vectara

The Executive **GenAI**: **Strategy Guide**

Business Executives are tasked with introducing AI to their organization, LOB's, and products. But the risks associated with introducing nascent technologies like "Generative AI" are high. As such, stakeholders are cautious and caught in a deep and neverending education cycle. This guide was developed to accelerate learning, minimize risk, and accelerate value derived from GenAI. You'll learn how to evaluate GenAI systems (e.g., build vs. buy), what GenAI approaches/architectures exist (RAG vs. Fine-Tuning and when to use each), functional use cases, and most importantly, how to introduce Generative AI into the enterprise wisely, economically, and deliver results quickly.

vectara

Table of Contents

Introduction

Gartner forecasts that by 2030, the majority of individuals, around 80%, will engage with smart robots on a daily basis. This trend is fueled by the rise of Generative AI (GenAI), a pivotal technology that is reshaping business operations, competition, and customer service delivery. Surprisingly, only a small fraction, just 9%, of organizations surveyed by Gartner have established an AI vision statement.

It's becoming increasingly rare for businesses not to have a vision statement or a clear GenAI strategy, especially with the advent of technologies like ChatGPT, which exploded into the business world in November of 2022 and grew to 1 million users in just five days. However, novel technologies are measured by business impact of applicable use cases. According to a recent Gartner webinar poll of 2,500 executives, 38% identified customer experience and retention as the primary focus of their investments in generative AI. This was followed by priorities such as revenue growth (26%), cost optimization (17%), and ensuring business continuity (7%).

What Is GenAI?

GenAI is a type of Artificial Intelligence (AI) that generates various types of content such as text, images, videos, audio, and 3D models. This remarkable ability is created by learning patterns from existing content leveraging advanced machine learning and deep-learning (DL) models. Pivotal models such as GANs (Generative Adversarial Networks), Transformers, and Stable Diffusion have enabled GenAI to produce content that is relevant and accurate to the intention of what the user requested, whether it is creating an image from a prompt (e.g., “create a photorealistic image of a dog with wings in 16:9 aspect ratio”) or a chatbot replying with the exact answer to the question which you were asking.

Types Of GenAI

Each generation type serves different purposes and is being applied to drive real business value in every vertical around the world.

Image Generation

Generates creative, visually appealing, and highly realistic art and images. Successful applications in this type of GenAI are DALL-E and Midjourney, which have become pivotal tools for artists, designers, and marketers for creating visuals. Image generation AI has made generating creative assets more efficient, cost-effective, eye-wateringly beautiful, and opened up new artistic avenues.

Text Generation

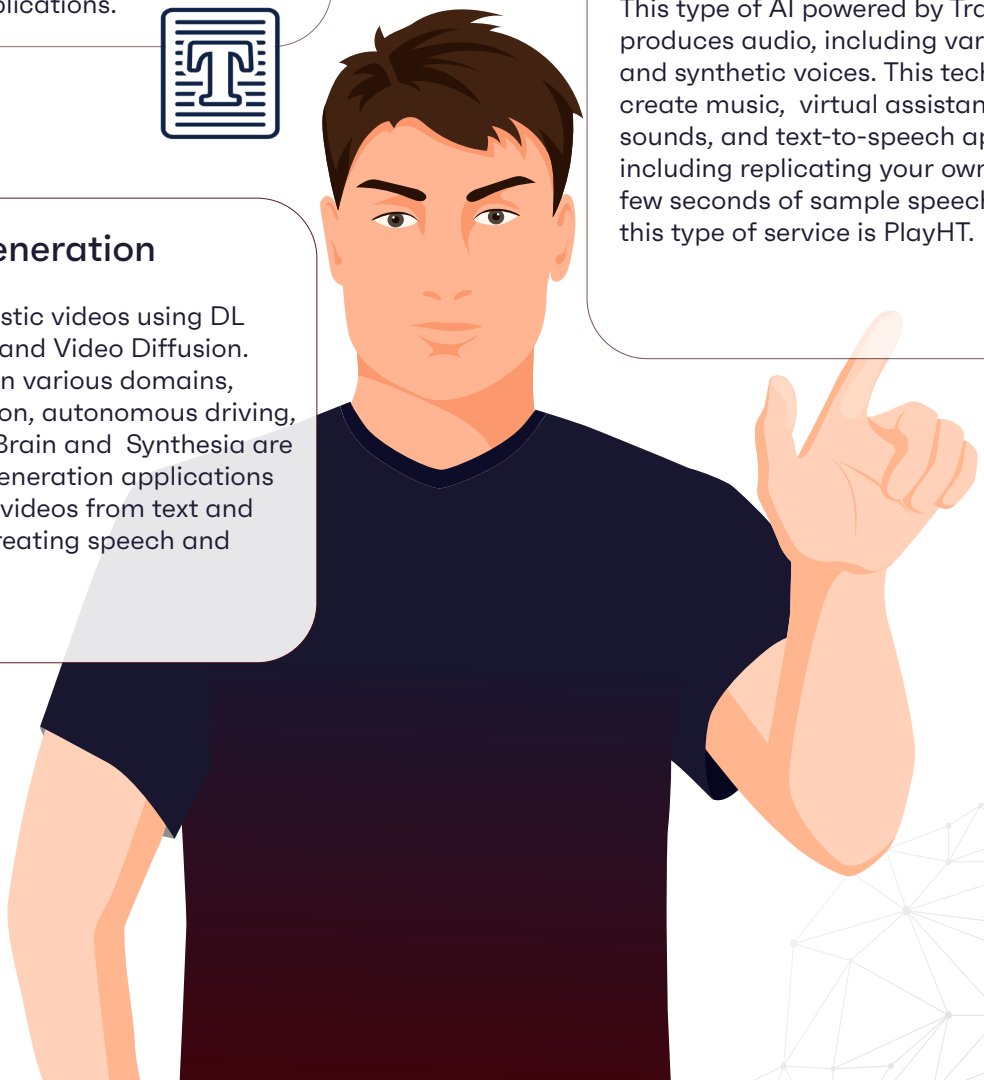
Text generation AI is widely used to create conversational chatbots, generate and debug code, and create new articles and essays. ChatGPT is the most innovative application developed so far using this type of GenAI that has revolutionized how people search for and interact with information. This type of GenAI makes search, knowledge discovery, and summarization easier and more efficient, improving productivity and creativity across various tasks and applications.

Audio Generation

This type of AI powered by Transformers produces audio, including various sound effects and synthetic voices. This technology helps create music, virtual assistants, engineering sounds, and text-to-speech applications, including replicating your own voice with just a few seconds of sample speech. An example of this type of service is PlayHT.

Video Generation

Creating new and realistic videos using DL models such as GANs and Video Diffusion. This type of AI is used in various domains, including film production, autonomous driving, and advertising. DeepBrain and Synthesia are popular GenAI video generation applications capable of generating videos from text and combining audio for creating speech and sounds.



Risks Of GenAI For Business

While GenAI brings many benefits to businesses, it also poses risks. Therefore, investing wisely in GenAI and managing those potential risks is important to eliminate complexities, unnecessary costs, and potential legal and ethical issues. Risks can be categorized into two categories: Business Risk and Technical Risks.

Business Risks:

Adapting to the changing landscape - GenAI can bring forth substantial changes in many industries. The way information is generated and processed will forever change. This will disrupt and change existing market segments, processes, and customer expectations. This also could open new market segments. Failing to adapt to this changing landscape could be fatal to your business.

Intellectual Property (IP) infringement - Copyrighted material can be used to train the models. This leads to the risk of the output containing copyrighted content, leading to legal complications. AI advancement is outpacing legislation, so we'll continue to see major lawsuits in this area for the foreseeable future.

Intellectual Property (IP) leakage - Different from IP infringement, this exposes you to being sued because you used an LLM that someone outputs someone else's copyrighted content. The leakage risk is more about accidental leakage of your (or your customers' sensitive information) into some external LLM or external AI system, the most notable case involving Samsung.

Malicious use of GenAI (e.g., DeepFakes) - Image and video generation AI can create realistic but fake images and videos impersonating actual people. Deep fakes can mislead people and can cause reputational damage, with the greatest avenue being election fraud.

