

An executive's guide to operationalizing generative AI

From experimentation to implementation: how to use generative AI for real-world scenarios



Table of contents

Start your generative AI journey 3

Part 1

Understanding the generative AI landscape 5

What is generative AI?	<u>6</u>
What is machine learning?	<u>6</u>
What is a large language model (LLM)?	<u>6</u>
What is retrieval augmented generation (RAG)?	<u>8</u>
What is a vector database?	<u>11</u>

What generative AI can do 12

Step 0: Find the why and determine what's possible	<u>13</u>
--	-----------

Adapting to your industry 16

Telecommunications	<u>19</u>
Financial services	<u>20</u>
Retail	<u>21</u>
Automotive and manufacturing	<u>23</u>
Public sector	<u>24</u>

Start here if you know what you want
to achieve with generative AI



Part 2

Operationalizing generative AI: one small step for the machine — one giant leap for your organization 27

Step 1: Identify your ideal outcome	<u>28</u>
Step 2: Figure out the impact. Measure success.	<u>29</u>
Step 3: Pick a model (way forward)	<u>30</u>
Step 4: Try fast, fail fast	<u>34</u>
Step 5: Governance and operations	<u>36</u>
The thing about data safety	<u>37</u>
Step 6: Set a timeline. Give it benchmarks.	<u>38</u>

The start of a new era 40

Start your generative AI journey

Generative AI was the most disruptive technology to emerge in 2023. Across industries, it's predicted that generative AI will shape pretty much everything for years to come — but how many can say they've cracked the code to put it to work now?

While companies are getting a handle on the generative AI wave, some have already begun seeing results. For example, at Cisco, support engineers can instantly find relevant, summarized answers from similar support cases, internal discussion forums, and knowledge articles related to customer issues. Cisco has already reaped the benefits of generative AI, with 90% of support requests resolved with its reimaged search solution and 5,000 support engineer hours saved per month.¹

In the realm of ecommerce, you've probably seen generative AI at work. Generative AI can analyze a customer's past purchases, browsing history, and preferences to generate personalized product recommendations with a chatbot. And on the back end, the use of generative AI promises to increase customer engagement and retention, improve fraud detection, and more.

To demystify generative AI's capabilities and decide how to make it work for you, you need a step-by-step guide on activating your data. In this ebook, we'll guide you along your journey from **wishful thinker to AI expert**. Think of this as a roadmap to help revolutionize your business outcomes using generative AI.

99% Yet only 32%

of organizations believe generative AI has the potential to drive change within their organization, be it internal or external²

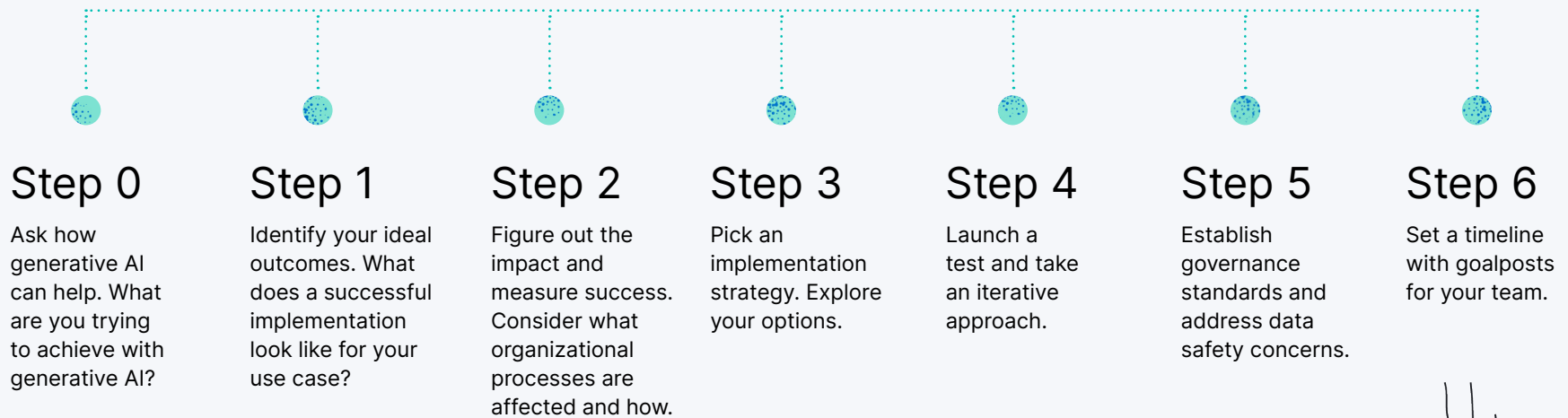
of leaders are confident in their ability to implement AI in their organizations³

¹ Elastic, Cisco creates AI-powered search experiences with Elastic on Google Cloud 2024

² Elastic, The Elastic Generative AI Report, (2024)

³ Russell Reynolds, Embracing the Unknown: How Leaders Engage with Generative AI in the Face of Uncertainty, (2024).

Here's what to expect from this guide:



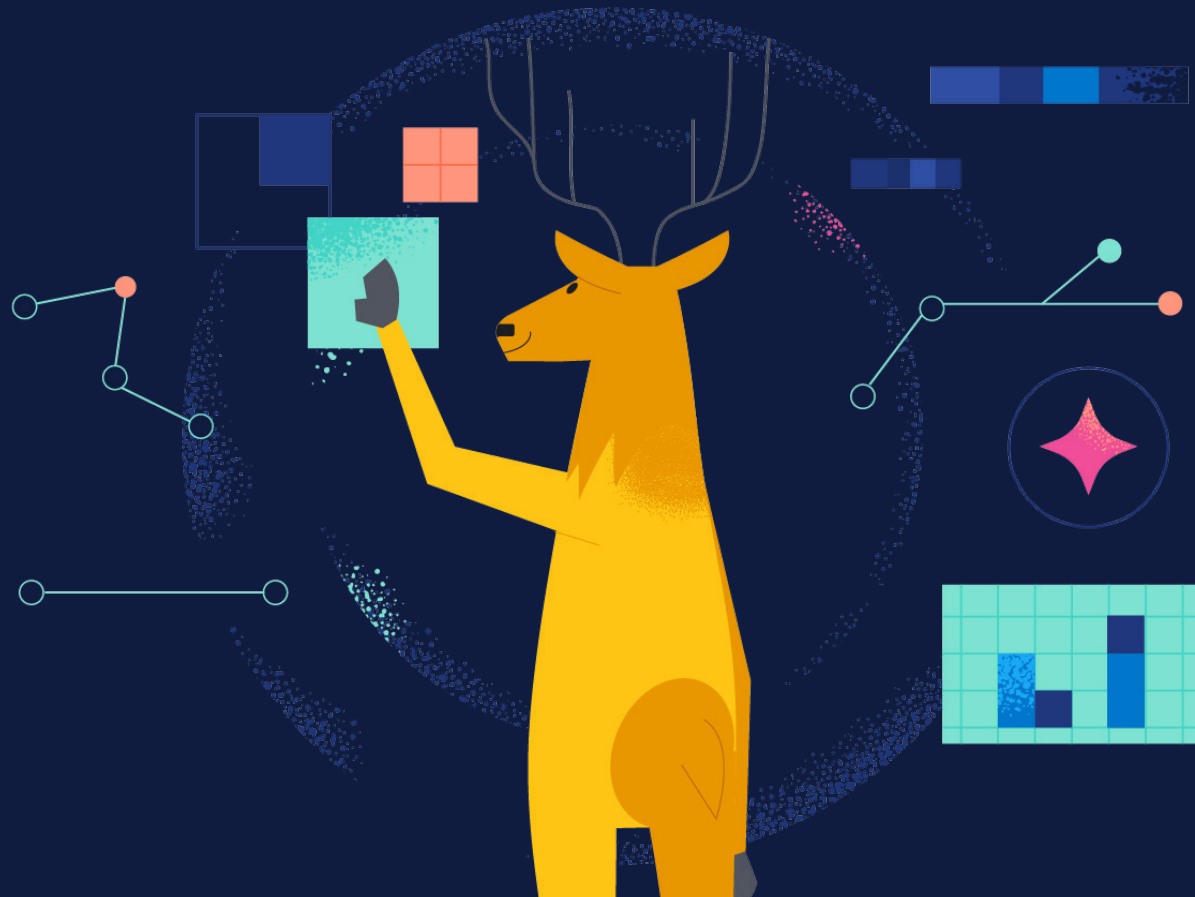
[Skip ahead if you know what you want to achieve with generative AI](#)

But first, let's brush up on the basics.



