elastic search labs

The Search AI Developer Survey

Search and AI developer insights

This report looks at the state of search experiences and trends, with a particular eye on generative AI.

- We learned what factors were driving development around keyword-based and semantic search, model selection and use, approaches to domain-specific data, and the favored tools and integrations.
- Several key use cases were highlighted, mostly centered around customer service, and participants cited their most common challenges with AI, such as data readiness and security.
- We also delved into how our customers were using our flagship product, Elasticsearch, to build AI search experiences.

 | © 2024 Elasticsearch B.V. All rights reserved.

Survey methodology

The survey was conducted in 2024 by Dimensional Research, on behalf of Elastic. The survey went to both Elasticsearch users and external participants. Topics included questions about search environments, AI, and Elasticsearch use.

Survey participants

Over 800 participants completed the survey. All participants had responsibility for search technology in a professional capacity. A wide range of job levels, roles (engineer, architect, product manager, DevOps, data science, admin/operator, other), industries, regions, search responsibilities, and company sizes were represented.

In this report...

Excitement for GenAI, semantic search approaches, and RAG

Variety of use cases, many of them customer service-related. **The three most frequently cited center around knowledge self-service.**

An abundance of caution and variety of challenges, many privacy-related. **Only 11% of respondents have a GenAI use case in production.**

How your peers are building search: patterns on ingestion, languages, and querying.

The most popular tools, integrations, and models used by your peers. **OpenAI's tools are being evaluated by over 70% of respondents.**

 | © 2024 Elasticsearch B.V. All rights reserved.

Why we ran this survey

We ran this survey to understand how generative AI and search tools are transforming the role of developers in delivering next-generation search applications.

Developers are prototyping and taking generative AI to production. They are being challenged to try and evaluate different models, vector databases, frameworks and more. Developers are also eager for AI to automate repetitive tasks, freeing up time for more creative and complex problem solving.

"[For a 1,000-line code review], which would take a human maybe a week to do, it can do it literally in one minute or less. So we are projecting that our code development process can be reduced by as much as 20% to 30% by using generative AI."

VP of IT
Retail firm in the U.S.

"Relevance is paramount, especially given the continuous flow of new reports. [We need to] consider how we would ground any future generative AI services. The quality of the output, in that regard, is largely influenced by the quality of the retrieval."

AI Strategy Specialist
Leading market research firm

 | © 2024 Elasticsearch B.V. All rights reserved.

I.

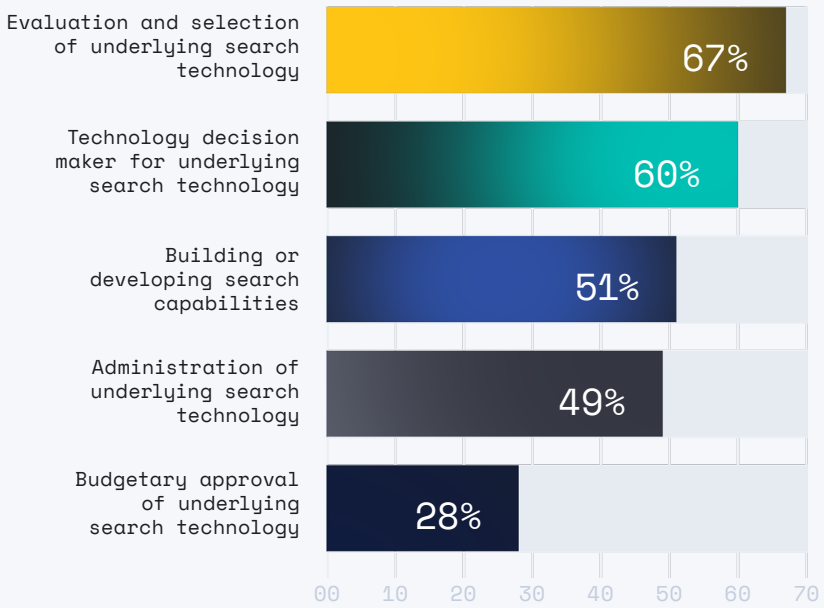
Survey Overview

 | © 2024 Elasticsearch B.V. All rights reserved.