Job displacement - GenAI will alter the job market. It will create ways to access, generate, and process data differently than what's done today. It may generate new workstreams as well as make others obsolete. How a business handles the people aspect of those implications is important. Companies can steer this as human-assist versus human-replace.

Technical Risks:

Hallucinations - Producing meaningless, incorrect, or misleading results. Common reasons for hallucinations include lack of proper training data and fictional content in training data, model complexities, and model overfitting.

Data privacy, ethical, and security issues - Training data may include Personally Identifiable Information (PII). This raises concerns about data breaches and the misuse of sensitive information.

Integration risks - There will be challenges in integrating GenAI into existing systems. For example, compatibility issues, lack of resources and expertise, and additional unplanned requirements. Only 15%-53% of ML models make it into production, and keeping them running can be even more challenging.

Benefits Of GenAI For Business

Despite the risks associated with GenAI, its benefits for businesses are massive. The key benefits of GenAI for businesses are:

Improve productivity and efficiency - A 2023 NATIONAL BUREAU OF ECONOMIC RESEARCH study found that Generative AI at work increased productivity by 14%. This includes the number of issues resolved per hour, improving customer sentiment, reducing requests for managerial intervention, and improving retention rate. One big benefit of the technology, according to the study, is that it helps capture the knowledge and behaviors of the top-performing employees and disseminates that information throughout the organization. GenAI can significantly reduce the time required for creating new content and ideas, and discovering knowledge, saving valuable time, effort, and costs.

Boost creativity - GenAI enhances creativity by generating new and innovative content. It helps businesses to shift their focus from traditional to novel ideas.

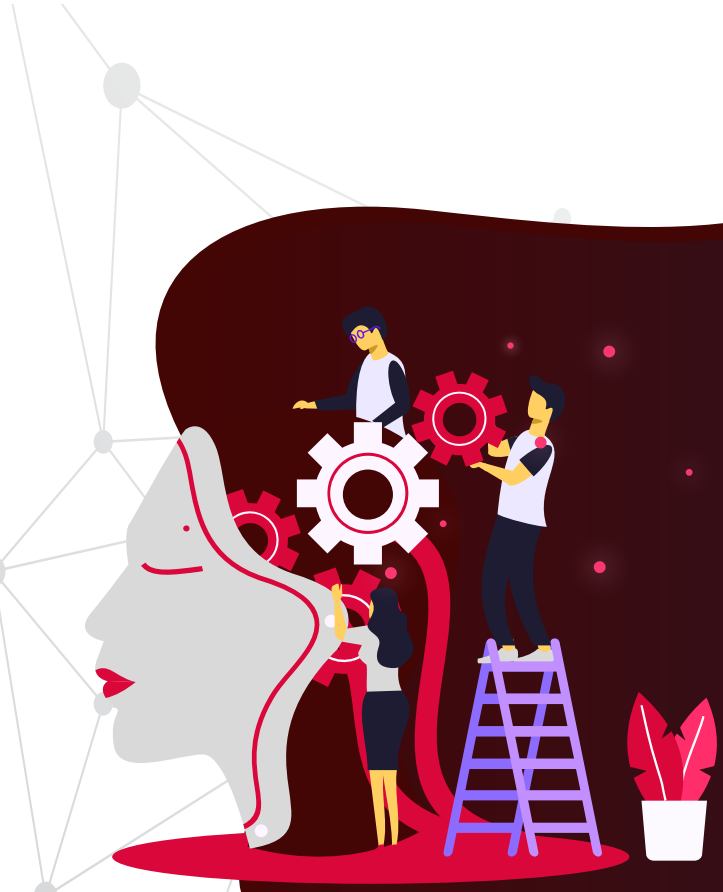
Improve customer experience - GenAI helps businesses to offer exemplary customer experiences by streamlining many customer operations. For example, AI chatbots that assist humans offer faster responses, higher quality answers, and multiply agent's output, all while exceeding customers' expectations. A McKinsey bank case study outlined a bank's initiative to address increasing customer complaints, slow resolution times, rising cost-to-service, and low uptake on self-service channels. The bank decided to revamp existing channels with AI. Customers flocked to self-serve channels resulting in a 40% reduction in service interactions and more than a 20% reduction in cost-to-serve, all of which contribute to improved customer satisfaction.

Hyper personalization - Helps create personalized content and marketing campaigns based on individual preferences, purchase histories, and user behavioral data. This hyper-personalized experience boosts customer satisfaction, trust, and engagement with the business.

Reduces issue resolution time - GenAI applications such as customer support chatbots help customers resolve issues faster. According to McKinsey's The Economic Potential of Generative AI, "research at one company with 5,000 customer service agents found that Generative AI increased issue resolution by 14 percent an hour and reduced the time spent handling an issue by 9 percent".

Knowledge-Oriented GenAI Use Cases

Knowledge-oriented GenAI provides several innovative use cases that revolutionize information search and knowledge discovery. This is the category of GenAI that the rest of this report will focus on.



The most impactful applications of GenAI include:

Question Answering (Q&A)

Rather than searching through a plethora of content to find the right answer for a specific question, GenAI Q&A solutions directly generate the answer. Just ask the question and the right answer will appear in seconds. They are even capable of providing answers to complex queries. Common applications include creating website search functions, product or service information search, educational material search, and workplace search functionalities.

Conversational AI

Enables machines to converse with humans through voice or text messages, leveraging Machine Learning and Natural Language Processing (NLP) techniques. It provides the capability of getting a compelling, relevant response, simulating human-like conversations rather than merely searching through the information. Common applications of conversational AI include customer support self-service virtual assistants, digital learning concierges, and any other system that generates fast answers for professionals swamped with information and demands.

Research and Analysis

GenAI can perform deep analysis from large amounts of data, which helps researchers and analysts use it for research purposes. Whether it is financial investment, legal, pharmaceutical, regulatory compliance, or product development research, GenAI helps users analyze large and diverse datasets so they can more effectively discover what they need, faster, and with greater relevance, including receiving summarized answers.

Semantic App Search

GenAI-based information retrieval systems go beyond using basic keyword approaches, instead leveraging semantic search techniques that produce higher-quality results. Semantic search helps understand the users' intentions regardless of what they ask, producing more accurate and relevant search results. This is used to modernize legacy search functions in existing apps, and to build greenfield apps that take advantage of GenAI's powerful semantic understanding from day one. Semantic App Search is commonly used in developing SaaS products, training and eLearning products, eCommerce, and digital media.

Is Your Company Ready For GenAI?

Having only a strong AI ambition does not mean you are fully prepared to embrace it. Gartner's IT Symposium/Xpo Opening Keynote talk emphasized the three pillars you need to focus on: your organization's GenAI preparedness, AI-ready data, and security of GenAI.

Org Preparedness

Even though many organizations are prepared to invest in GenAI, they lack clear AI vision statements and principles. In fact, according to results from a Gartner survey on organizational preparedness for AI, only 9% of organizations had an AI vision statement.

To be better prepared for GenAI, organizations must establish “Lighthouse principles” that focus on the organization’s values to make better AI-related decisions. Today, people are forming relationships with AI-based systems. Therefore, clear guidelines or principles must be established to define AI’s acceptable and unacceptable use.

Data Preparedness

A Gartner survey reveals that only 4% of CIOs report that their data is AI-ready. Therefore, 96% of them did not meet the criteria. While all your data may not need to be AI-ready, it should be governed by principles.

For example, AI-ready data must include rules and tags to be fed into large language models, so that models can avoid being biased and can suppress copyrighted information. Also, tagging metadata is critical in improving the model’s accuracy. The more governed the data is, the more accurate, enriched, and fair the generated content becomes.

Trust, Security, and Safety

Several security threats are associated with GenAI. For example, exposing sensitive information in the training set, providing misleading information to the AI model, an indirect prompt injection attack that modifies the prompt input by a user, and deliberate generation of content that makes it difficult to distinguish between real and fake information.

Traditional security measures are not adequate to overcome evolving security challenges. Therefore, organizations must invest in models that have emerging security techniques, such as digital watermarking, which reveals the content source, and LLM Grounding (Large Language Model Grounding, AKA Retrieval Augmented Generation or RAG), which reduces inaccurate or inappropriate responses.

Build VS. Buy Strategies: Where Do I Start?