Part 1: Understanding the generative AI landscape

You don't have to be an expert on generative AI to instrument a plan to operationalize it. However, understanding what components are at play enables you to make informed and strategic decisions throughout the process. Let's lay out the building blocks.



What is generative AI?

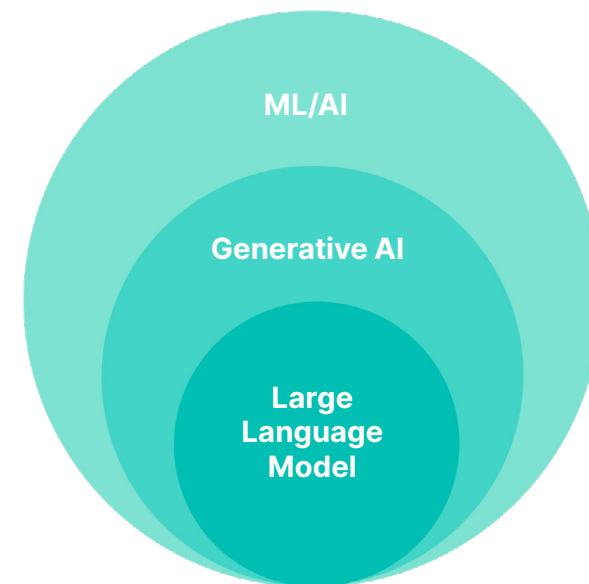
Generative artificial intelligence, or generative AI, refers to deep learning models that can generate outputs when prompted. It's important to understand that the generative capability of this technology rests on its ability to predict statistically probable outputs, which it achieves with the help of machine learning. **Data is at the heart of operationalizing generative AI and is the key to success in both its implementation and results.** More on this shortly.

What is machine learning?

Machine learning, or ML, is a subset of artificial intelligence that uses algorithms to derive knowledge from data. These algorithms parse through data and “learn” — in a supervised, semi-supervised, or unsupervised context — the patterns and similarities that enable them to make decisions. Machine learning is the underlying technology that gives generative AI, such as large language models, the ability to “learn” continuously.

What is a large language model (LLM)?

A large language model, or LLM, is a computational model that sits under machine learning. It is a type of generative AI that deals with human language specifically. Having been trained on vast sets of primarily public language datasets, an LLM can perform a variety of natural language processing (NLP) tasks, including recognizing, analyzing, summarizing, predicting, translating, or generating text. In the context of operationalizing generative AI, LLMs are what enable generative AI to communicate in natural (or human) language.





Let's talk about hallucinations

A hallucination is an incorrect or misleading result generated by an LLM. You're probably wise to ChatGPT's sometimes questionable responses. The output seems legitimate... but is it really? If the LLM (ChatGPT is powered by an LLM) can't find the answer, it tends to make one up. This blind spot is important to consider when discussing using LLMs in enterprise applications. How do you ensure that the outputs generated are relevant and accurate? This is where retrieval augmented generation (RAG) comes in.

You: How much PTO do I have left?



AI: There are 200 days left in the year.

You: How do I fix my video doorbell that won't connect to Wi-Fi?

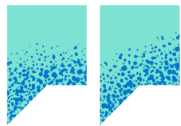


AI: The best video doorbells offer 4K recording and instant...

What is retrieval augmented generation (RAG)?

Think of retrieval augmented generation, or RAG, as your line of defense against hallucinations. The output generated by an LLM is augmented, or “checked,” by retrieving information from a particular dataset, or data context, that you provide using highly relevant search powered by a vector database. For example, via RAG and in response to a user query, an organization searches its policy documents and provides relevant responses to an LLM so that it may respond to questions using the organization’s policies. Beyond a defense against hallucinations, **RAG lets you use generative AI with your proprietary dataset** — this is its biggest benefit.

In the context of operationalizing generative AI for business applications, RAG is important for multiple reasons: it can deliver better, more relevant results, and offers a quick way to bootstrap or utilize your own proprietary data. It’s also more cost-efficient than training or building your own LLM. In this way, RAG is the key to successful generative AI integration. RAG goes beyond the limits of general LLMs to create “the next-generation search engine.”

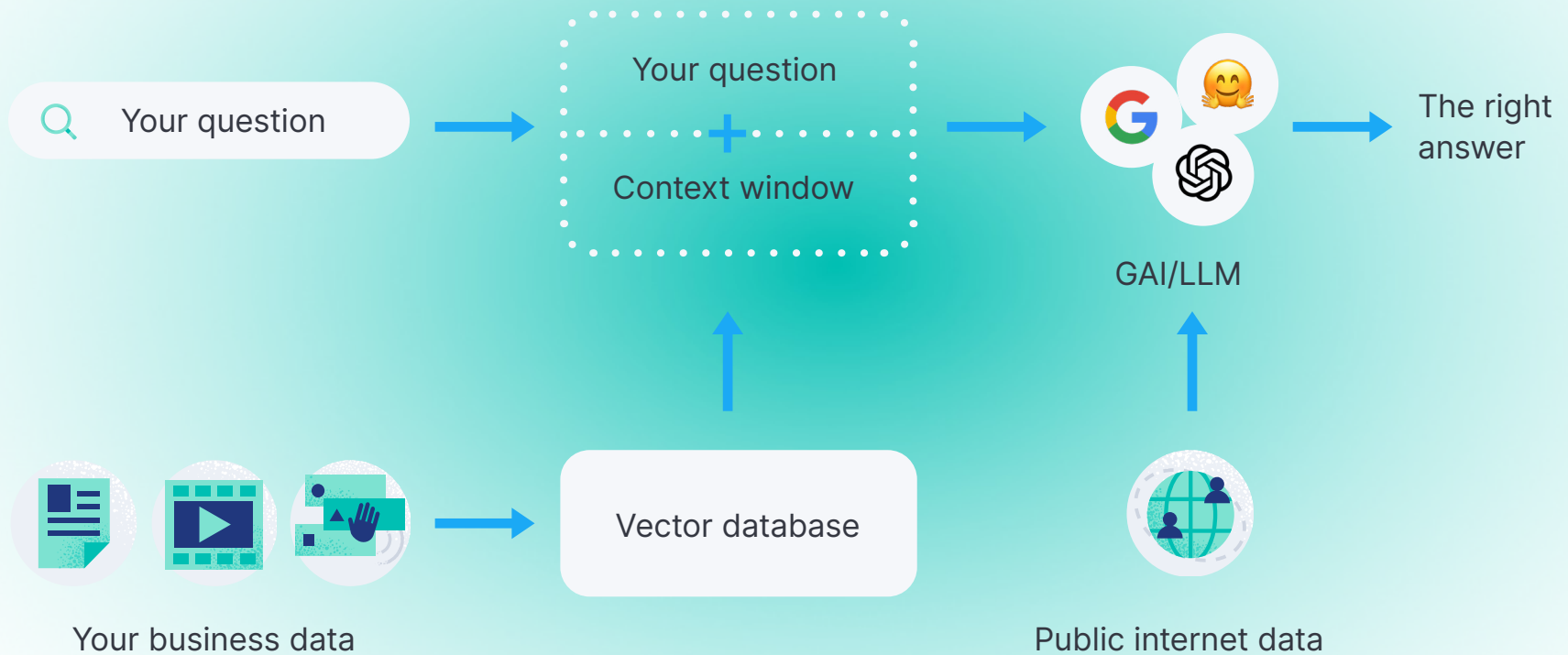


“RAG helps to create the next-generation search engine”

— **Baha Azarmi**
VP, Global Customer Engineering at Elastic

Retrieval augmented generation (RAG)


RAG lets you use generative AI with your proprietary dataset.