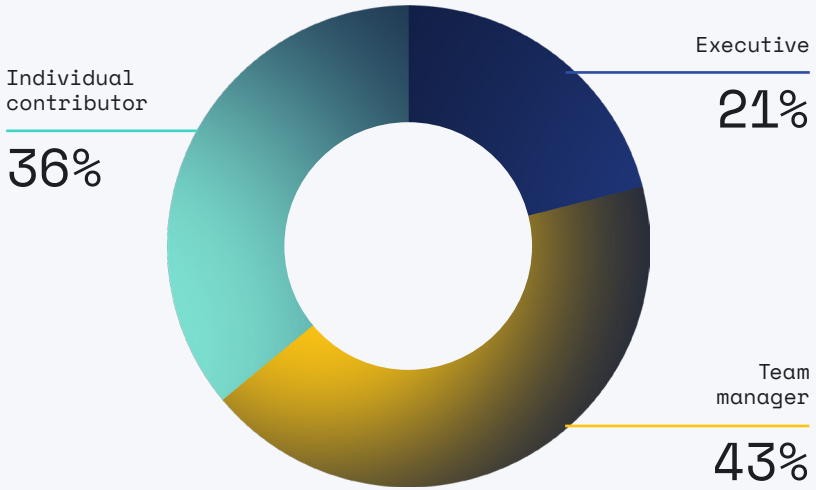
800+ developers and tech professionals surveyed

Survey respondents spanned a number of roles beyond just developers, including data scientists, DevOps pros, and product managers. Many of the respondents held responsibilities beyond coding, including evaluation, selection, and budgetary approval of search tools.