In your business's GenAI journey, the most important step is choosing between the most suitable strategy: building vs. buying a GenAI solution. However, given the major LLM foundation model builders, who can add a feature that in one swipe renders smaller point solution builders' business models meaningless, it is important to not only evaluate the tool, but the underlying broader framework and synergies with these LLM builders (GCP, Azure/OpenAI, AWS). Some organizations build customized AI solutions from scratch using open-source tools like Langchain or Pinecone.io, while others leverage pre-built GenAI solutions from vendors like Vectara and others including cloud providers AWS, Microsoft Azure, and GCP.

Additionally, these "Big Cloud" vendors are creating robust, composable frameworks/workflows and extending partner solutions via cloud marketplaces and managed services that allow "plug-and-play" capabilities of individual vendor components, such as a retrieval model from one vendor and a generative summarization model from another vendor for building GenAI enabled applications.

In all cases, your decision highly influences your applications' success and significantly influences your business's competitive standing in the market. Therefore, businesses must carefully consider the pros and cons when deciding what strategy best suits them to achieve their business goals.

Build

Building can be done in several ways: Use proprietary models as a base that you Fine-Tune to extend with your own data, Use Open-Access models (e.g., LLaMA from Meta AI, Stable-Diffusion from Stability AI), Use Open-Source Models, or train your own model from scratch.

What's really required to build your own GenAI system?

Building your own complete GenAI solution requires building and operating both the underlying platform infrastructure and your actual business applications, not to mention possibly building a team of AI/ML/DS/MLE experts. Put another way, you build and manage the plumbing/electricity/framing/heating for the house (i.e. the platform infrastructure)... and you then design the rooms (i.e. your business applications).

The Process

Several tasks and components are involved with building the infrastructure:

ETL - Extract, Transform, and Load (ETL) processes that include input data parsing and chunking, and which require significant technical expertise and resources to implement.

Retrieval LLM - Using a Retrieval Large Language Model (LLM) (a.k.a Embeddings Model) such as OpenAI Ada or Cohere Embed requires the selection of the best model for your data set and integration code. This is the most important component of a RAG system (discussed below).

Vector data storage - GenAI systems require fast data storage and retrieval capabilities. Thus, a vector storage solution such as Pinecone or Qdrant must be implemented, adding yet another complicated resource to manage.

Other data storage - Other data storage systems like MariaDB or Postgres are required to store the entire amount of data required for a complete GenAI application. This includes metadata, structured data fields, and raw text, and doing it properly requires specialized expertise and adds complexity to the end-to-end platform.

Generative LLM - Models such as OpenAI GPT or Meta's Llama2 are required for content generation.

Integration of platform components - Code is required to orchestrate the entire pipeline and integrate all the components. This can be done via open-source frameworks, such as LangChain, or via custom code. In either approach, attention must be paid to ensuring high quality of service, low end-to-end latency, and comprehensive monitoring. This is generally where organizations abandon this approach; discovering how difficult this path is, either through overestimating open-source capabilities or underestimating the effort required to keep this highly technical system operational.

Platform maintenance, enhancement and support costs (ongoing) - After building the solution, the organization is responsible for ongoing maintenance, platform enhancement to add new capabilities that the business will inevitably require, and costs to support the end users. This usually requires a dedicated technical team.

Security and legal/regulatory compliance costs - Security and legal/regulatory compliance costs are essential to implement robust security measures to protect the business and its clients.

The Benefits

What makes building so hard to initially resist:

Greater flexibility and customizability - Building a GenAI platform yourself will likely provide complete control over its functionality and features, offering greater flexibility and customizability. These capabilities enable you to tailor your solution to your organization's specific business requirements.

The Drawbacks

Why do Builders most often have buyer's remorse:

Higher Total Cost of Ownership (TCO) and risk - Building your own GenAI platform usually results in higher TCO than buying an equivalent existing one. The development, modification, update, and ongoing maintenance costs that are incurred result in higher risks, such as project delays, technical difficulties, privacy and security risks, and integration risks.

Longer time to value - Several tasks are involved in building your own GenAI platform solution. Thus, it will take a longer time to get your business applications released to your users, compared to purchasing a pre-built platform. Therefore, getting the value for your investments will take longer.

Harder to staff the projects - Integrating complex AI models and specialized storage and compute components requires people with specialized knowledge in those areas. Such expertise comes at a great cost. Data scientists (1.7-year average tenure) and ML Engineers are in high turnover, so expect to be in a constant hiring cycle, especially as the GenAI space shows no sign of cooling off in the near term.

Buy

Buying a solution like Vectara & others (Big Cloud; AWS, Azure, GCP) has many benefits that might be worth considering, including faster Time To Value (TTV), lower TCO over the lifetime of the project, reduced expertise and team hiring requirements, faster realization of new features and updates, and lower maintenance costs. Negatives might include a higher upfront cost but should be weighed against near-term business benefits and lifetime cost/time savings.

What's required when buying a GenAI system?

Buying a complete GenAI solution offers many benefits that may not be expected beyond head-to-head comparisons. For example, you may think that you only get the benefits of composability (swapping out or selecting bespoke components) if you DIY or build your own, however, platforms may offer the ability to select the best LLM for your specific use case or vertical (e.g., a model trained for the Legal industry).

The Process

Several tasks and components are involved with building the infrastructure:

Connect your data stores - Connecting your data to the platform of choice kicks off the process to make it useful in your GenAI application. No need to research vector databases, embeddings models, retrieval models, or LLMs.

Test your user queries - Easily validate whether the solution performs well on the types of requests that will be made by your users.

Build your app - Now that your data is indexed/embedded, simply build your app and connect it to the GenAI platform via simple APIs. That's it.

Path to production - Due to the holistic nature of the GenAI solution, where the platform infrastructure is typically already included, the path to get to production is typically very short, requiring only DevOps config changes.

The Benefits

What makes buying the more viable, durable, and economical choice:

The vendor provides the underlying platform infrastructure- if you use a pre-built solution from a vendor, you don't have to worry about building and maintaining physical infrastructure or resources, as the vendor provides all necessary platform infrastructure. Therefore, you only have to focus on building and deploying your business applications.

To continue the earlier metaphor, in this case, a third party provides you with a prefabricated house, which they maintain and upgrade over time... and all you have to do is design the rooms.

Lower TCO and risks - Using pre-built GenAI solutions typically result in a lower TCO than using your own GenAI solution, since there are no platform development costs, and the vendor handles a large portion of operational and maintenance costs. Also, it reduces risks because providers handle most infrastructure security, maintenance, and uptime guarantees.

Receive ongoing maintenance, upgrades, and technical support - GenAI solution vendors will provide ongoing updates and improvements such as security, bug fixes, and feature enhancements. Also, they often come with dedicated ongoing technical support in case of any issues or incidents related to the solution. In the rapidly evolving AI space, keeping up with these on your own could be nearly impossible.

Shorter time to value - Since the vendor already provides the bulk of what is required for your solution, you simply have to focus on implementing your own business requirements to complete the picture. This lets you bring your GenAI applications to market faster, so you can realize the value sooner.

Easier to staff the projects - Vendors of ready-made GenAI solutions have experts in those platforms. Therefore, it is generally easier to staff these projects, as the skillset required is lower. When your internal experts have questions, where do you go if you build? Buying from an expert partner, you always have someone there to help you.

The Drawbacks

Why do Buyers most often have buyer's remorse

Less flexibility and customizability - Buying a pre-built GenAI solution is less flexible and customizable than building your own end-to-end solution from scratch. Although it will satisfy many of your requirements, modifying the solution or implementing your specialized requirements might be harder. Therefore, there are situations where you might need to adapt your applications to fit the constraints of the chosen solution. However, smaller, more nimble startups will integrate feature requests into the product roadmap sooner, possibly rendering this negligible.

As emphasized above, building a GenAI solution requires significant time, effort, costs, and resources. Buying can significantly reduce this burden, offering a more efficient path to business value.

Build vs. Buy
Head-to-Head Comparison

Building a GenAI solution is suitable in some scenarios, such as when:

- No commercial product is available.
- An in-house AI solution's success heavily depends on your staff's specialized knowledge and expertise. Building your AI solutions makes sense if your organization can attract such world-class domain talent.
- There is a strong business momentum that can retain the top talent for continuous innovation and development in their AI initiatives.

