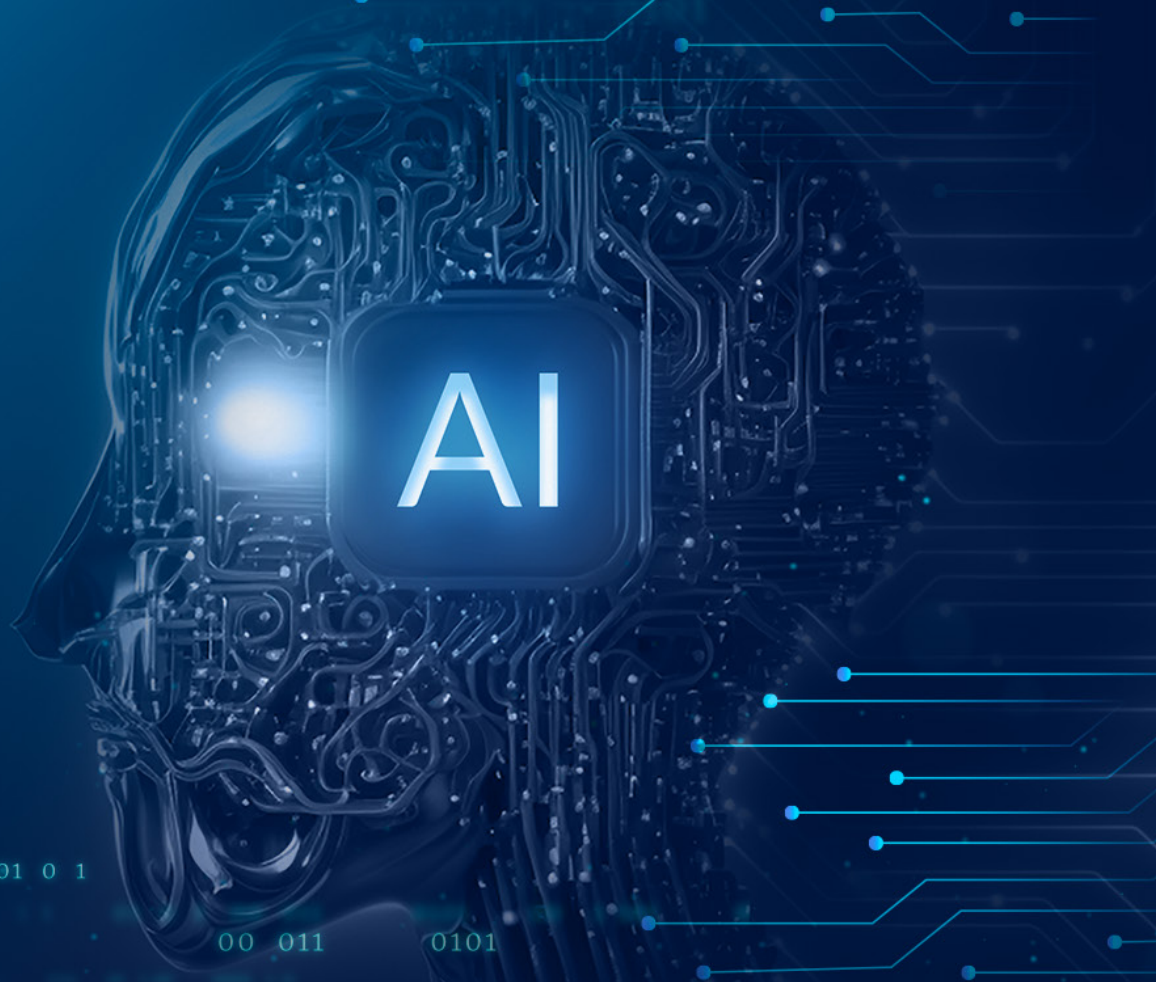
KMS TECHNOLOGY

CTO GUIDE:

Choosing the Right Generative AI Tech Stack for Your Business Data



Written by:
Milan Luketic
Solution Architect, KMS Technology

KMS TECHNOLOGY

TABLE OF CONTENTS

Introduction:
A Glance Into Generative AI Today

> The generative AI market is projected to reach **$110.8 billion USD** by 2030 (Acumen)

“How do we get ChatGPT to work with our data?”