Responsibility for Search

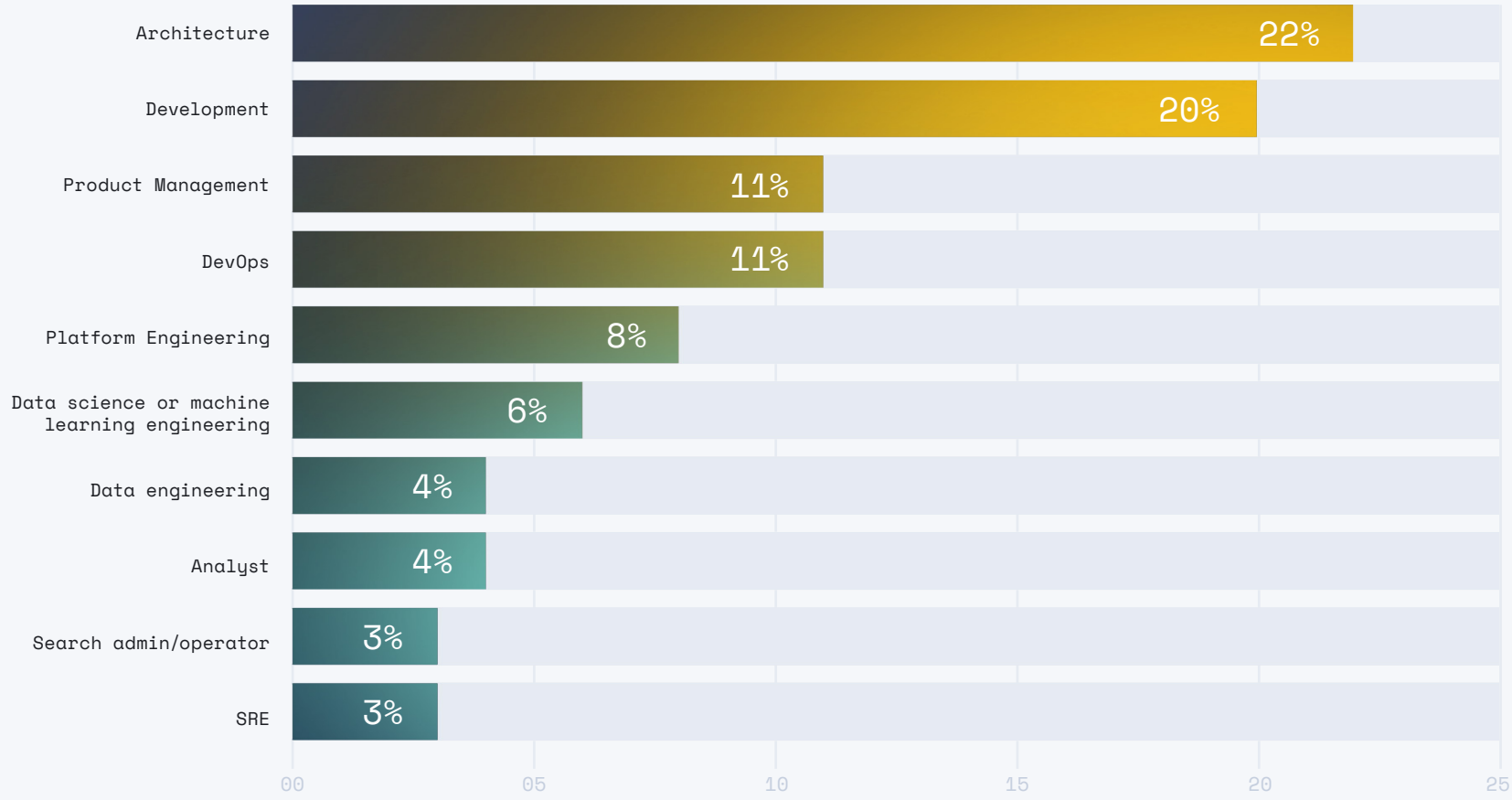


Job Level



 | © 2024 Elasticsearch B.V. All rights reserved.

Role



 | © 2024 Elasticsearch B.V. All rights reserved.

II.

GenAI for your applications: Uses and challenges

Key takeaways:

- The survey highlights anticipated use cases, such as customer self-service or chatbots, which help reduce operational costs and improve customer satisfaction.
- However, respondents also cite challenges such as ensuring accuracy of AI responses, maintaining data privacy, and complexities of integration with existing systems.

 | © 2024 Elasticsearch B.V. All rights reserved.

A few notes on terminology

Retrieval augmented generation

Retrieval augmented generation (RAG) is an approach in natural language processing that combines the strengths of information retrieval and text generation. In a RAG workflow, a retrieval technology plays a crucial role by first retrieving relevant documents or data based on the query. This retrieved information is then used as context to generate more accurate, informative, and contextually appropriate responses or content. RAG enhances the quality and relevance of generated content by grounding it in real-world information sourced by the retrieval system.

Keyword search

Keyword search is a search method that retrieves documents or information by matching user queries with exact words or phrases in the text.

Vector search

Vector search uses mathematical vectors to represent data points and retrieves information based on the proximity of these vectors in a multidimensional space, often used for more complex and nuanced searches, involving multi-modal data (text, images, video, and more).

Semantic search

Semantic search focuses on understanding the meaning and context of search queries and content, allowing for more accurate and relevant results by interpreting the intent behind the words.

Hybrid search

Hybrid search combines multiple search techniques, such as keyword search and vector search, to leverage the strengths of each method and provide more comprehensive and relevant results.

 | © 2024 Elasticsearch B.V. All rights reserved.

Do you agree or disagree with the following statements:

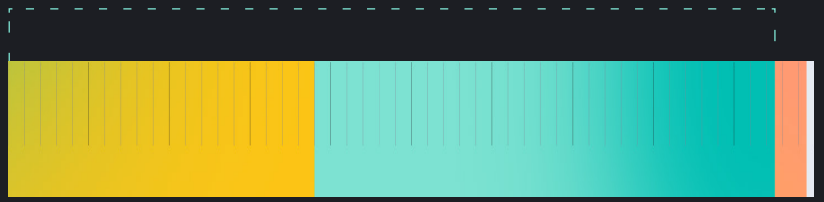
- Strongly agree
- Agree somewhat
- Disagree somewhat
- Strongly disagree

Generative AI is important for our business, but needs to be augmented with our business data to deliver maximum business value

95% agree internal business data is needed to deliver maximum value from generative AI

Retrieval augmentation technologies (that use a search engine for proprietary data) are important for getting value out of generative AI

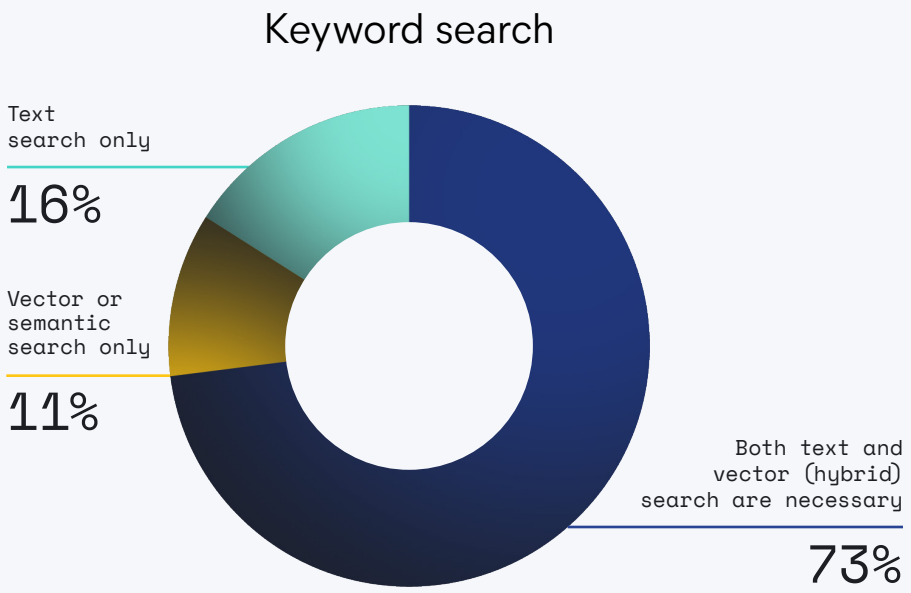
93% agree that retrieval augmentation technologies are important to get value from generative AI