Even in those scenarios, significant challenges exist with DIY approaches, such as:

- From the design of the solution, DIY approaches take more time to launch into a fully functional product.
- The selected solution approach must be properly validated to ensure feasibility, reduce risks, and align with the business goals.
- The domain experts may be talented enough to build such complex systems. However, they are often not much inclined towards documentation and maintenance. Therefore, proper attention must be given to these important, yet tedious tasks.

Even if the organization has full control and flexibility with in-house solutions, such solutions can become costly to maintain and adapt to evolving technologies, causing a 'self-lock-in' effect. If a need arises to change from building to buying a solution, it can be challenging with organizational and political situations until there is a change in leadership. Therefore, it is important for organizations to carefully consider the build vs. buy decision.

Total Cost Of Ownership (TCO) Analysis

Buyers and builders alike are always interested in an apple-to-apple comparison for each of these GenAI solution strategies, so we've built a TCO calculator that shows a simplified comparison for an app that has 1GB of data and 100k queries per month. While there are many, many factors and every organization is unique, this encompasses the most common configuration that we have encountered and should provide a viable starting point for most use cases. Notice that this does not include human capital costs such as building and staffing a DS/ML team to support the Build option and the smaller ML Engineer IC for the Buy option.

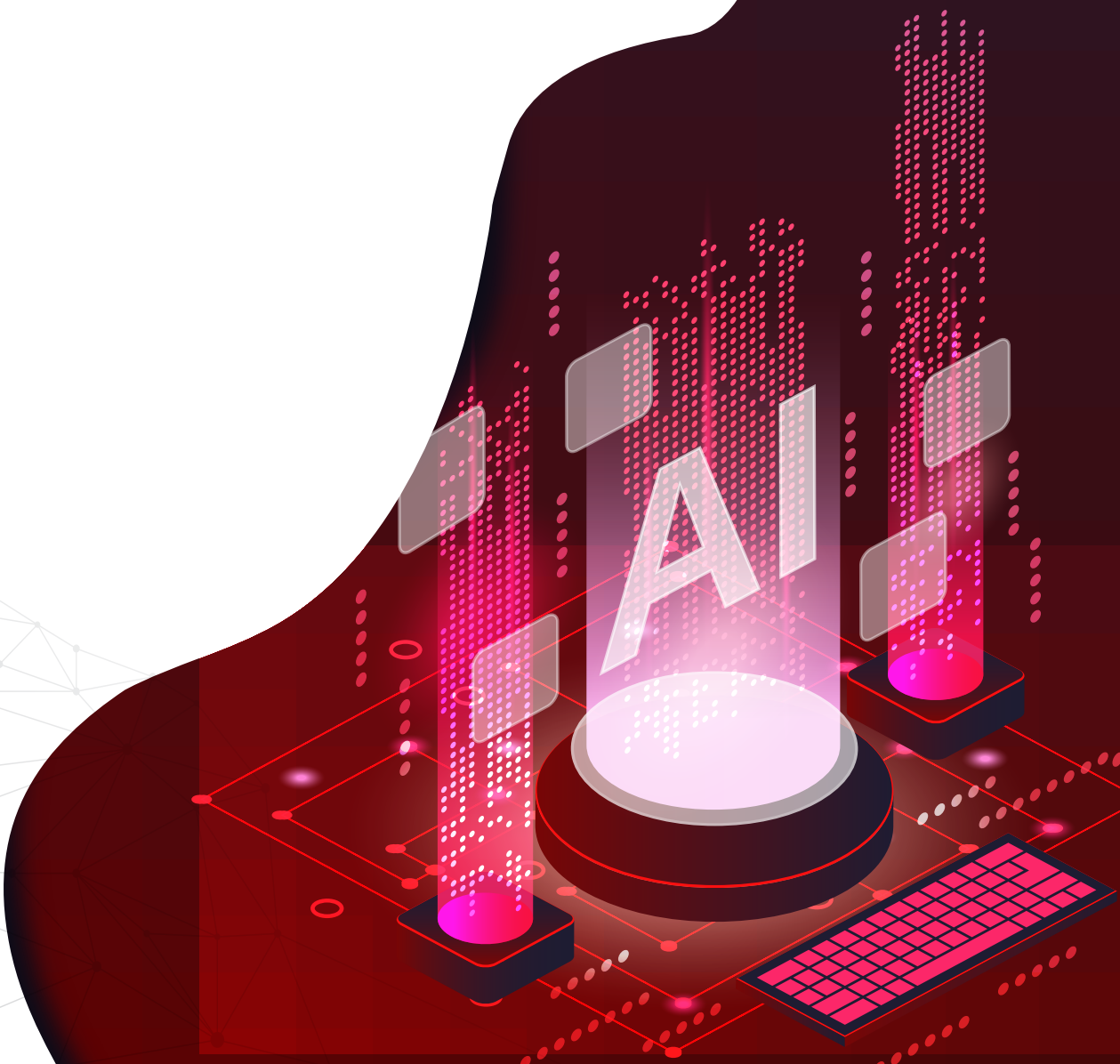
Less Flexibility and Customizability (Notes: 1GB of text data, 100k chat messages/mo.)	Build Year 1 Cost	Buy Year 1 Cost
Platform Implementation Costs	$61,538	$0
Business App Implementation Costs	$11,538	$11,538
Hardware, Software, API calls, Subscriptions	$117,302	$19,826
Annual Platform Maintenance, Enhancement, and Support Costs	$40,000	$0
Security, Legal, Regulatory Compliance Costs	$17,500	$0
Total	$247,878	$31,364

Table 1: Representative GenAI Solution TCO Comparison for Build vs Buy Options

Assumptions: Business application is a ChatGPT-style system for customers to more easily learn about a company's products and address their own issues. 1 GB of text data in the knowledge base. 100,000 chat messages per month.

GenAI Approaches/Architectures

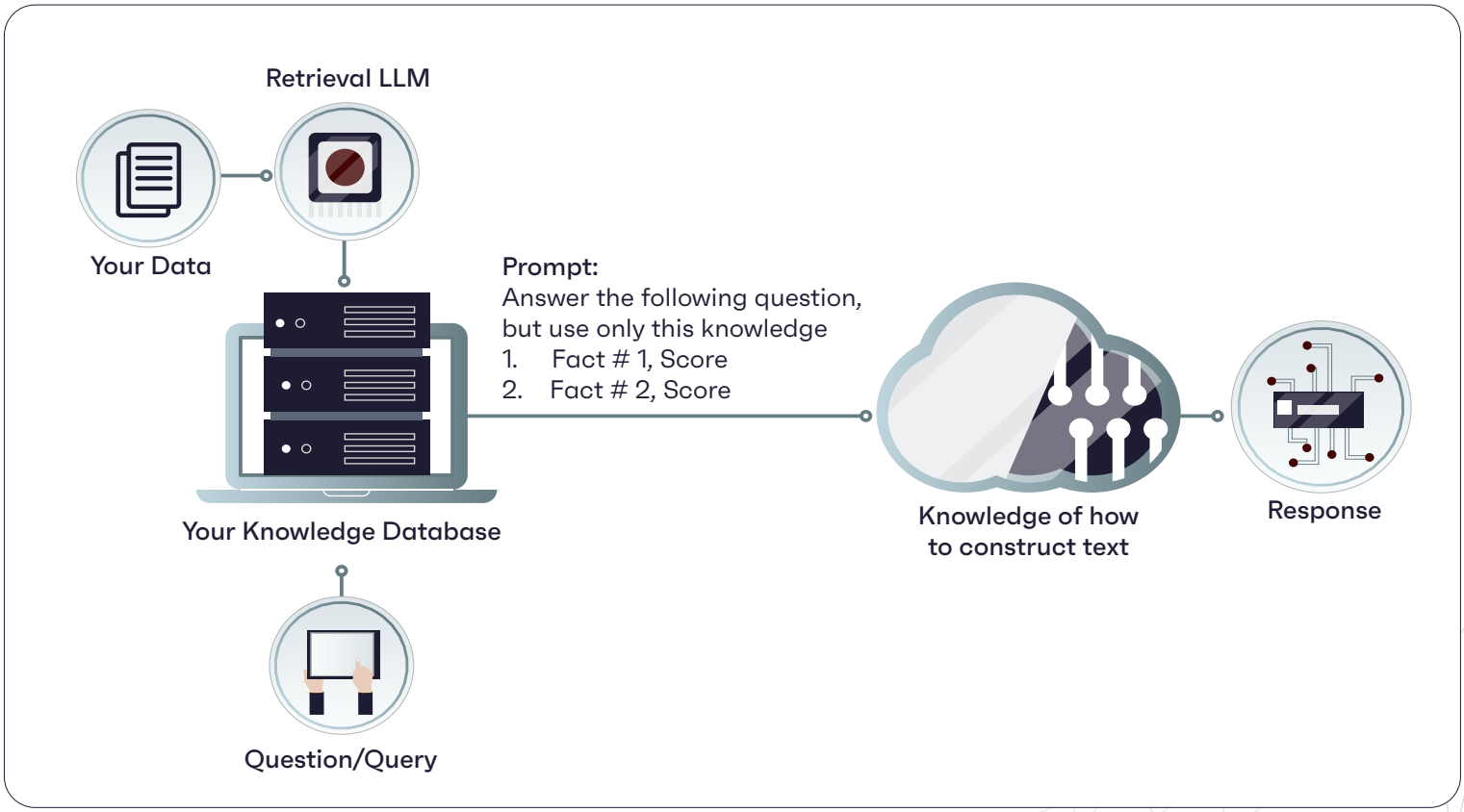
Performance and accuracy are crucial to successfully adapting the LLM models for your GenAI initiatives. Generative approaches alone won't produce the best possible content due to their limitations, including lack of domain expertise and accuracy and relevancy of the generated content. Luckily, advances in GenAI have introduced several architectures to overcome those challenges. Retrieval Augmented Generation (RAG) and Fine-Tuning are the two prominent approaches for improving the outcomes of GenAI solutions.