RAG is a new way to answer user questions:

Regular search

A person searches for **a term**

 Work from home policy

A query is performed

```
query = {
  "bool": {
    "should": [
      {
        "text_expansion": {
          "ml.inference.text_expanded_predicted_value": {
            "model_id": model_id,
            "model_text": question
          }
        }
      }
    ]
  }
}
```

Results are presented


	Doc Title	Date added
1	Employee code of conduct	01/01/2010
2	IT usage policy	01/01/2015
3	Home page work in progress	01/01/2022
	etc...etc...	

Users choose a document and read its contents

Employee code of conduct
 Lorem ipsum Lorem ipsum
 Lorem ipsum Lorem ipsum
 Lorem ipsum Lorem ipsum
 Lorem ipsum Lorem ipsum
 Lorem ipsum Lorem ipsum
 Lorem ipsum

Generative AI without RAG

A person asks **a question**

 What is our work from home policy


A query is performed

```
query = {
  "bool": {
    "should": [
      {
        "text_expansion": {
          "ml.inference.text_expanded_predicted_value": {
            "model_id": model_id,
            "model_text": question
          }
        }
      }
    ]
  }
}
```

An answer is derived that has no context to your domain

A work from home policy is necessary when you have employees who work hybrid or...

A person asks **a question**

 What is our work from home policy


A query is performed

```
query = {
  "bool": {
    "should": [
      {
        "text_expansion": {
          "ml.inference.text_expanded_predicted_value": {
            "model_id": model_id,
            "model_text": question
          }
        }
      }
    ]
  },
  {
    "match": {
      "text": question
    }
  }
}
```

Results are provided as context

	Doc title	Date added
1	Employee code of conduct	01/01/2010
2	IT usage policy	01/01/2015
3	Home page work in progress	01/01/2022
	etc...etc...	

An answer is derived from search results by an LLM

 Employees are encouraged to work from home, provided that they are able to effectively... etc...etc...

RAG-enabled

What is a vector database?

A vector database stores vector embeddings, or numerical representations of words, images, or video. These embeddings are multi-dimensional and enable semantic search, a type of search that looks for the intent and contextual meaning of a query. In contrast, a textual search looks only for results that match keywords in the search query.

In the context of RAG, a vector database enables fast semantic search based on the prompt provided to the generative AI. This is what makes RAG possible.

Generative AI is good at NLP, but traditional keyword search is not able to take in natural language and serve up the best results to feed to the generative AI. Therefore, with a vector database feeding generative AI with search results that are semantically similar to the original prompt, the generative AI can generate answers that are more relevant. Think of vector databases as the knowledge bank that enables generative AI to answer questions with accurate information.

However, generative AI is not limited to vector databases. Using RAG, generative AI can tap into relational databases, graph databases, document-based databases, or even keyword search engines. The best database for you often depends on the nature of the data, the specific algorithms being used, and the performance requirements of the system. For example, relational databases can be used for structured data, while graph databases are well-suited for data with complex relationships and traditional search engines for full-text search.

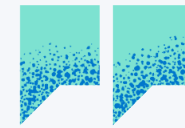


Vector

Semantic

Keyword

Hybrid



“All roads lead to hybrid search.”

— **Serena Chou, Director,
Product Management at Elastic**

While **semantic search** serves up results that match the meaning of a query, **keyword search** still plays an important role in matching results to exact keywords from the queries. Hybrid search is the practice of using both vector — frequently used for semantic search — and keyword search technologies to provide what are often the most relevant results to a generative AI.

Bottom line: a hybrid search solution will likely serve up the most relevant results for generative AI experiences in your organization.



What generative AI can do

We've talked a lot about the underlying technologies and fundamental concepts, but what exactly can generative AI do?



Create

By learning patterns in its training data, generative AI can “create” or *generate* outputs. It iterates on existing data to produce new ideas, images, insights, and more.



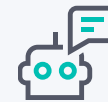
Summarize

Thanks to its natural language processing capabilities, generative AI can analyze a text and summarize it. Need to review lengthy documents in a short time span? Generative AI to the rescue.



Discover

The key to generative AI is its underlying search technology. This enables the generative AI tool to receive a query, search a vast set of private or public data, and produce a reply.



Automate

Say your organization uses two different cloud platforms for different services. Each cloud platform will generate logs in different formats. By automatically transforming this data into the same format and mapping it with AI, your team can summarize and ask questions about the data with generative AI. Your IT team can focus on monitoring and managing your systems instead of performing labor-intensive tasks.

Step 0: Find the why and determine what's possible

With so many possible ways to add value, zeroing in on a starting point is crucial. How do you harness the power of generative AI to empower your team, meet the rapidly evolving expectations of customers, and propel your company to new heights? By zeroing in on the one area in which generative AI can add the most value to your organization.

Consider these questions:

1

What problem am I trying to solve?

Is there an area in your business that is especially inefficient? Where are your employees spending significant time on repetitive tasks? Are they constantly searching for existing information in internal databases or external engines?



For example:

Do your employees have trouble finding information — whether it's project updates or HR-related information? Does your security team have tasks they can automate so they can free up time to take a more proactive stance? Is there entropy in your customer service processes because customer service engineers aren't always communicating in real time with customer service agents about known issues and recent fixes?

Friendly reminder:

You will need to work with the team whose workflow you will be impacting. For example, if you decide to update an inefficient HR process, bringing in HR from the start will be important to get stakeholder approval and bolster support.



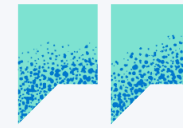
2 Is this a problem I can solve with a knowledge base?

A knowledge base is a collection of informative content — be it support articles or internal processes.



Consider what content you have to draw from, and whether that content can be mined to solve your problem more efficiently. Is it sufficient to automate and personalize responses?

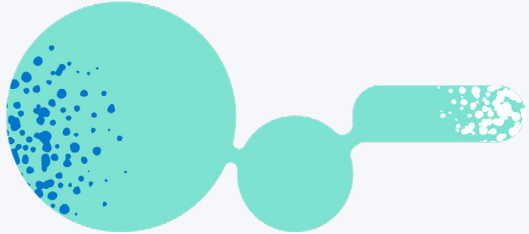

For example: You've identified that employees spend a considerable amount of working hours looking for HR-related information. Your team doesn't have an HR expert, so employees are being redirected to the company intranet, where they end up having to read through a variety of policy documents to, say, find out how many vacation days they have left in the year. To solve this problem, you'll need a knowledge base, such as your HR policy documents, as well as access to employee personal data to personalize the response.



"When you pull a thread, a thousand [threads] keep coming."

— **Baha Azarmi,**
VP, Global Customer
Engineering at Elastic

Step 0 is understanding what processes are limiting productivity. By cutting the mundane tasks, you free up space for your employees to be creative. A well-executed implementation is a responsible one. **Investing in your employees, upskilling them accordingly, and planning for adjusted workstreams and processes are all crucial parts of a successful marriage with generative AI.**



As you search for the why, **K.I.S.S. — keep it simple and specific**. Identifying which generative AI use case you want to tackle first is a great initial step in operationalizing generative AI. Then, smaller projects will set you up for effective implementation.