 | © 2024 Elasticsearch B.V. All rights reserved.

73% see value in hybrid search for GenAI

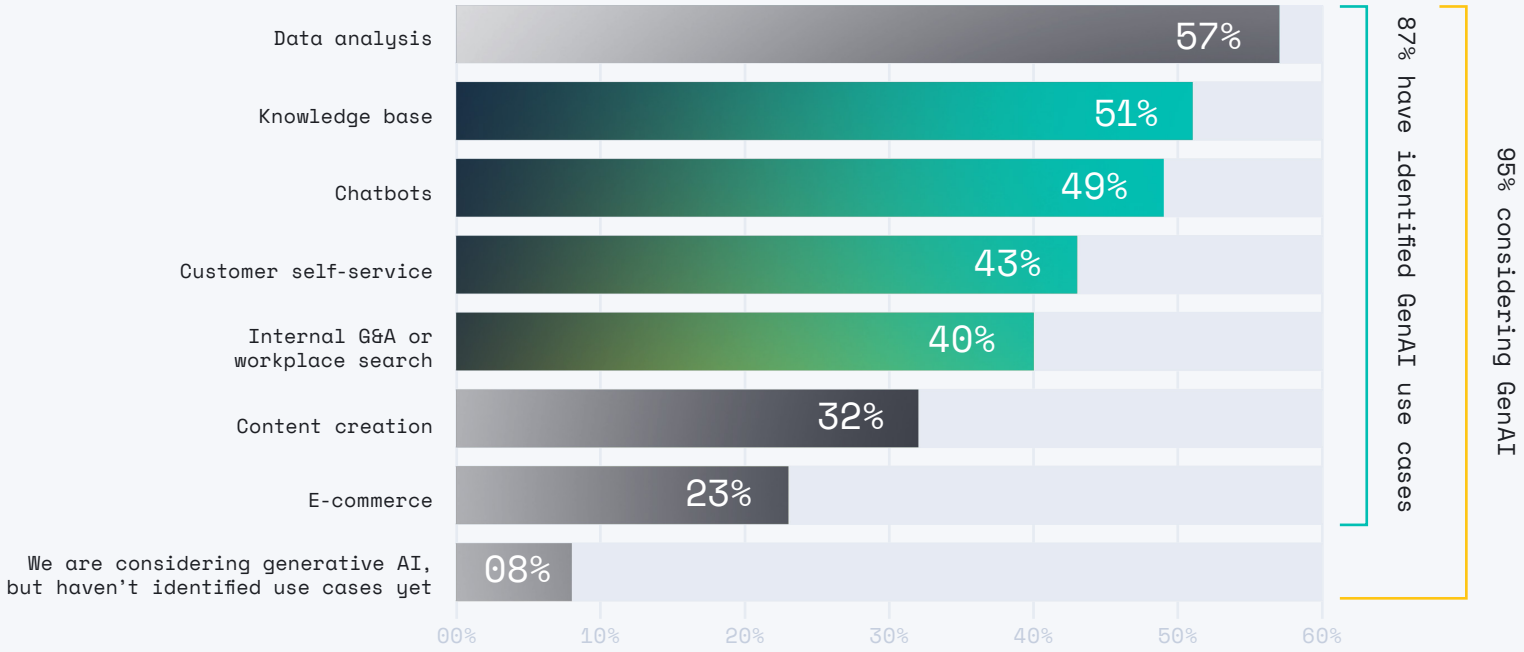
What types of search capabilities are important to ensure internal data sources are included in your company's generative AI search strategy?



 © 2024 Elasticsearch B.V. All rights reserved.