RAG (Retrieval Augmented Generation)

A hybrid approach for content generation that combines the capabilities of pre-trained generative AI models with retrieval techniques. First, the retrieval model takes a specific input query and searches and retrieves relevant information from a large knowledge base, which has been defined by the application owner and includes data external to what was used in the generative LLM's training set (e.g., a company's knowledge base and product support docs for a customer service chatbot application). Then, the generative LLM uses the retrieved information to generate content, such as a summarized answer or a chatbot response.

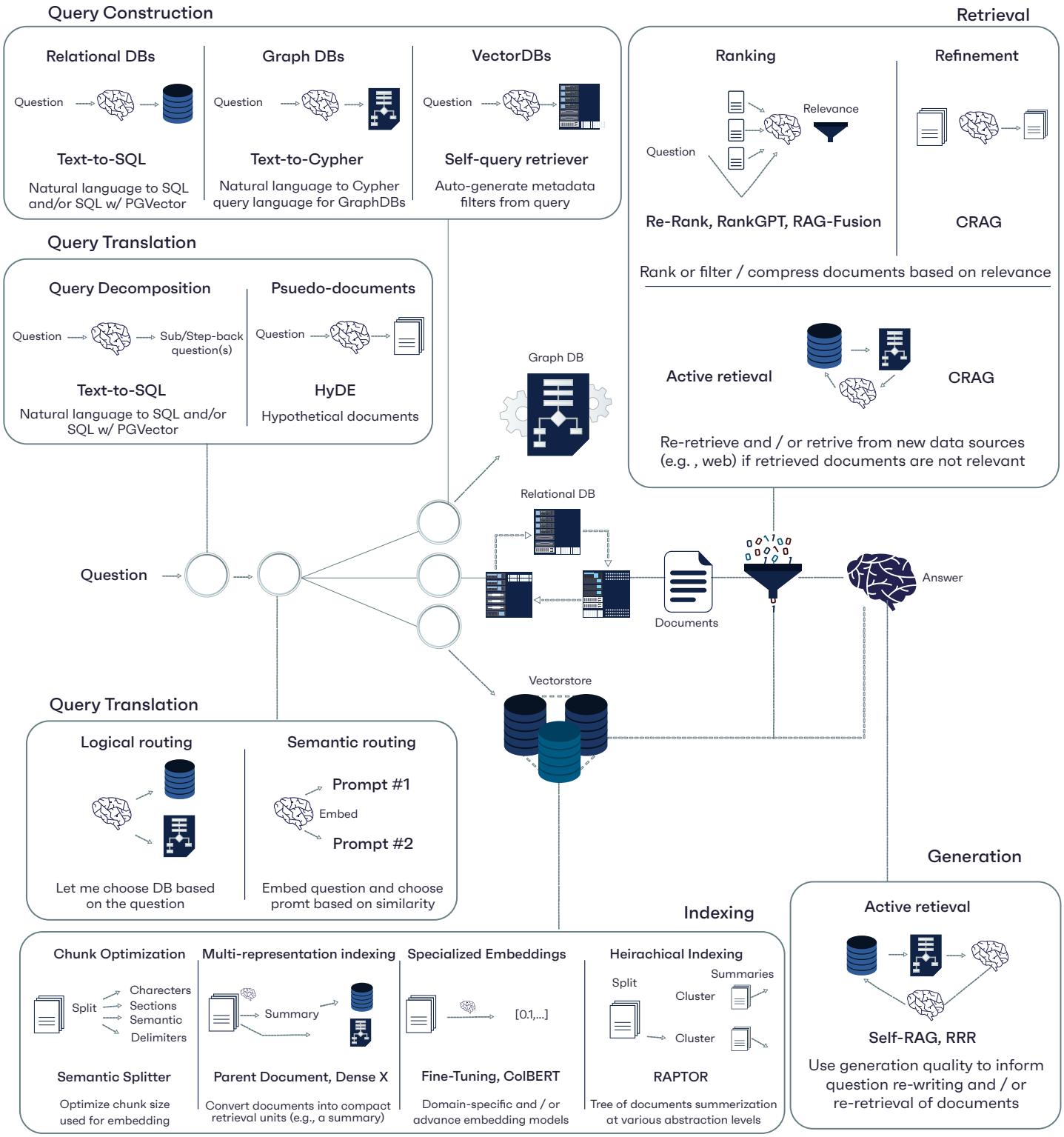
Vectara's Retrieval Augmented Generation How It Works



Now that we've given the executive a high-level overview of RAG, let's dig a bit deeper and look under the hood to explore what all is required to build such a system, first, for a developer looking to DIY or build it all on their own stitching open-source solutions together. We won't go into much detail here as this is beyond the scope of this executive guide.