For example, in the previous HR scenario, generative AI could serve multiple use cases:

1

Discovery

An employee queries the interface — How many vacation days do I have left in a year? To respond, the AI needs to conduct a search and present the documents that are relevant to the query, pulling in HR policy documents and the employee's record.

2

Summarization

Taking it a step further, the generative AI might analyze the documents and summarize them for the employee into a conversational response. “You have 10 vacation days left for this year and four floating holidays you can use. Check out the intranet page on paid time off policies.”

3

Creation and automation

The chatbot could save managers time by creating a response to approve or deny a request to use vacation time, and even give a reason why. It could also create calendar invites and log the PTO request in the system.

Adapting to your industry

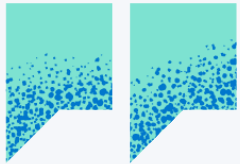
Copilot, assistant, bot — generative AI takes various “forms” that provide valuable productivity-enhancing services in a variety of fields and industries. As a security and observability copilot, or with internal and external apps, generative AI helps companies increase their efficiency, upgrade their security efforts, improve their customer experience, and accelerate their competitive differentiation.

By using data to unlock the best AI responses and harness the power of underlying search technology, **companies can decrease time spent on repetitive tasks, reduce response times, and overall, improve productivity.** Adding RAG to the equation allows you to tap into your proprietary data for generative AI responses that are secure while respecting document and user-level permissions.

Suddenly, you’re looking at next-level speed and relevance — the same speed and relevance that increasingly tech-savvy customers have come to expect. Ensuring that your services match those expectations is critical. So is using generative AI in the ways most relevant to their user experience. There’s nothing worse than investing in buzzy build-outs that nobody uses.

Generative AI is most commonly integrated into a company’s IT infrastructure as an AI assistant or a security and/or observability copilot.





Every business has opportunities with generative AI because generative AI is fundamentally a much more human, intuitive way to get information from information systems.

 **Ash Kulkarni**
CEO at Elastic





AI assistants

Make the most of generative AI conversation skills for employees and customers with internal and external apps. AI assistants provide every user flexible, adaptive, and personal help as an on-hand expert, a personal shopper, or even a schedule minder.



Security and observability copilot

Boost your observability and security capabilities with generative AI copilots. Designed to work in tandem with IT teams, generative AI copilots function as expert problem-solving partners. For example, you can prompt your copilot for a detailed description of why a security alert was triggered and get recommended steps to triage and remediate the attack (based on previous similar attacks that your organization has encountered). This type of prompt can generate a dynamic runbook for an organization.

These integrations enable companies across industries to increase their personalization, automation, and productivity potential, leading to **three major generative AI use cases:**

Improving operational resilience

Operational resilience is vital to keeping systems running smoothly. With a boost from generative AI, IT teams can accelerate root cause analysis, correlate more data across all environments to pinpoint issues faster, and have a dedicated discovery tool on hand to speed up their responses — all in favor of business continuity.

Elevating customer experiences

Customer satisfaction is at the core of every business. Generative AI provides your teams with the tools to resolve issues faster and get the information they need while also giving your customers personalized attention and fast access to relevant information. The result? Improved customer experiences and better business outcomes.

Mitigating security risk

As the digital world evolves at a breakneck pace, new sophisticated security threats emerge. Dealing with them requires dynamic and proactive measures, not to mention the expertise to respond to and manage the threats. Generative AI can not only bolster your security team and operations but also automate alerts and maintain a proactive posture.

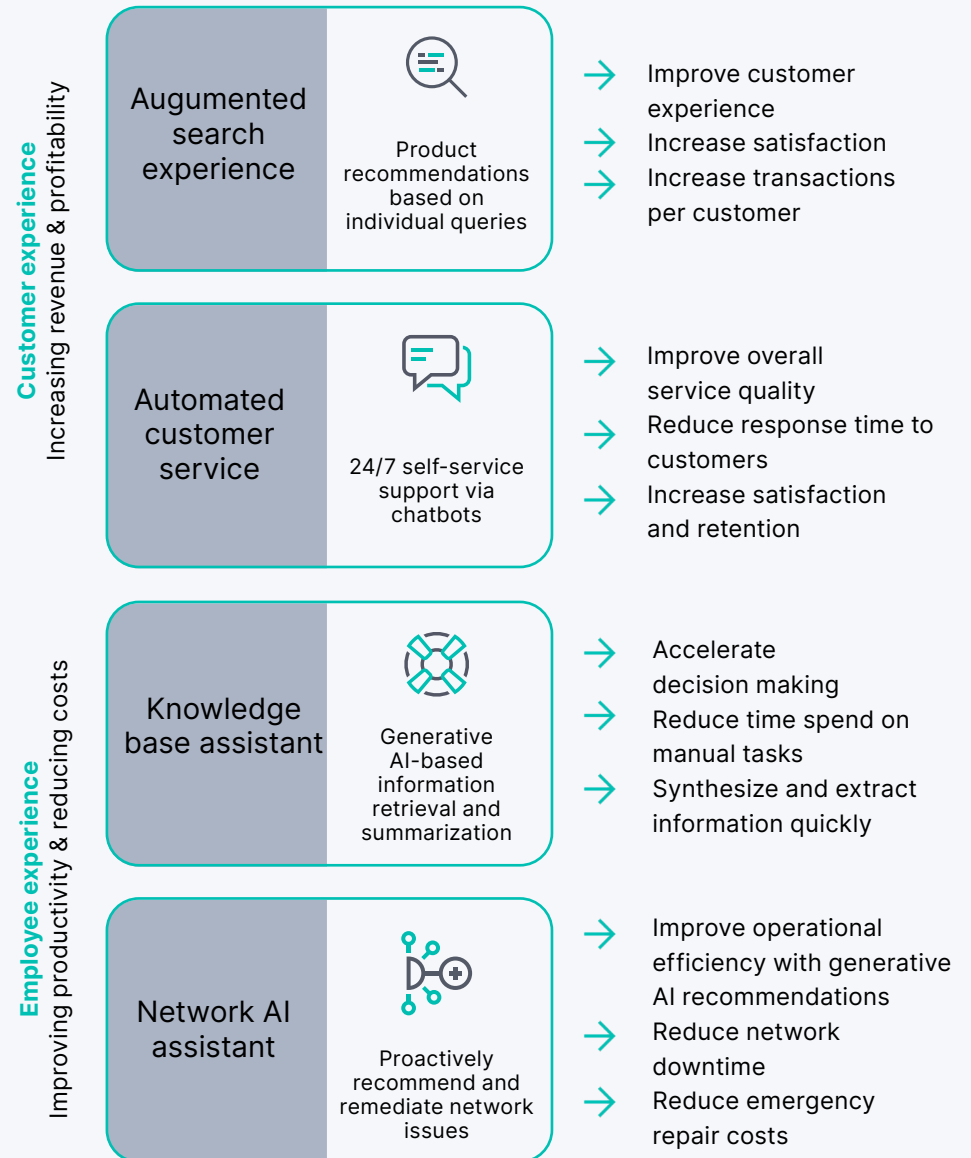
Across industries, generative AI can augment existing employee and customer experiences by delivering personalized, relevant, and prescriptive responses to their queries. No matter what sector you're in, there's a way to operationalize generative AI to power your search and unlock new capabilities of your data.



Telecommunications

For telecommunications companies, generative AI is projected to create economic value exceeding \$60 billion.⁴ With generative AI, telecommunications companies can enable their employees and customers to query their website or internal workplace knowledge base to get personalized and relevant responses, fast. The result? Better customer service and improved productivity.

⁴ McKinsey, Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom, (2024).