95% are implementing or considering GenAI *87%* have already identified use cases

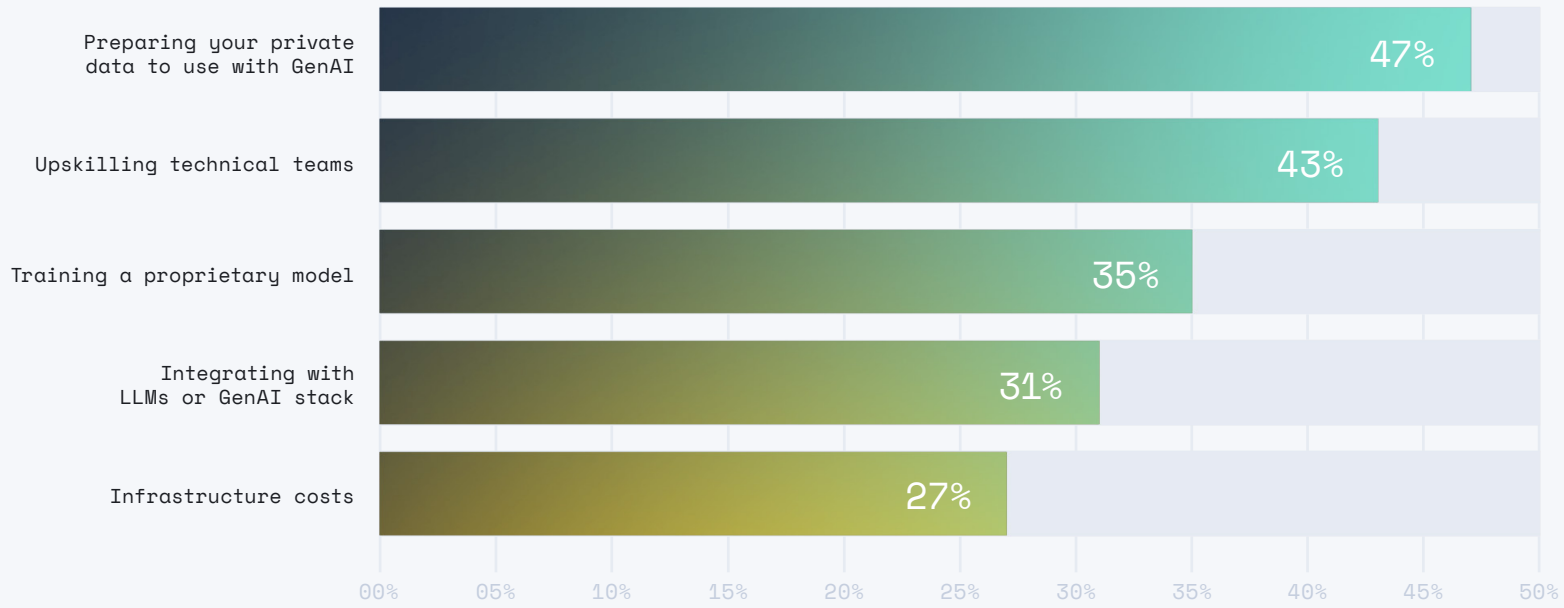
What types of generative AI use cases is your organization implementing or considering creating apps for?



 | © 2024 Elasticsearch B.V. All rights reserved.