This question has been asked of almost every tech leader in the last 6 months. The answer differs depending on specific situations: data size, regulatory requirements, privacy, budgets, and internal AI resources.

There’s no doubt that the generative AI ecosystem is moving quickly and we’re witnessing innovation demands like never before. Given today’s data-driven business landscape, companies are pursuing stronger pathways to extract valuable data insights and leverage their existing data assets.

This guide aims to be a starting point for tech decision-makers trying to understand which generative AI models fit the needs of their organizational data. KMS Technology has outlined seven best practices for selecting a tech stack, and provided key considerations in your research process. Our experts have selected four use cases to illustrate the different factors that drive technology stack options.

7 Tips to Selecting the Perfect Generative AI Tech Stack

When choosing a generative AI tech stack for your business, below are some steps you can take to ensure an informed decision:

1 Outline your business objectives:

Start by clearly understanding and defining the problems you want to solve or the outcomes your teams want to achieve by leveraging generative AI.

2 Evaluate your existing data:

Assess the nature, quality, and availability of your current business data. Determine whether your data is structured or unstructured and what the level of preprocessing is required.

3 Consider scalability and flexibility:

Confirm that the chosen tech stack can scale as your data grows and adapt to evolving organizational requirements. Take a deeper look into the flexibility of the framework in terms of algorithm customization, model training, and integration with your existing systems.

4 Leverage community support:

Research and engage with active communities, developer forums, and documentation for the generative AI tech stack options you are considering. Strong community support can provide valuable resources, assistance, and opportunities for collaboration.

5 Conduct proof-of-concept (POC) projects:

Execute small-scale POC projects to test and evaluate the performance and capabilities of the shortlisted tech stack options. You can use real or simulated data to interpret effectiveness.

6 Evaluate ethical and legal considerations:

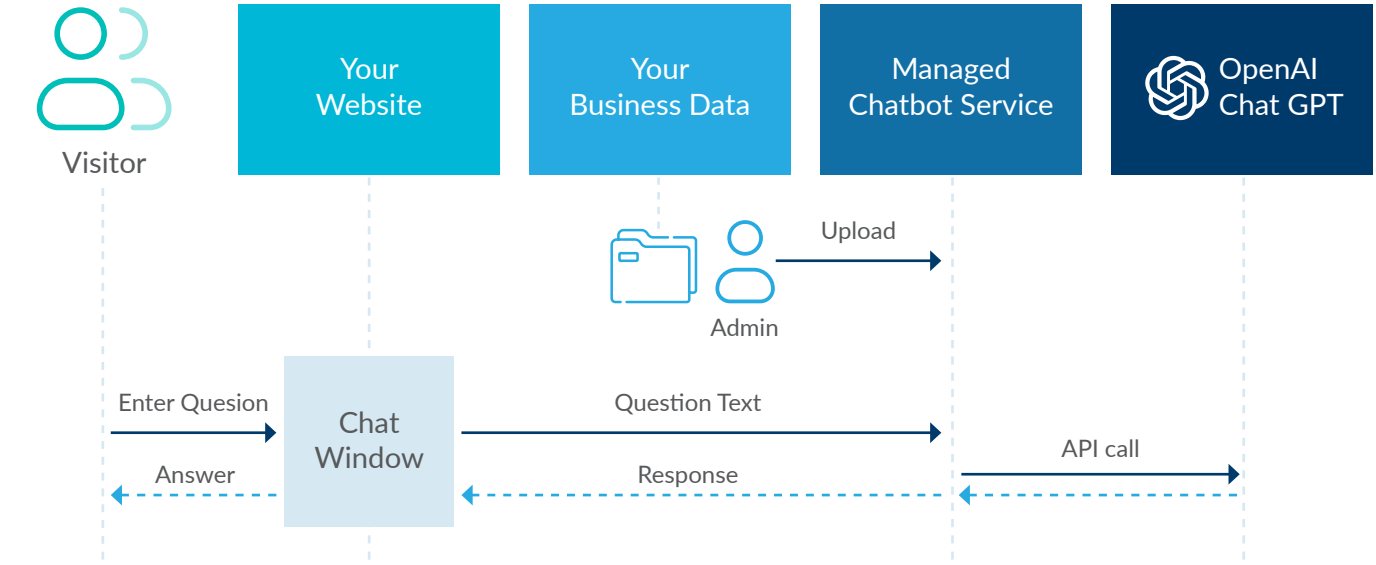
This is one of the most crucial best practices for businesses. Conduct your due diligence to uncover the ethical implications and legal requirements associated with generative AI. Ensure that your chosen tech stack aligns with privacy regulations, data protection laws, and fairness principles.

