to your agency for integration
- **Security** on data you provide to the LLM tool
- **Ownership** of the data you provide to the LLM tool, and ownership of the LLM's outputs
- **Indemnification**, including protection from copyright infringement.
  - It is common for LLM providers such as [Microsoft](#) and [Adobe](#) to protect their users from copyright claims. However, it is also important to consider the limitation of liability, or the maximum dollar amount the vendor will cover per claim. We recommend carefully reading the fine print to make sure you are comfortable with the specific terms.

To address the above concerns, we recommend attaching a version of the Coalition's [Vendor Agreement template](#) to your purchasing agreement.

## Agency Readiness

Before considering an enterprise or other generative AI procurement, public agencies should take stock of their existing governance programs for data, privacy, and cybersecurity. Each of these programs should be leveraged in the procurement and governance of generative AI tools. Agencies should engage the staff of these programs in planning for procurements and the subsequent operation and oversight of generative AI use across the agency to proactively mitigate against risks (as described in preceding sections of this guide).

For education tools and resources, refer to the GovAI Coalition [Adoption Support Committee](#) and the [Education Working group](#). For a more comprehensive set of AI governance tools, refer to the GovAI Coalition [webpage](#).