The *most time consuming tasks* when building GenAI use cases

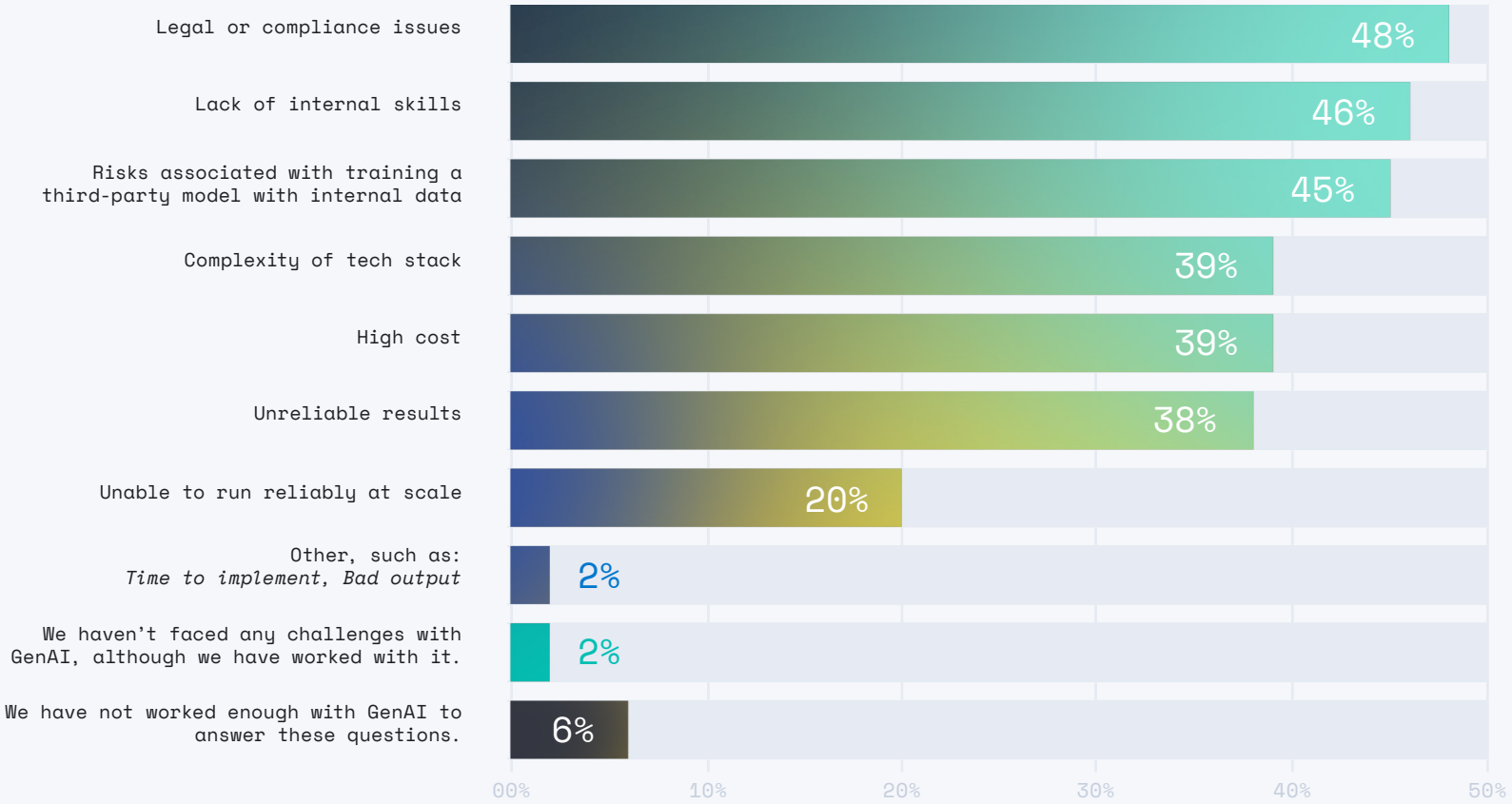
Where do you expect that the MOST time and resources will be spent when building generative AI use cases for your applications?



 © 2024 Elasticsearch B.V. All rights reserved.