7 Develop an implementation and deployment plan:

Align teams with the steps, resources, and timelines required to integrate the generative AI tech stack into your existing infrastructure. Consider long-term support and update needs, along with any additional training or expertise required for your team.

USE CASE 1:

Marketing Information Delivered by Website Chatbot



**Managed AI Chatbot Service*

About the technology stack:

This is a managed SaaS solution that provides chatbot code to be embedded into websites and handles communication with the backend server. The backend communicates with the APIs of a cloud-hosted LLM, such as OpenAI's ChatGPT, and your business data is uploaded manually through the managed service UI or with API endpoints. Some popular providers that build on the ChatGPT service are Chatbase.co, Writesonic's Botsonic AI, or CustomGPT.

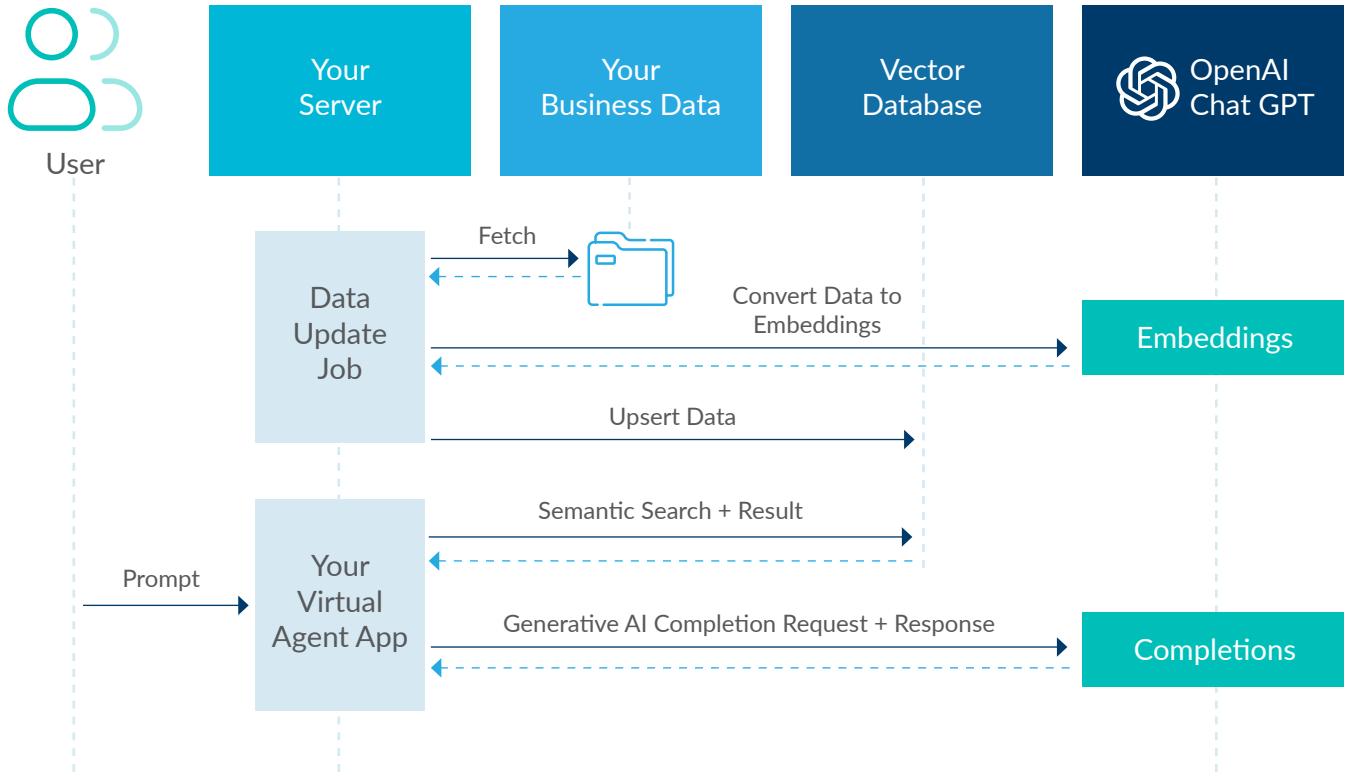
Should you choose this technology stack?