Financial services

With generative AI, financial services companies can further personalize their customer and employee experiences. Enhancements in customer experience, fraud prevention, and automation are projected to create economic value exceeding \$250 billion for the financial services industry.⁵

⁵McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Customer experience Increasing revenue & profitability

Retail banking assistant

Generative AI-based information retrieval and summarization

- Expand visibility of personal finances
- Provide tailored offers for higher conversion
- Increase satisfaction and retention

Enhanced customer service

Proactively recommend and remediate network issues

- Improve overall service quality
- Reduce response time to customers
- Increase retention

Employee experience Improving productivity & reducing costs and risks

Fraud detection summarization

Anomaly detection/transaction summary and next best action

- Improve accuracy and speed of fraud detection
- Reduce costs by automating tasks
- Reduce financial loss

Virtual assistant

NLP-based information retrieval and summarization

- Accelerate decision making
- Reduce time spend on manual tasks
- Synthesize and extract information quickly



Retail


Most compelling to retail, generative AI promises to increase customer retention by increasing search relevance, recommending additional products, and sending personalized follow-ups across channels. Have you ever received one of those “You forgot something in your cart!” emails? AI can automate and improve these for better recommendations and more personalized product discovery.

Whether it’s building next-generation customer experiences to drive e-commerce sales or empowering employees with the latest technology to improve productivity, generative AI is projected to create economic value exceeding \$240 billion for retailers.⁵

⁵McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Customer experience Increasing revenue & profitability


Personalized product search and discovery



Question answering, tailored search experience

- Improve website conversion rates
- Increase transactions per customer
- Increase satisfaction

Enhanced customer service



Self-service interactions via chatbots

- Reduce response time to customers
- Improve service, reduce churn
- Increase retention

Employee experience Improving productivity & reducing costs


Enhanced customer service



Enhanced agent experience and interaction

- First contact resolution
- Faster onboarding
- Reduced agent turnover

Predictive maintenance



Assess health of critical systems to prioritize critical maintenance tasks

- Reduce downtime of equipment and system
- Reduce emergency repair costs
- Improve operational efficiency

Case study: HSE

HSE is one of the leading brands in the European live commerce sector.⁶

“For Home Shopping Europe [HSE], commercial success starts with website personalization and relevance.”

Peter Strasser
Software Developer at HSE



The opportunity

Like any ecommerce business, the search capability is fundamental to the customer experience and sales. HSE must cater for customer journeys that originate from many channels, resulting in diverse search terms that reflect where the customer was introduced to the product.

HSE used generative AI and LLMs to extract the semantic meaning of a customer query and generate results that complement traditional keyword matching.



The result

HSE saw a **4% increase in click-through rate** and an **8% increase in customer satisfaction** thanks to more accurate and relevant search results.



Insight

Focus on an area you're already looking to enhance, like the customer search experience. See how you can integrate generative AI to take the experience to the next level with personalization and relevance.

⁶ Elastic, HSE increases customer satisfaction while reducing maintenance time by 42%, using Elasticsearch on AWS, (2024).




Automotive and manufacturing

Every step of the automotive and manufacturing industry process can be streamlined with AI, giving it a projected economic value exceeding \$170 billion.⁵ Generative AI has the potential to transform the industry from product research and development innovation to personalized customer retention strategies. Flying cars? Maybe!

⁵McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Customer experience Increasing revenue & profitability


Interactive digital manuals



Virtual product assistant

- Real-time answers on product features, maintenance, and troubleshooting
- Reduce support queries
- Improve satisfaction

Enhanced customer service




Self-service interactions via chatbots

- Reduce response time to customers
- Improve service, reduce churn
- Increase retention

Employee experience Improving productivity & reducing costs

Operational technology optimization



Predictive maintenance: Issue and resolution summarization

- Identify and address issues quickly
- Improve operational efficiency and decision making
- Reduce manufacturing costs

Product sentiment analysis



Summarize product percentage and recommend improvements

- Improve product offerings with customer PoV
- Decrease time-to-value of new product offerings



Public sector

Generative AI can significantly accelerate mission outcomes, improve citizen services, and better connect government analysts and security professionals to the right data at the right time by securely connecting generative AI with agency data.



Workload reduction

Automate manual processes and workflows



Compliance

Enable role-based data access



Real-time situational awareness

Make more accurate decisions



Employee productivity

Find the right information at the right time



Citizen experiences

Build trust via personalized digital interactions



Public services

Increase accessibility and self-service options



Dynamic intelligence

Accelerate mission search and insights



Cybersecurity

Conduct real-time risk assessment and analysis

Citizen-facing applications include:

- Personalized access to public services
- Streamlined online citizen experiences
- Increased accessibility and self-service options

Employee-facing applications include:

- More accurate investigations and intelligence
- Improved productivity by automating manual processes and workflows
- More efficient procurement processes

Case study: Relativity

Relativity helps corporations, law firms, and agencies store and utilize data for e-discovery and legal search.⁷

“The biggest challenge Relativity customers are facing right now is data explosion from heterogeneous data sources. The challenge is really compounded by the differences in data generated from different modes of communication.”

Brittany Roush
Senior Product Manager



The opportunity

Relativity needed to consolidate its data while maintaining a security-first stance. As a result of this explosion of data, sources, and complexity, traditional keyword search approaches were ineffective. Enter RAG.



The result

Together, with RAG and a vector database, Relativity implemented search experiences built on proprietary data, and offered users a fast, relevant, and accurate search experience. Its generative AI solution meets compliance standards such as PCI, DSS, SOC2, and HIPAA.



Insight

Build with scale in mind. Starting small can help identify generative AI's capabilities and hone in on the most relevant applications. Once you find the sweet spot, the sky's the limit.

⁷Elastic, Relativity uses Elasticsearch and Azure OpenAI to build futuristic search experiences, today (2024).

So you understand the immense economic potential of generative AI across industries. You might even have some potential use cases in mind. Hopefully, you also have your “why.”



Yet, implementing generative AI can seem like an arduous and disruptive process. There are privacy concerns, some legwork to be done in the realm of compliance, and changes to how people do their jobs. Responsible operationalization requires training, upskilling, and a partial reorganization of the workforce.

Despite those challenges, the value generative AI can bring to a company is undeniable. Implementation is inevitable to remain competitive. The good news? You can achieve quick time to value with tests that aren't completely production ready. In other words, it's time to get started.

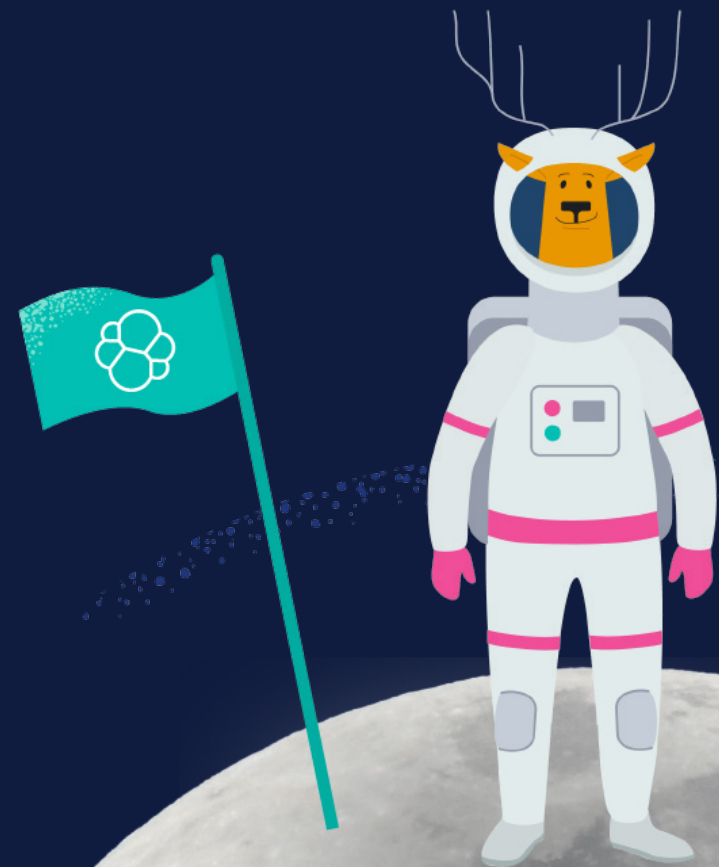
Part 2:

Operationalizing generative AI One small step for the machine — one giant leap for your organization

Operationalizing generative AI doesn't happen all at once. It's an iterative approach that requires planning and a clear outcome. By starting with a single generative AI project — a small step — you empower your team through the inevitable learning curve and allow them to workshop the processes, fine-tune the technology, and address concerns, setting them and your company up for success — the coveted giant leap.