List of top GenAI *challenges:* compliance, data risks, and upskilling teams

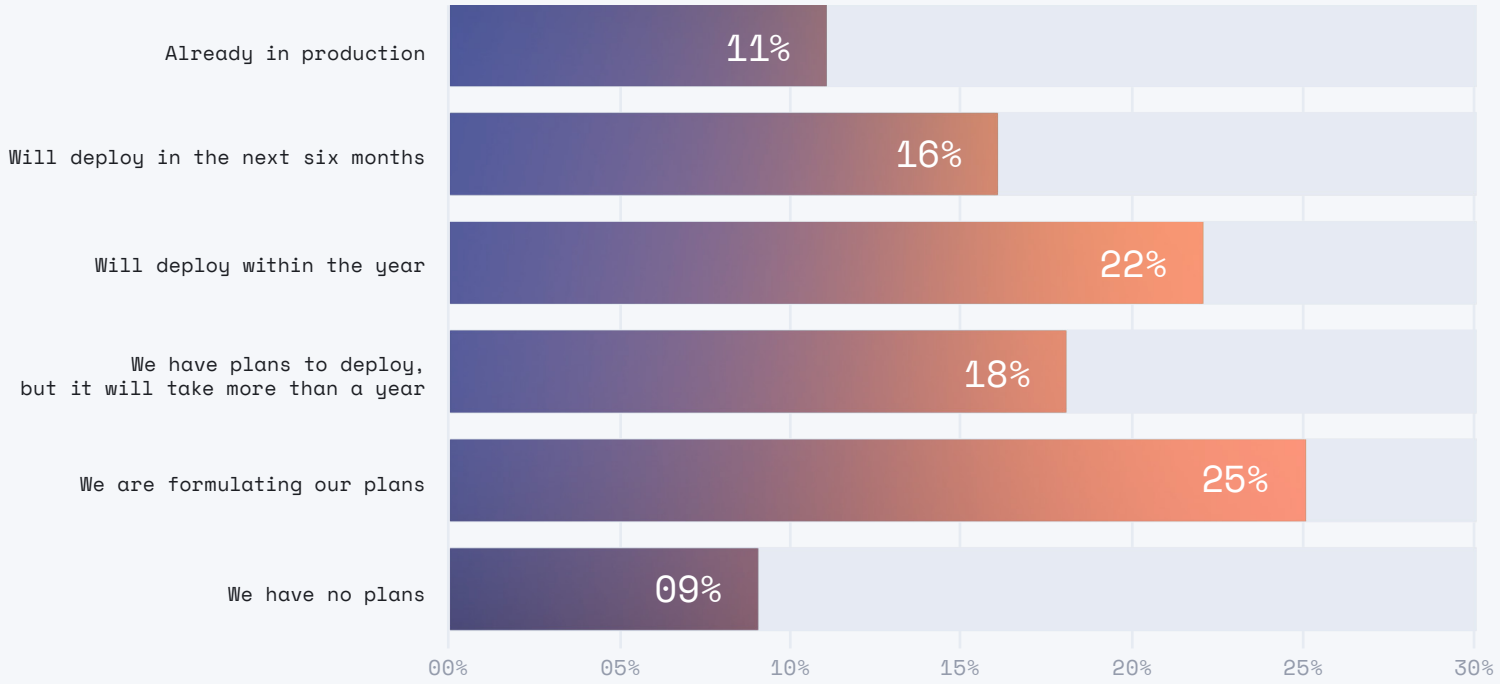
What challenges has your organization experienced or expects to experience when working with generative AI use cases?



 | © 2024 Elasticsearch B.V. All rights reserved.

Companies considering GenAI are also planning RAG-based GenAI, *but few are in production*

What is the status of your organization's plans for search based on generative AI with retrieval augmented search for internal sources?



 | © 2024 Elasticsearch B.V. All rights reserved.

III.

GenAI and vector search technologies

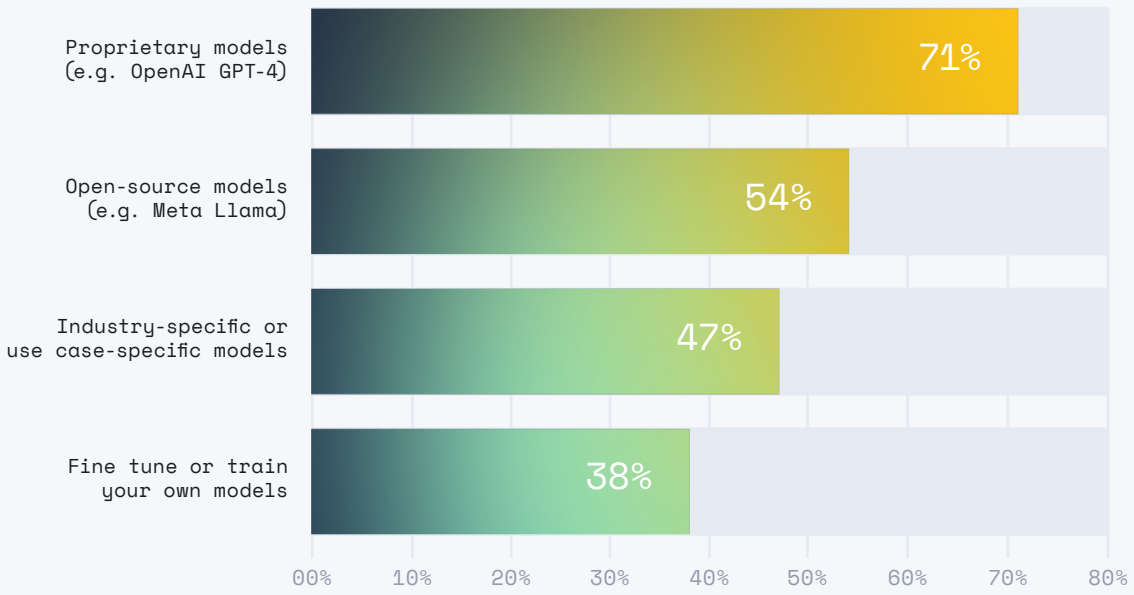
Key takeaways:

- Respondents are assessing a number of technologies, such as advanced NLP frameworks and large language models (LLMs).
- Users value the importance of tools that facilitate integration of AI into existing systems. Scalability, performance, and ease of use are also important criteria.

 | © 2024 Elasticsearch B.V. All rights reserved.