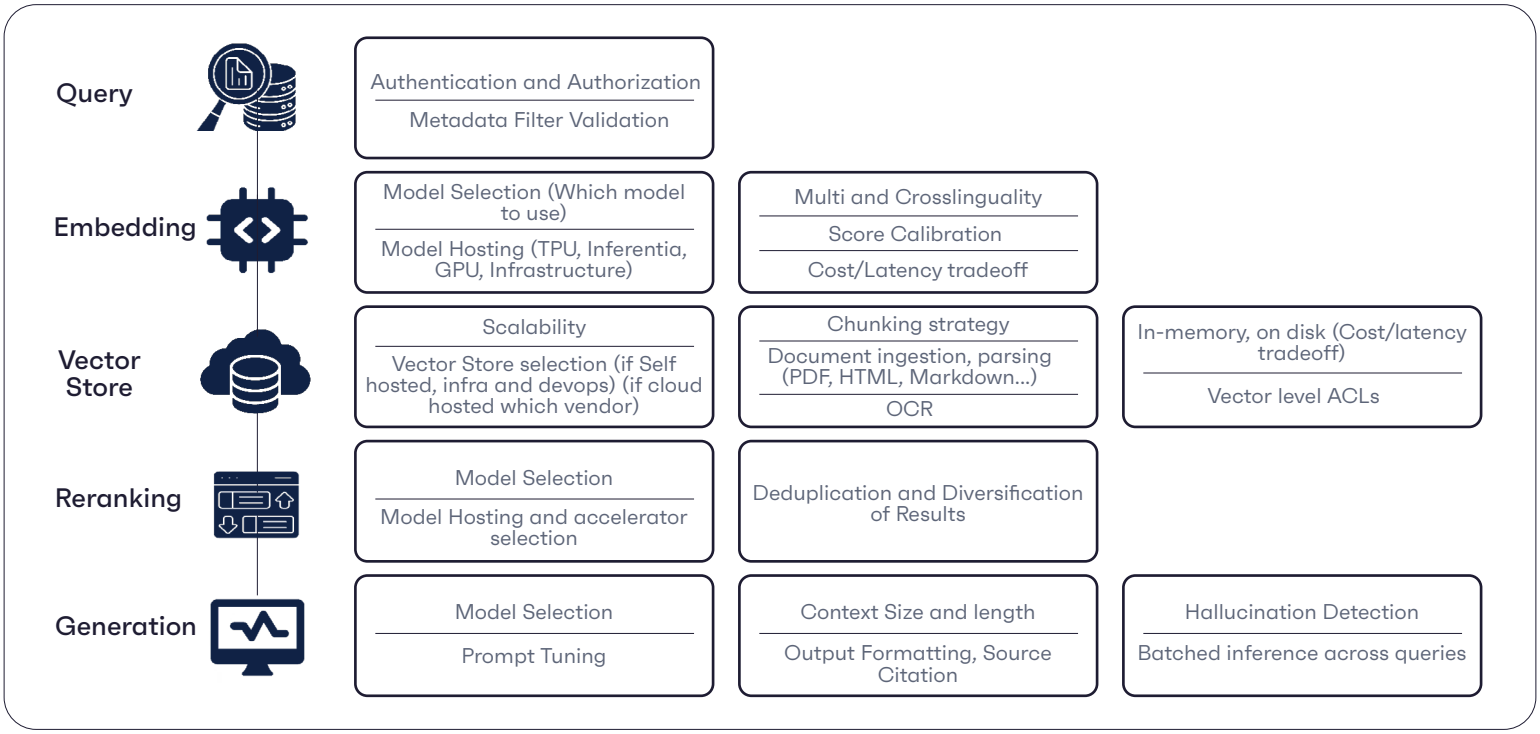
RAG From The Developer's Perspective

As you can see from the complexity in Figure 3., building a RAG system is complicated, has many possible failure points, and should not be taken on lightly ("free" always has a cost). Many who start on this journey encounter issues when pushing to Production, and re-evaluate or abandon this approach altogether.



RAG With Vectara

Now let's take a look at a RAG system built on Vectara. First, you'll notice the absence of complexity as Vectara offloads much of the technical overhead from the developer's workload by encapsulating and obfuscating many of the key components (embeddings model, vector database, retrieval, summarization), allowing the developer to focus on building their application. But don't interpret simplicity to mean a lack of features, as Vectara offers many features offered by open-source, and additional enterprise features that open-source lacks (MMR, User Access Controls, APIs for API Management , etc.)



Advantages of RAG

- Improves the accuracy of generated content.
- Access to information drawn from only the data sources specified by the user allows the generation of content that is more relevant to the search query.
- Can adapt to changing information by incorporating the latest data sources. Therefore, particularly useful in domains that have dynamically updating data.
- Response sources can be easily tracked. Therefore, it increases the transparency of data sources and reduces the risks of biases and hallucinations.

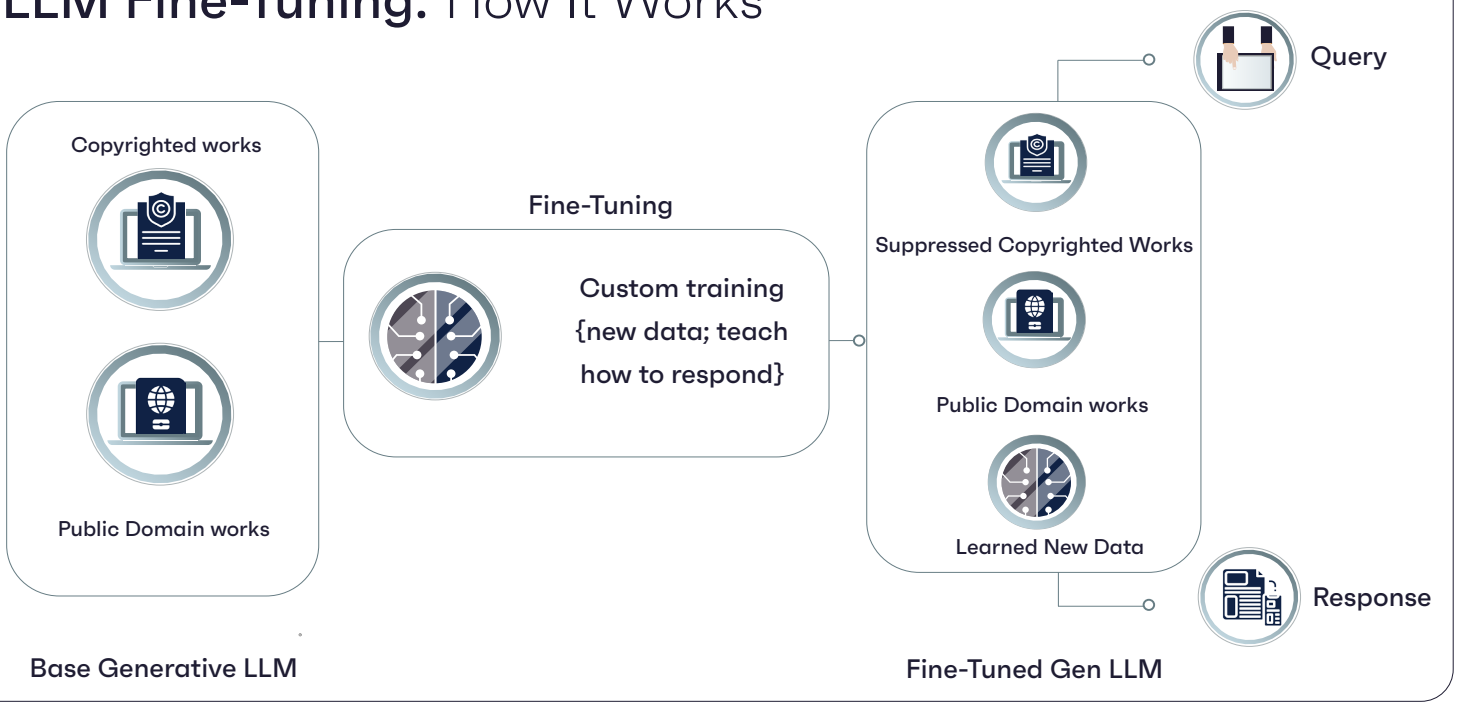
Disadvantages of RAG

- Highly dependent on the quality and scope of the retrieval database.
- Additional architectural complexity due to the need to run the knowledge database and retrieval engine, as well as the generative LLM.

Fine-Tuning

This approach starts with a pre-trained generative LLM. It improves the LLM's performance by further training it on a labeled dataset of a particular domain, tailoring it to perform specialized tasks. During the training phase, the model will be iteratively trained until it achieves the highest accuracy, adjusting its weights, parameters, and layers.

LLM Fine-Tuning: How It Works



Advantages of Fine-Tuning

- The model becomes highly specialized in generating content for a particular domain.
- Generate more accurate content than an untrained generative LLM.
- Requires a smaller amount of training data than training a generative model from scratch.
- Architectural simplicity, requiring only the generative LLM calls to perform the required tasks.

Disadvantages of Fine-Tuning

- Costly and slow.
- Hallucinations.
- Hard to explain the generated output.
- Can't enforce per-person access control on generated output.
- Hard to delete/update/maintain knowledge stored in LLM.
- Hard to suppress copyrighted information from being generated.
- Hard to minimize inherent bias from training.
- Hard to update/maintain the base generative LLM (will require fine-tuning).
- The weights in the LLM "know" your IP (Model Pollution).

RAG vs. Fine-Tuning

There are several dimensions against which to evaluate whether RAG or Fine-Tuning is the best approach for your application:

Type of content - RAG outperforms Fine-Tuning in handling dynamic datasets because it can easily incorporate datasets that change over time. No frequent re-training is needed for RAG, but it is still able to generate content using the most up-to-date information. In contrast, Fine-Tuning is based on static data, which can quickly become outdated in dynamic applications. To overcome this a Fine-Tuned LLM will accept a small amount of additional data to be passed in at run time as "context", providing new information to the LLM but increasing latency and cost.