Now, here's how you make it happen.

Leaping forward
thanks to RAG!



Step Identify your ideal outcome

1

You've honed in on a problem. You know you are looking to optimize an inefficient process. You now need to think about how users will interact with your solution. Are you augmenting a search application or a chatbot? Are you looking for a new way to interact with your teams or customers?

Your thought process might look something like this:

- ||→ You're seeking more customer retention.
- ||→ You've decided to implement a personalized product search and discovery application.
- ||→ You create metrics for success. Think of this as your "sub-why."

You want to utilize generative AI. Why? To personalize product search and discovery. Why? **Here is your ideal outcome:**

By interacting in this new way with our data, customers will effortlessly find the products they need, and discover products that they might want, based on their search history and location. As a result, customer retention will improve.


- ||→ Now you start the great big task of operationalizing your first generative AI project.



Ask yourself this:

What actions and outcomes can be created out of this new way to interact with your data?

The answer to this question will help set your goalposts. Identifying your ideal outcome determines what "good" looks like for your project, and on a larger scale, what it looks like for your company.



Step 2

Figure out the impact. Measure success.

To succeed in operationalizing generative AI, you'll need to establish a set of KPIs that help you measure what "good" means to you. Understanding how generative AI is moving the productivity needle in your organization is only one of many performance indicators.

Others might include an increase in customer satisfaction, measured by reviews in a customer support context, a decrease in support tickets, or faster resolution times. Depending on the use case you're testing, you'll need to establish corresponding performance indicators. Embedding these into every step of the testing process is vital to understanding the progress you and your team are making.

Basic performance indicators

1

Productivity impact

Measure the changes in productivity that result from your use case. Compare the time required to complete certain tasks against the time required without the use of generative AI.

2

Scalability

Evaluate how well the model scales to an increase in usage and demand. Is it still performing reliably and accurately?

3

Bottom line

Assess what impact the implementation of generative AI has had in terms of business costs. You might want to include certain business metrics in this review, such as the number of customer complaints recorded, or any changes to your sales.

4

Compliance

Continually monitor the generative AI's adherence to data privacy regulations.

5

Customer satisfaction

Review business metrics such as customer churn, increased sales, and maintained brand loyalty, and examine feedback from customers.

Use these indicators to determine whether a project is feasible, actionable, scalable, and affordable. These indicators will help you determine your ROI and can be broadened in the future as you expand your use cases.

Step Pick a model (way forward)

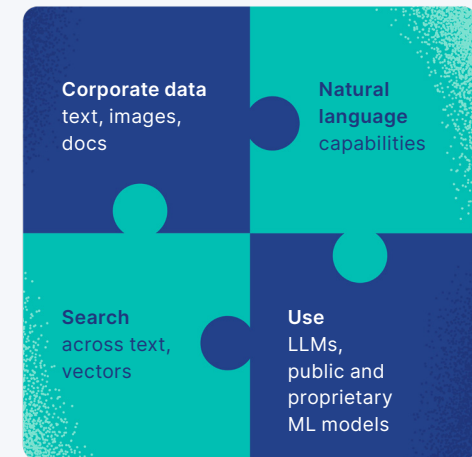
3

How do you build a generative AI architecture that meets your business needs? Many things will influence the choice you make: cost, language, your IT ecosystem, your deployment capabilities and timeline, data privacy regulations, and governance. For this reason, taking a narrow stance — starting with a simple, specific use case — is crucial.

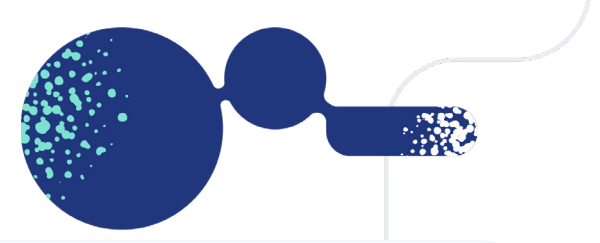
To operationalize generative AI, you'll need these components:

- ||→ **A fully managed cloud infrastructure** will increase agility, improve cost efficiency, and reduce wasted resources. Chips and hardware are evolving at breakneck speeds. If you invest in building out your own AI data center, it might become obsolete in a few months.
- ||→ **An LLM** will be the foundation that enables generative AI to communicate in and understand natural language.
- ||→ **A data platform** that includes vector, hybrid, and traditional keyword search that can be used to enrich the LLM with the right context from your proprietary data.
- ||→ **Extensive APIs** that enable you to enrich and pass your data to the LLM and your search engine.

The ingredients enterprises need for AI search



How you combine these components — whether you fine-tune your own model, bring your own vector database, bring your own model, or any combination thereof — will determine your implementation scheme, affecting your timeline, the complexity of your test, and whether you may need to supplement your team.



Pre-train an LLM

Fine-tune a model

RAG

This resource-intensive approach entails starting from scratch by training a large language model on a large set of data.

This approach uses an existing LLM with your search engine and a vector database to provide your proprietary data with context.

This process uses an existing pre-trained LLM and a set of techniques to tune the model to meet your needs.

Cost

\$\$\$\$

\$\$\$

\$\$

Deployment time

Long, deployed in months

Moderate, deployed in weeks

Fast, deployed in days

Data privacy

Do you have a large enough dataset to provide the LLM significant learning material? If not, you'll need public data. Will you combine public and private data?

Do you have a large enough dataset to provide the LLM significant learning material? If not, you'll need public data. Will you combine public and private data?

This approach enables you to keep your private data private.