✓ GOOD CHOICE	✕ NOT A GOOD FIT
Internal teams lack AI skills	You have large business data sets
Data is not private	Chatbots are limited to approximately 100MB-200MB of data (many use cases will hit this limit early on)
Organization is moving fast	

This managed service solution saves you from the substantial effort to deploy an end-to-end generative AI application platform from the ground up. The managed service takes care of 90% of the stack while the remaining 10% is embedding the code in your website.

USE CASE 2:

Large Knowledge Base Powering a Virtual Agent App



**Client/Server app with an OpenAI ChatGPT backend*

About the technology stack:

This is a direct integration with ChatGPT and a vector database to handle a large knowledge base—which is a typical tech stack for managed services. The knowledge base data must be converted to a vector format known as embeddings and the vectorized embedding data is stored in vector databases like ChromaDB or Pinecone. Vector databases are becoming increasingly popular due to their role in generative AI tech stacks.

Should you choose this technology stack?

✓ GOOD CHOICE	✕ NOT A GOOD FIT
You have large volumes of frequently updated business data / Data is too large to fit into a single ChatGPT request	You have private business data

Understanding Embeddings

Users can encounter an error when they input too much information during an LLM interaction. For example, if you want ChatGPT to summarize a report for you, the size of that report you can input to ChatGPT is limited. In GPT-3.5, the report can be approximately 3,100 words (4,000 tokens) and in GPT-4 -4 the report can be approximately 25,000 words (32,000 tokens). The amount of data that can be inputted is called input length. Input length is measured in tokens, and a single token may represent a character, sub-word, or word, so there are nearly always more tokens than words.

So, if the size of the input is limited, and the size of your business data is greater than 25,000 words, how can an LLM understand all of your business data? The common approach to solving this problem is splitting the interaction with the LLM into two steps: 1) query your business data to return only information relevant to the interaction, and 2) pass only the relevant information to the LLM. Embeddings assist with step 1 by converting your business data from words to the meaning and context of words, which is a significant improvement over keyword search in returning the most relevant parts of your business data. Then, in step 2, the results of your query are passed into the LLM.