Accuracy of enerated content - RAG provides very accurate results since the content it generates is grounded on the information contained within the data sources specified by the user. Fine-Tuning also can provide accurate results because its domain-specific understanding is high.

Hallucinations - By restricting the generated output to only be derived from the information contained within the data specified by the user, RAG is more capable of reducing hallucinations than Fine-Tuned models.

Costs - Fine-Tuning consumes a significant amount of computational resources compared to RAG. Re-training deep learning models requires powerful compute, such as Graphic Processing Units (GPU), which are costly. Any new information or updating generally requires re-training the Fine-Tuned model, which is time-consuming and costly.

Transparency (explainability) - RAG provides more transparency than Fine-Tuning because, in the RAG approach, the generated content can refer back to the specific data sources from which the content was drawn, called "citations," allowing users to track the source and veracity of the information.

Total Cost Of Ownership (TCO) Analysis

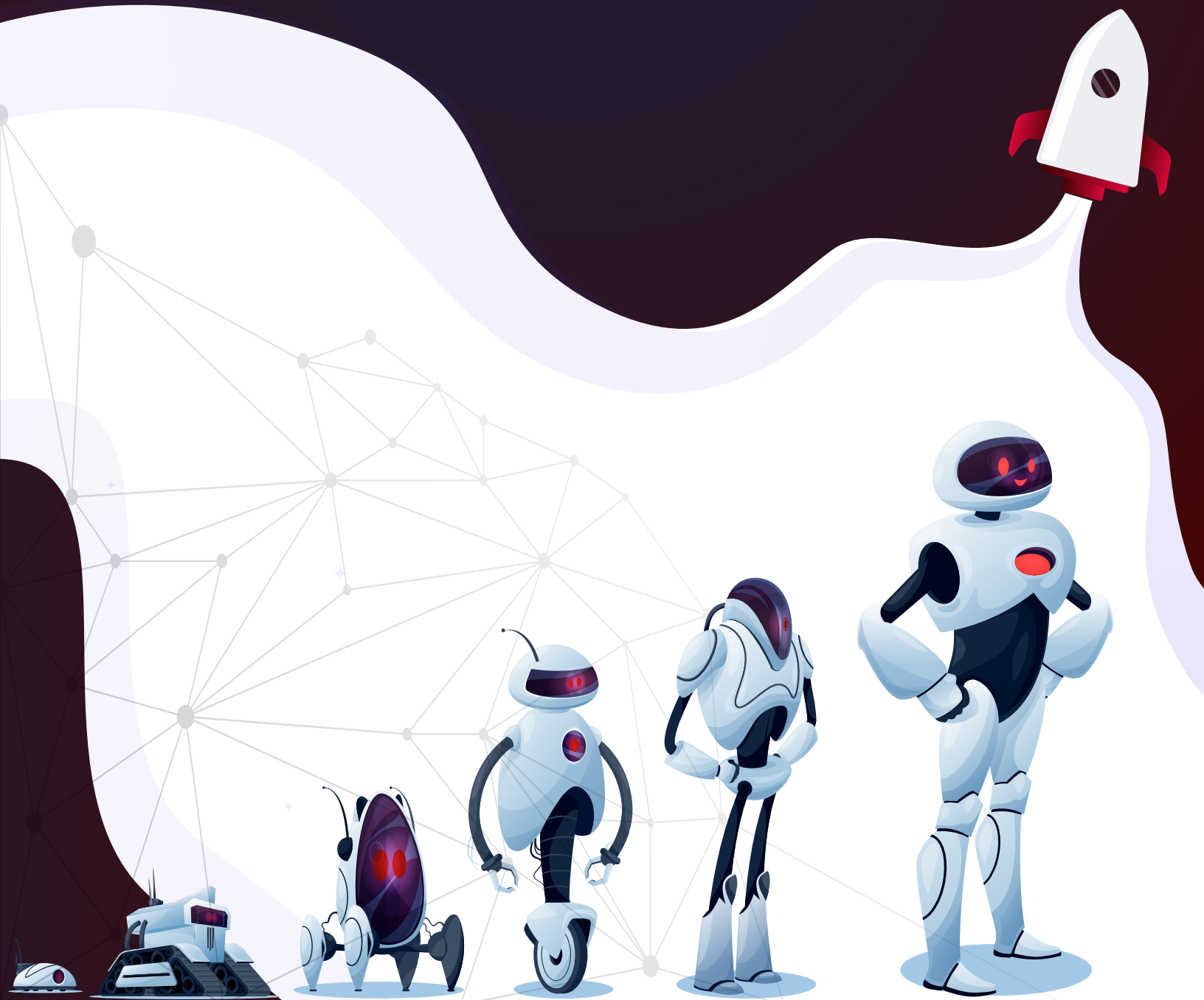
Feature Comparison	RAG	Fine-Tuning
Knowledge Updates	Directly updates the retrieval knowledge base, ensuring information remains current without the need for frequent retraining, suitable for dynamic data environments.	Stores static data, requiring retraining for knowledge and data updates.
External Knowledge	Proficient in utilizing external, user-specified resources, particularly suitable for documents or other structured/unstructured databases.	Can be applied to align the externally learned knowledge from pretraining with large language models, but may be less practical for frequently changing data sources.
Data Processing	Requires minimal data processing and handling.	Relies on constructing high-quality datasets; limited datasets may not yield significant performance improvements, and could provide even worse results than the original LLM.
Model Customization	Focuses on information retrieval and integrating external knowledge but may not fully customize model behavior or writing style.	Allows adjustments of LLM behavior, writing style, or specific domain knowledge based on specific tones or terms.
Interpretability	Answers can be traced back to specific data sources, providing higher interpretability and traceability.	Like a black box, not always clear why the model reacts a certain way, with relatively lower interpretability.
Computational Resources	Requires computational resources to support retrieval strategies and technologies related to databases. External data source integration and updates need to be maintained.	Preparation and curation of high-quality training datasets, definition of Fine-Tuning objectives, and provision of corresponding computational resources are necessary.
Latency Requirements	Knowledge retrieval followed by LLM generation frequently provides the lowest end-to-end latency.	Relying solely on the LLM to generate the output usually requires additional data to be passed in as context, increasing end-to-end latency.
Reducing Hallucinations	Inherently less prone to hallucinations as each answer is grounded in retrieved evidence.	Can help reduce hallucinations by training the model based on specific domain data but may still exhibit hallucinations when faced with unfamiliar input.
Ethical and Privacy Issues	Ethical and privacy concerns arise from storing and retrieving text from external databases.	Ethical and privacy concerns may arise due to sensitive content in the training data and the LLM context.

Table 2. Comparison between RAG and Fine-Tuning

Source: Retrieval-Augmented Generation for Large Language Models: A Survey (with additional edits by Vectara)

The GenAI Maturity Roadmap

Achieving GenAI maturity in your business requires strategically implementing use cases to balance return on investment (ROI) and the risks associated with the initiative. Always begin with smaller use cases, build confidence, and move forward gradually with moderate and advanced use cases. Along this journey, continuously evaluate the performance of each use case and improve the GenAI applications with user feedback.



Starter Use Cases

Initially, focus on internal use cases beneficial for employees that are easier to implement and have low-risk but produce high ROI. Starting from smaller internal use cases enables organizations to experience the value GenAI brings to the organization much faster. Also, such use cases give adequate time to learn the underlying concepts and prepare the organization to embrace the technological and cultural shifts.

Examples

Internal employee Q&A on HR policies or regulatory compliance - The HR department must handle common and frequently repeated employee queries. This is an example of a good area of focus in which to start your GenAI maturity journey. You can develop an employee Q&A application that answers frequently asked HR or regulatory compliance queries. Demonstrate how it can significantly improve the response time and reduce the workload of the HR employees.

Customer sales/support chatbots - Integrate GenAI into internal human-assist systems that support/sales agents use to generate quick responses to customer issues. It helps to enhance the user experience and reduce the burden on customer support teams but does not expose this technology to outside uses, as it uses the internal agent as another QA source to filter generated answers.

Moderate Use Cases

Next, leverage GenAI on moderate use cases that focus on the user experience of your customers and employees.