Accuracy and relevance

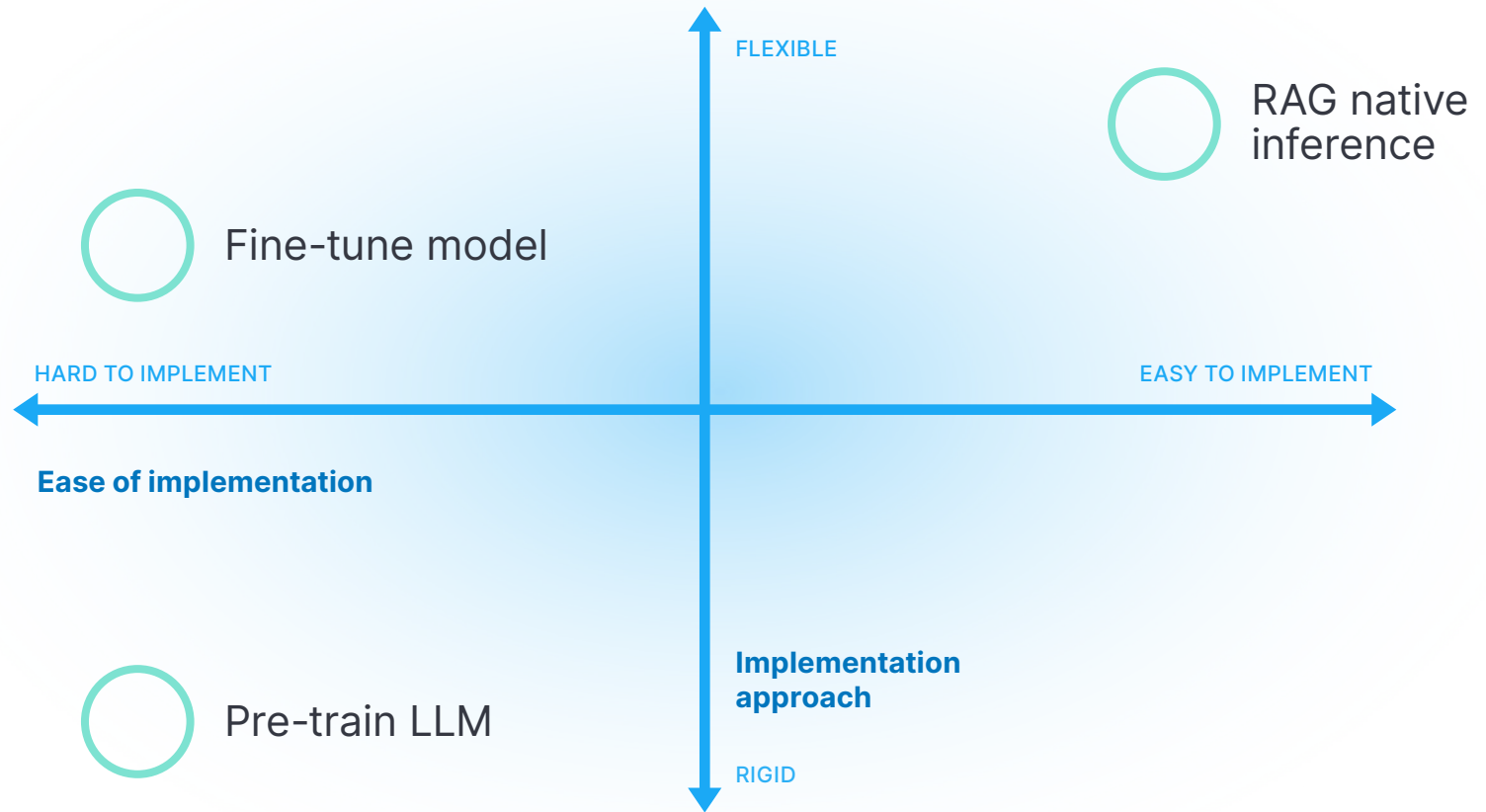
Difficult to ensure consistently

Easier to ensure accuracy and relevance for the specific tasks the model has been tuned for.

The benefit of RAG rests especially on its ability to "field" hallucinations by quoting sources or letting the user know when it doesn't have the answer.

Consider these options:

Pre-train,
fine-tune,
RAG



Choosing the right way forward

There's no right way for everyone, so make sure you're checking your decisions against the goals outlined in previous steps — and that those goals are aligned with the concerned parties. Ultimately, you need a path you can clearly outline to your business stakeholders and your team.

Unless you're seeking a more extensive overhaul and larger-scale project, building an LLM from scratch and fine-tuning are too resource-intensive. You'll be inundated with questions: do you need a vector database to supplement your search engine? Can you upgrade your search engine, create and store embeddings in it, and build logic to continue to power your search? How do you expand this into a recommendations system?

For cloud-based solutions with hybrid search and semantic search, it can be pretty simple. By connecting to an existing LLM you can use RAG to create a more relevant search experience for your customers.

Here's what to consider:

- ||→ Take stock of what you already have in your IT environment. Oftentimes, rearchitecting the infrastructure isn't necessary at all.
- ||→ Consider out-of-the-box (OOTB) solutions like an OOTB LLM, an OOTB vector database, and an OOTB all-in package.

Pro: Black box technologies are likely to get you up and running faster.

Con: You can't scale in the same way because they're minimally customizable.

- ||→ Find a supplemental product that offers you flexibility and minimal disruption. You'll want to benchmark your search relevance and performance, and perhaps swap models to see which works best for you.

Step Try fast, fail fast

4

A rapidly evolving digital ecosystem means that there are a lot of moving parts in a generative AI project. There is only so much control that you can have on a pre-trained LLM, as well as a limited amount of flexibility to manipulate your architecture.

This is the time to take an iterative approach: you have your use case, you've established desired outcomes, you've set KPIs, and you've considered how to implement your generative AI project.



Remember

At its core, operationalizing generative AI is about getting answers from *your* data. Be sure to keep compliance top of mind. Are you setting up a test that may compromise your privacy policies? Is this a low-risk test?

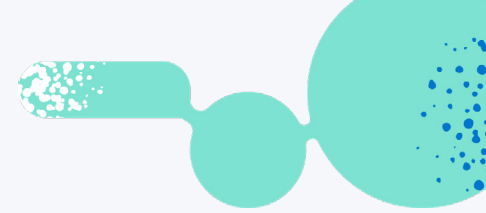


At this stage, you're looking to ...

- ||→ **Build a feedback loop:** Establish who reports what to whom and identify the key stakeholders in the project.
- ||→ **Enrich your LLM:** Ensure your LLM has access to the right information that's stored in a vector database. A vector database will allow you to quickly serve up the most relevant pieces of information to enrich your LLM.
- ||→ **Fine-tune the user experience:** Work on a user-friendly interface and keep testing it. Ultimately, generative AI is there to serve your employees and customers. Building an interface that is suited to the application and the user is critical to a successful generative AI project and ensures its scalability.
- ||→ **Establish a reference architecture that can scale:** While you're testing your generative AI project, keep an eye on the big picture. What will your architecture look like when you scale the project, and what will it look like when you expand into further use cases?

For example, if building a vector database from scratch feels like a heavy lift, you might look into a downloadable one — yes, they exist. With that vector database, you can unlock the next level: hybrid search. Using semantic search in your search applications lets you test your next-gen AI project prototype. This is an example of the power of starting small and iterating.





Step Governance and operations

5

Generative AI projects bring their own set of challenges — from data privacy and compliance to ethical considerations, quality control, and risk management. You need to anticipate potential obstacles and ensure that your project aligns with your business objectives.

As part of your governance and operations review, you'll need to consider a host of elements:

- ||→ **Cost management:** You get billed per thousand tokens; one cost for prompting, one for responses.
- ||→ **Logging:** You'll need to log every response to see the communication between your model and your customer for quality control.
- ||→ **Establish response sentiment:** Determine the sentiment of the LLM responses so they are on-brand with your company's tone of voice (another important quality control step).
- ||→ **Monitor for hallucinations:** Hallucinations include incorrect or misleading information but also can include hate speech and antisocial behavior from a chatbot.
- ||→ **Flag inconclusive answers:** Monitoring the quality and relevance of the responses is vital to quality control. This is an opportunity to understand which applications will require more human involvement than others and to plan accordingly when it's time to scale.

On bias in AI

Generative AI models rely on the data they are trained on. If the training data contains biases and limitations, these will be reflected in the outputs.

Organizations can mitigate these risks by carefully considering and limiting the data their models are trained on, or by using customized and specialized models specific to their needs. That said, the humans who program this technology or who curate the data the model is trained on also have biases.

Bias, like in any context, is difficult to eradicate. This doesn't mean organizations shouldn't try to address this challenge and educate users to exercise some critical thinking as part of the solution.

Additionally, expect involvement from your legal team, and make sure you factor their work into your proof of concept. Though their involvement might seem like it slows down a testing phase, it is crucial to establish review processes that are thorough and efficient for a responsible, ethical, and compliant implementation.

The thing about data safety

With security threats affecting organizations every day, data safety is paramount. Your customers trust you with their data — which is why many companies operate from a zero trust framework. This takes the principle that users and devices should never be automatically or implicitly trusted, whether inside or outside an organization's network perimeter.