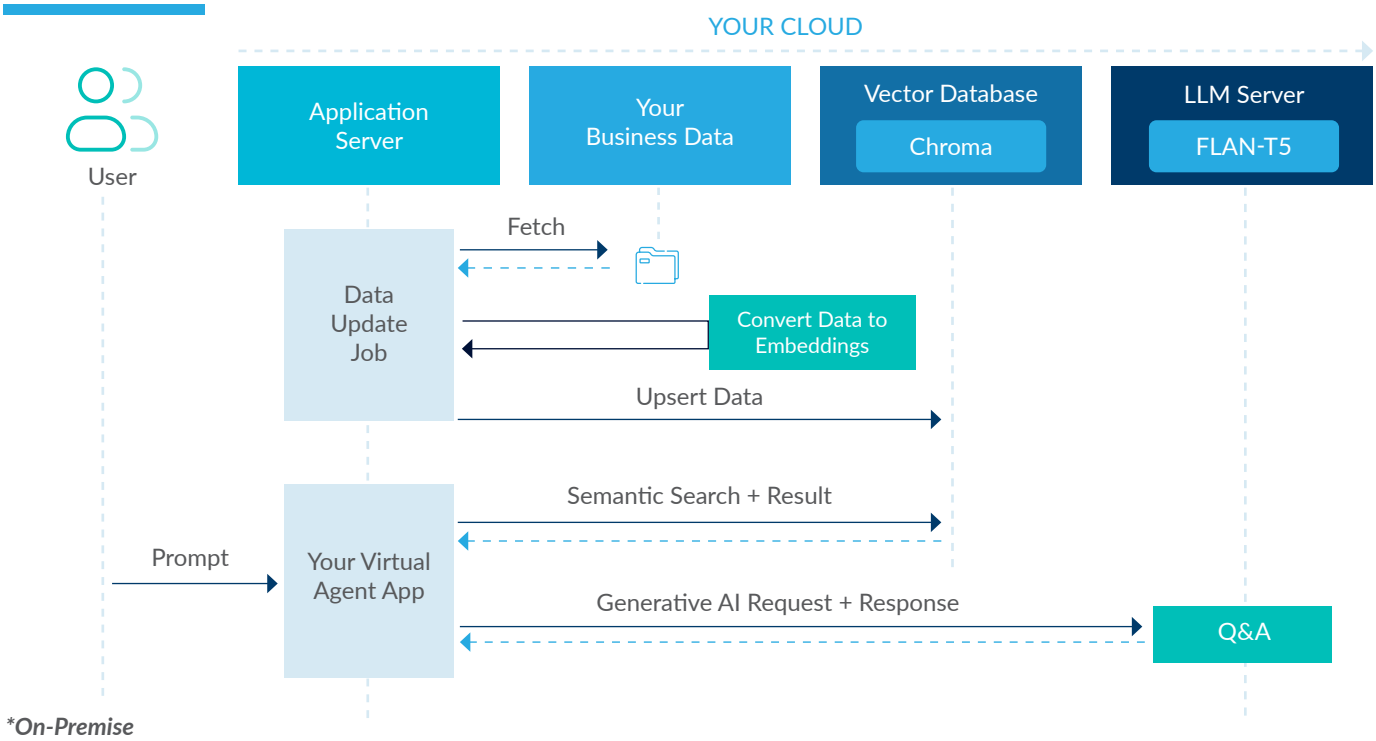
Technically speaking, embeddings are words and sentences translated into tokens and organized into numerical lists, or vectors. For example, the word “dog” may be translated into a list of numbers capturing its meaning and relationships, such as it being a pet, animal, or mammal. This allows the computer to understand similarities and differences between words. For instance, the word “cat” will have a similar numerical list to “dog” since they share features. The word ‘server’ is another example of when an LLM uses embeddings to decide if the meaning is ‘food server’ or ‘hardware device’. After words and sentences are converted into embeddings, they are much more searchable, and results contain words and sentences that are close in meaning.

Businesses might consider this tech stack (in Use Case 2) if their data sets contain more than 3,000 words, implying the data is too large to input as a single prompt. In this case, they should create a subset of the data, less than 3,000 words, by querying the vector store. This subset can then be sent to the LLM for generative AI tasks such as summarization or Q&A.

Keep in mind that this solution is based on OpenAI’s ChatGPT, for which there is limited information about security controls. OpenAI does not share who has access to the data on OpenAI servers.

USE CASE 3:

Private and Restricted Data in a Closed System



About the technology stack:

This stack is designed for serving Generative-AI on-premise. The LLM is Google's FLAN T5 model and is released under the Apache 2.0 license, which is a permissive open-source license that allows for commercial use. The model is capable of various tasks, including text generation, translation, summarization, and question answering. It is not as powerful as ChatGPT 3.5 or 4, but is among the leaders of open-source models. Chroma is an open-source database with a permissive Apache 2.0 license.