Examples

Related content recommendations - Develop GenAI solutions to recommend relevant content based not just on user preferences and behavioral patterns, but on the inherent semantics of the data within your organization. This can improve user satisfaction and engagement, and increase sales.

Automatic form fill-out - many applications, such as user subscriptions and product purchases, require filling out forms of user details. Use GenAI to automatically fill out the details in such forms using past information and user data entries that reduce the time required for such tasks.

Advanced Use Cases

Gen AI integration into simple and moderate use cases equips organizations with adequate experience, knowledge, and confidence to apply it in more complex scenarios. As your business becomes more mature in using GenAI, venture into more advanced use cases that can yield more transformational gains in revenue, efficiency, and cost reductions.

Examples

Automated actions - Leverage GenAI in automating manual business processes such as sending emails and submitting tickets.

Automatic document creation - Deploy GenAI application that can generate complex documents and customize them, such as Requests for Proposals (RFPs), Statements of Work (SOWs), Non-Disclosure Agreements (NDAs), or learning curriculums. These systems will help the business to save time, effort, and costs required for manually generating mostly repeated work and improve the quality and accuracy of such content.

Evaluating GenAI Technologies

Choosing the GenAI technology that best fits your business needs is critical in achieving your intended goals. Therefore, evaluating the potential technologies against several critical factors is essential to decide their suitability for the business context. Here's a high-level breakdown of these factors:

Accuracy, precision, recall - Accuracy is a measurement of how correct the content generated by a GenAI system is when compared to an acceptable response. Precision measures how many true positives the system outputs out of the total number of true and false positives in the dataset - i.e. how much relevant content it can find. Recall shows the ability of the GenAI solution to identify all relevant information.

For a GenAI solution to be effective, it should consist of high accuracy, precision, and recall. For example, Vectara's industry-leading Boomerang Retrieval model has shown higher performance in those areas. In fact, according to Vectera, its new Boomerang model provides a 54% relative improvement in Precision@1 and a 39% relative improvement in Recall@20 compared to Vectara's legacy retrieval model. Therefore, it is ideal for businesses seeking accurate and precise GenAI solutions.

Performance - Research performance comparisons of the chosen approach with others to understand the respective execution and response times. Also, check the performance of handling complex multi-language setups with industry-standard benchmark datasets. In addition, examining resource consumption data, such as CPU and memory usage, is useful to understand its scalability for large data sets.

Security - Ensure the AI system can store and handle data securely, eliminating unauthorized access. It must comply with mandatory cybersecurity regulations. Also, check for advanced security techniques, such as digital watermarking and LLM Grounding, that can combat evolving security risks.

Cost - What initial investments are required to implement the GenAI solution? Also, evaluate the implementation, modification, and ongoing maintenance costs.

Data Privacy - GenAI solutions must comply with data privacy regulations such as HIPAA, GDPR, and CCPA. There should be safe data storage and processing practices, and transparency for users on how collected data is used.

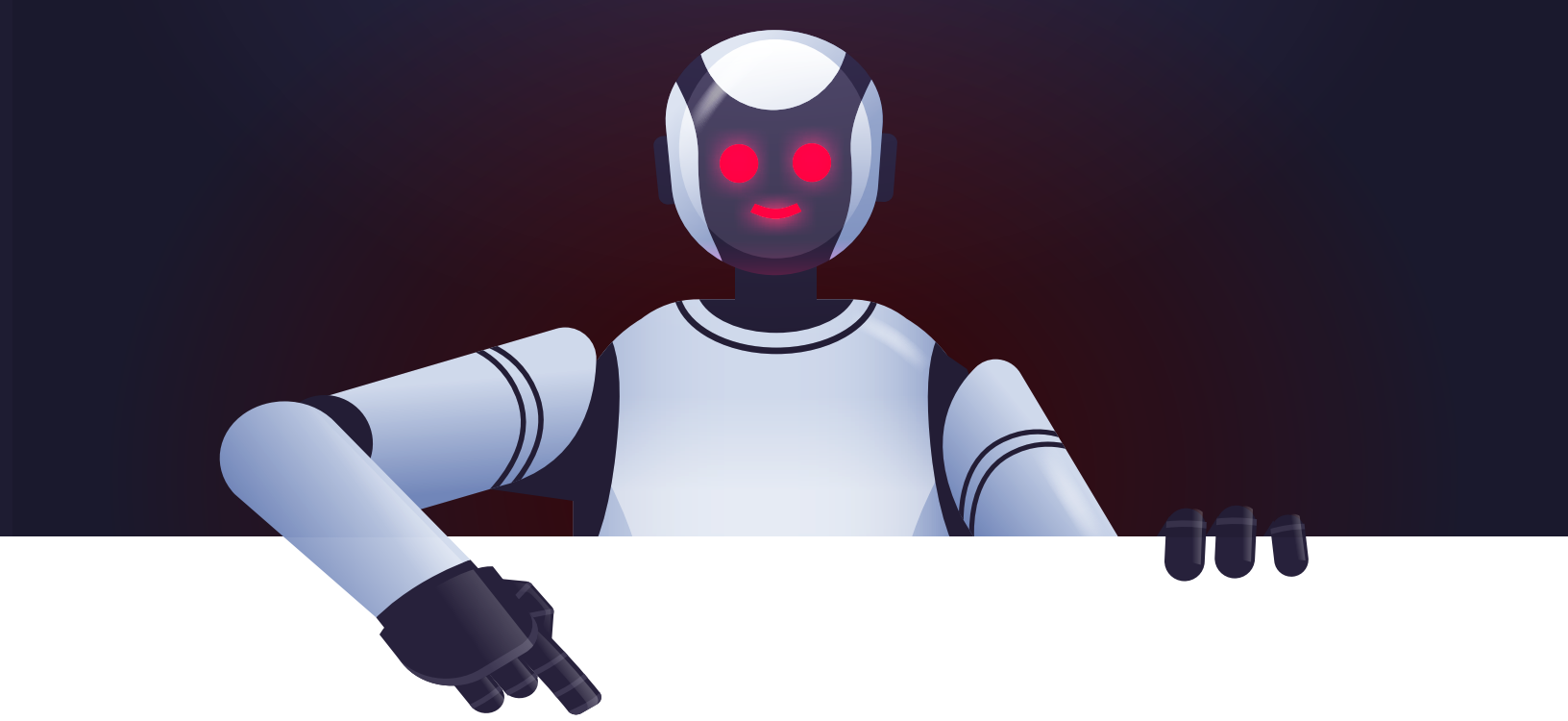
Conclusion And Next Steps

Integrating GenAI into your business operations could be exciting, but simultaneously, it will be challenging. GenAI has many use cases that can elevate your business to the next level. When leveraged effectively, it can immensely benefit the organization and customers.

In this journey, key points to consider are:

- Perform a cost-benefit analysis to understand the risks and benefits of integrating GenAI into the business.
- Understand the benefits of buying over implementing a solution from scratch.
- Evaluating your GenAI readiness in terms of organizational, data, and security preparedness.
- Select the right GenAI strategy that aligns with your business goals.
- Choose between the RAG and Fine-Tuning approaches, understanding their pros and cons.
- Evaluate the GenAI technologies before deciding on one.
- Achieve GenAI maturity by starting with smaller GenAI use cases and proceeding with more complex ones.

As the next step towards realizing your vision, you can discuss with experts in this field, such as Vectara, who can guide you on investing in your GenAI solution. Also, provide your employees with the necessary knowledge and training to prepare them for the revolution. Finally, build GenAI prototypes of your planned use cases, get user feedback, and iterate to improve the solution.



About Vectara

Vectara is an end-to-end platform that empowers product builders to embed powerful Generative AI features into their applications with extraordinary results. Built on a solid hybrid search core, Vectara delivers the shortest path to an answer or action through a safe, secure, and trusted entry point. Vectara is built for product managers and developers with an easily leveraged API that gives full access to the platform's powerful features. Vectara's Retrieval Augmented Generation (RAG) allows businesses to quickly, safely, and affordably integrate best-in-class conversational AI and question-answering into their application. Vectara never trains its models on customer data, allowing businesses to embed generative AI capabilities without the risk of data or privacy violations. To learn more about Vectara, visit www.vectara.com

Palo Alto, CA, USA

www.vectara.com

Book a Demo

vectara