To optimize security, you can:

- 1. Take a RAG approach:** RAG models leverage retrieval mechanisms to better understand the context of the input prompt, which leads to more contextually appropriate responses that omit sensitive details. Using RAG with a data platform that has document-level and role-based security will ensure that permissions are respected.
- 2. Invest in or expand your observability solution:** Address the trust question. Follow the data trail with monitoring capabilities, and monitor the responses produced by your generative AI project. Where is your data going, and what is generative AI saying to your customers?

Ultimately, bringing generative AI into your ecosystem will require that you establish new operations protocols and accordingly, new policies. With more efficient processes and higher revenues, the time saved performing mundane tasks can be redirected to these efforts.

Step Set a timeline. Give it benchmarks.

6

Outline a time frame — let's say one quarter. Within that time frame, set goalposts at the 30-day and 90-day marks. Use the quarter to prove the value of your generative AI-boosted use case.



By day 30, you'll want to have launched your first test. What does that look like?

- You've chosen a use case
- You've assigned a small team to the task
- You've facilitated training sessions as necessary
- You've established desired outcomes
- You've built a prototype interface



By day 90, you'll be ready to launch your first use case. What does that look like?

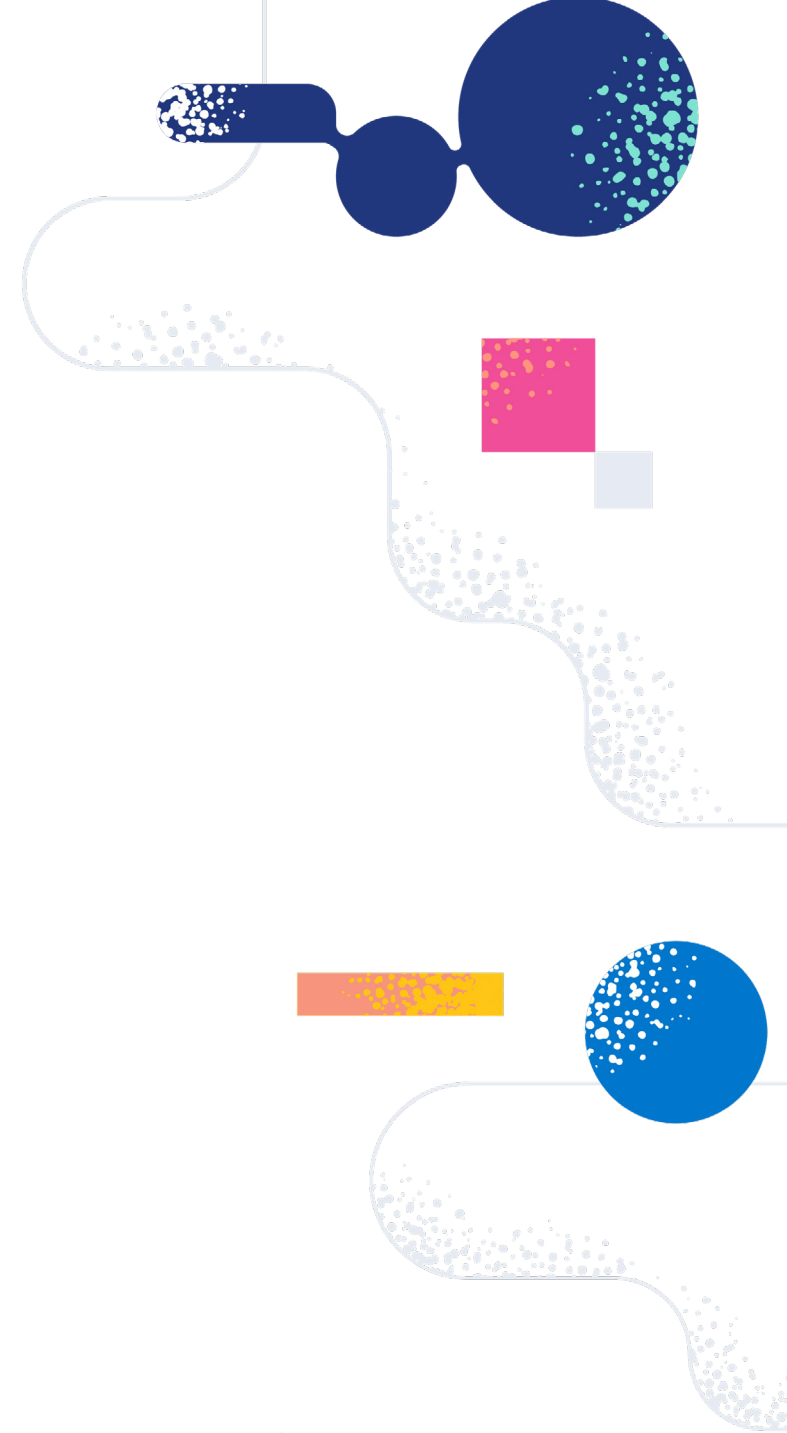
- You've opened up the test to a few internal people
- You've tested, tuned, and measured the outputs generated
- You've continually monitored the way users interact with the interface
- You've established a set of guidelines for what constitutes quality outputs
- You've collected data on some key performance metrics
- You've measured the value of the initiative

These tasks should serve as rough benchmarks. Your company's specific needs — the makeup of your team, and the tech they're working on or adding to your stack — will affect the speed at which you can deploy your first use case and gather insights.

At this stage, consider:

- 1. The error rate:** Measure the error rate. Is the generative AI producing correct and relevant outputs? This is crucial to fine-tuning the generative AI.
- 2. Training time and cost:** Measure the time and resources needed to train your model. Doing so helps ensure an efficient testing period, and therefore a faster time to operationalize.
- 3. Human intervention:** Does the generative AI perform only with a human-in-the-loop? How much oversight is required to maintain reliability and accuracy?
- 4. Response time and quality of outputs:** Measure how quickly the generative AI provides outputs and compare the quality of the outputs against a set of established rules or guidelines.

And just like that, you'll be ready to operationalize and scale your success.



The start of a new era

So many industry leaders are already seeing the benefits of generative AI — and many more are trying to replicate that success and keep up with shifting customer expectations. Innovation in generative AI is moving fast. But you can't go anywhere without the basics.

Strategizing the most aligned ways to implement generative AI can help you harness the powers of your data without getting distracted by exciting but irrelevant innovations. To most effectively operationalize generative AI, dedicate time and resources to implementing it in stages. Integrating new technology with purpose is the recipe for the highest return on your investment. What's more, customizing and adapting AI tools to meet your operational needs ensures relevance and effectiveness — which is generative AI's "calling" in the first place.

Keep in mind the need to responsibly implement generative AI, from data privacy protection and security to sensitivity and ethics. Aside from adding up to trillions of dollars in value to the global economy, generative AI presents an opportunity to democratize the workforce and to upskill workers. Position yourself not only as a generative AI trailblazer but also as a pioneer in the development of new business processes for your company.





Now let's get started.

Identifying your first generative AI use case is a cross-team endeavor. You need your security team, IT team, dev team, and line of business team to work together — on day one. **Here's how Elastic can help:**

With your security team

You can boost practitioner productivity and reduce risk. Operationalizing generative AI for your security use cases starts with [a unified approach on an open platform](#). You can harness the power of generative AI with Elastic Security to create an experience that's tailored to your security team's needs.

Meet Elastic Security

With your SRE and IT operations team

Equip your SREs and engineers with the power to utilize an interactive natural language chat interface that allows them to zero in on the most relevant information, faster. Discover how you can combine conversational AI with Elastic Observability and advanced machine learning for a [context-aware interactive chat experience](#) based on your proprietary data and runbooks.

Meet Elastic Observability

With your dev team

You can enhance your dev team's toolkit to help them provide customer support through self-service options — like through highly personalized chatbots with Elastic Search. Empower your customer service agents with the same great search tools to solve cases quickly — including generative AI experiences that help them find answers from disparate sources of data. Discover how to [implement powerful search for your knowledge base](#).

Meet Elastic Search