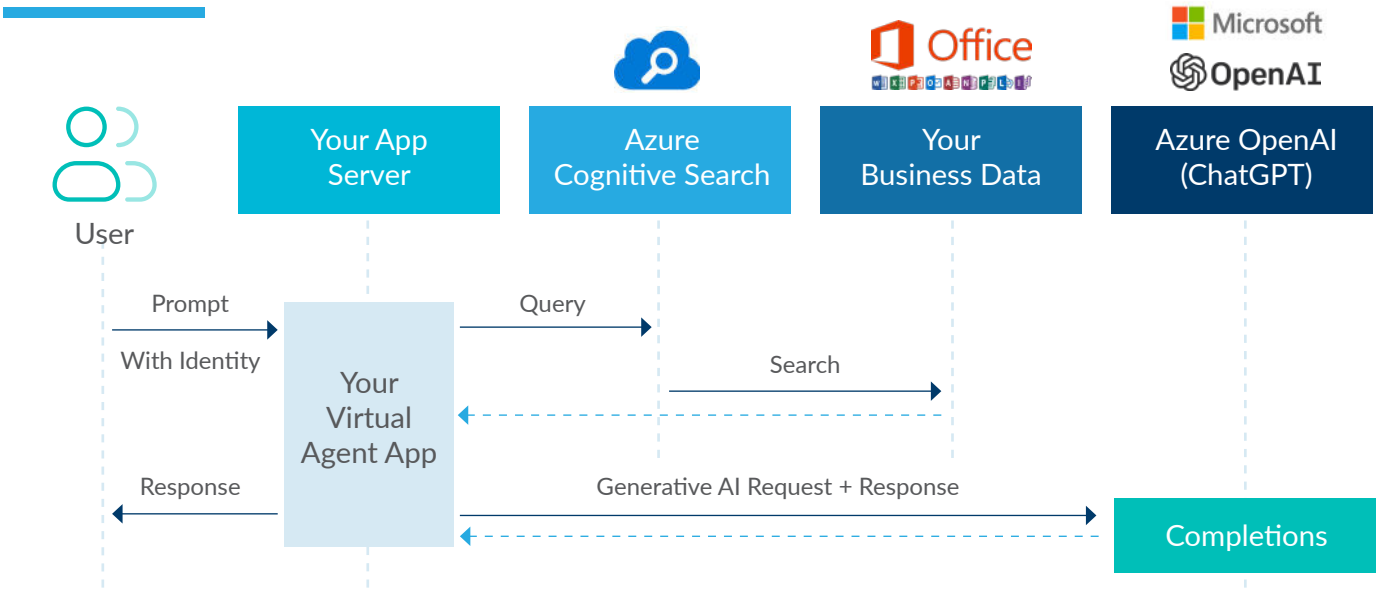
Should you choose this technology stack?

✓ GOOD CHOICE	✕ NOT A GOOD FIT
Security compliance or regulations prevent hosting in a public cloud	The organization lacks the skills or knowledge required to deploy generative AI locally
You have access to staff with AI skills and computing capacity	Generative AI open-source solutions frequently have limited documentation, support, & tooling.

Overall, an on-premise solution can be very cost-effective, especially in a high-traffic environment. However, your teams can expect some level of custom development to wire up the solution for an end-to-end integration with internal systems. It's critical to allocate a significant amount of time to experiment with different models to determine which will provide the best responses for your use case.

USE CASE 4:

Microsoft Files with Permissions



**Azure Hosted*

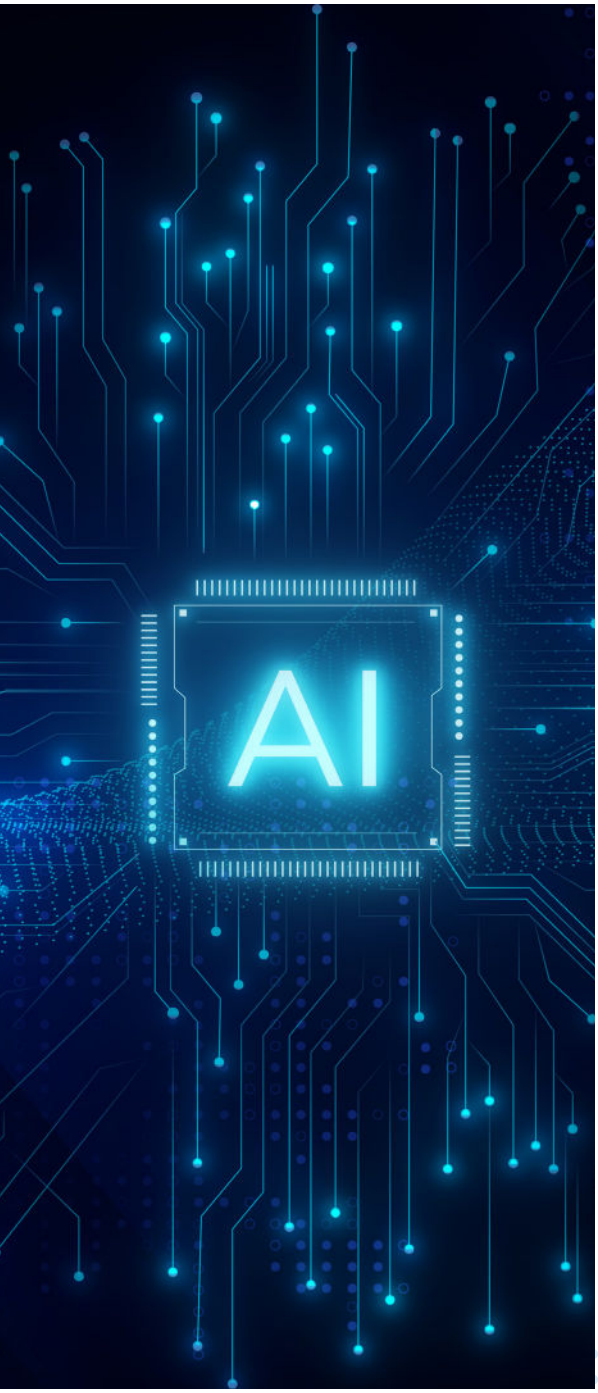
About the technology stack:

This is the security and privacy-compliant version of the OpenAI stack. Microsoft has partnered with OpenAI to provide ChatGPT models on Azure with all of the security and compliance guarantees that come with other Azure services. This solution works with your Office 365 data and respects AD identity and user permissions. In other words, it will not display document-level data to a user unless their AD-level roles and permissions allow it. Azure's Cognitive Search replaces embeddings for searching your business data.

✓ GOOD CHOICE	✕ NOT A GOOD FIT
Highly regulated environment or already in Azure	Limited budget

Microsoft has done a remarkable job integrating OpenAI's models into the Azure ecosystem in a short period of time and continues to add new capabilities. Furthermore, Microsoft has designed the OpenAI integration in a way that avoids ETL-type engineering work and most of the generative AI is available with your Office 365 data out-of-the-box. This is made possible by leveraging Azure's Cognitive Search, which has many AI capabilities already baked in.

However, keep in mind that Azure's OpenAI and surrounding components like Cognitive Search are expensive. The less expensive tiers have rate limits that are quickly reached, and prompts for upgrades follow. Prepare to spend time tuning the solution with a hybrid of open-source and Azure components to reduce costs. At this time, Google and Amazon do not have comparable out-of-the-box solutions and therefore there is no price pressure on Azure.



Conclusion

The landscape of Generative AI technology is evolving rapidly. The market is eagerly awaiting tech giants like Google and AWS to present technologies that can compete with Microsoft's Azure OpenAI. However, many software companies have already integrated Generative AI into their solutions to ensure competitiveness. Tech leaders must be prepared with roadmaps and architectures, should the business decide to invest in Generative AI. This guide serves as a starting point for those leaders in determining the next steps.

Selecting the appropriate generative AI tech stack for your business data is a critical decision. Technology leaders must thoroughly consider the nature and characteristics of their organization's data, including volume, variety, quality, and permissions. When evaluating different generative AI technology stacks, companies should prioritize factors such as cost, internal skill capabilities, and regulatory requirements.

Overall, the decision-making process should involve comprehensive testing and evaluation of potential generative AI tech stacks. POC (proof-of-concept) projects and pilot studies can offer valuable insights into the capabilities, performance, and suitability of various options for your specific use cases. Choosing the best generative AI tech stack for your business data requires a strategic approach, taking into account your unique business needs, data characteristics, ethical considerations, and community support.

By carefully considering these factors and conducting thorough evaluations, you can confidently select a tech stack that enables your organization to unlock the full potential of generative AI.

About KMS Technology

KMS Technology is a leading software services company that specializes in providing innovative and cutting-edge solutions to our clients. Our suite of services includes software development, Salesforce consulting and development, and Generative AI offerings.

Based in Atlanta, GA with award-winning offices across Vietnam and Mexico, our teams are dedicated to delivering high-quality products and technologies that achieve our customer's business and technical outcomes to excel in market. For more information, visit www.kms-technology.com.

Awards & Recognition