More organizations are using or plan to use proprietary models than other options

What types of models does your organization use or expect to use in the future?



elastic search labs

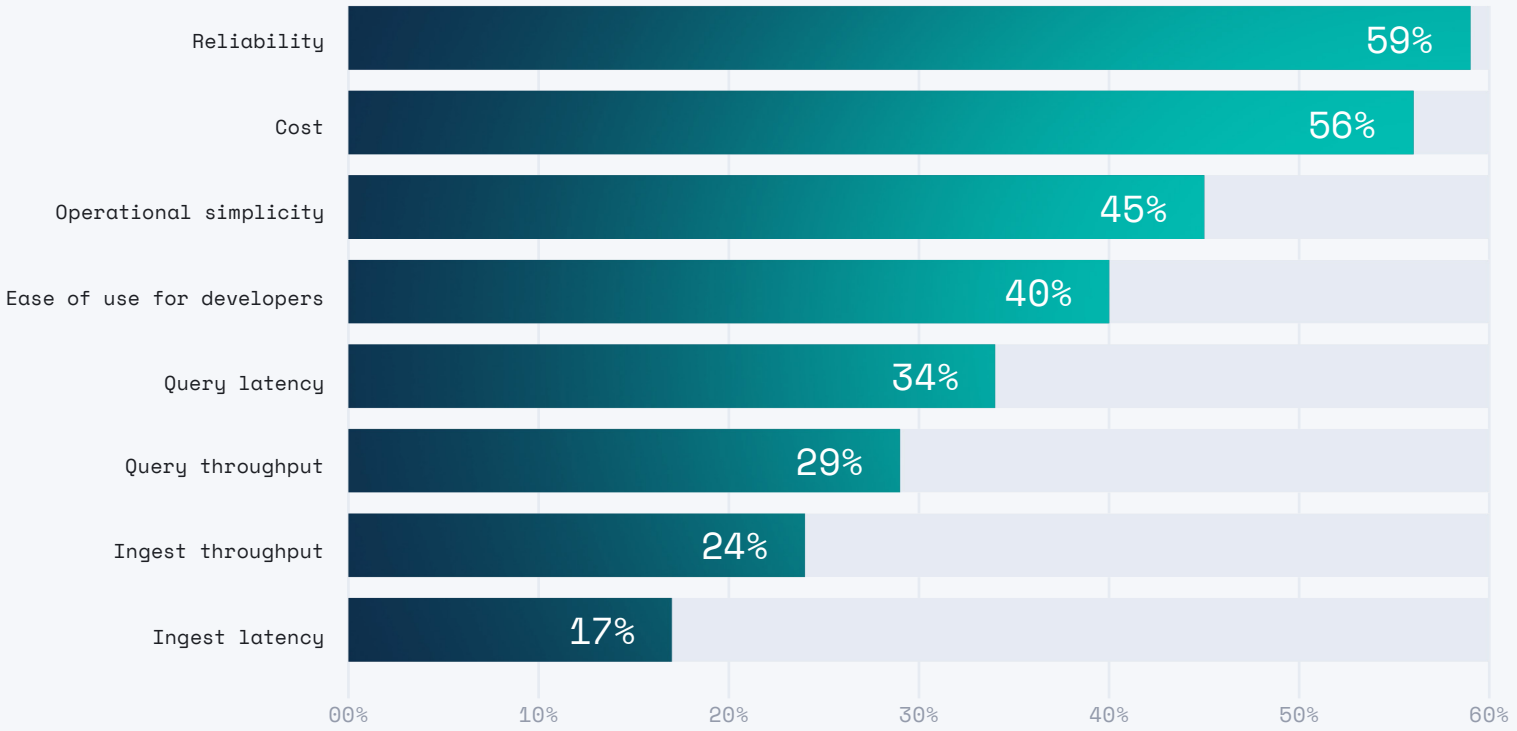
GenAI blogs and tutorials for developers

Blog in focus:
Fine tuning vs. pre-training vs. RAG

 © 2024 Elasticsearch B.V. All rights reserved.

Selecting a vector search engine: reliability, cost, ease of use over speed and throughput

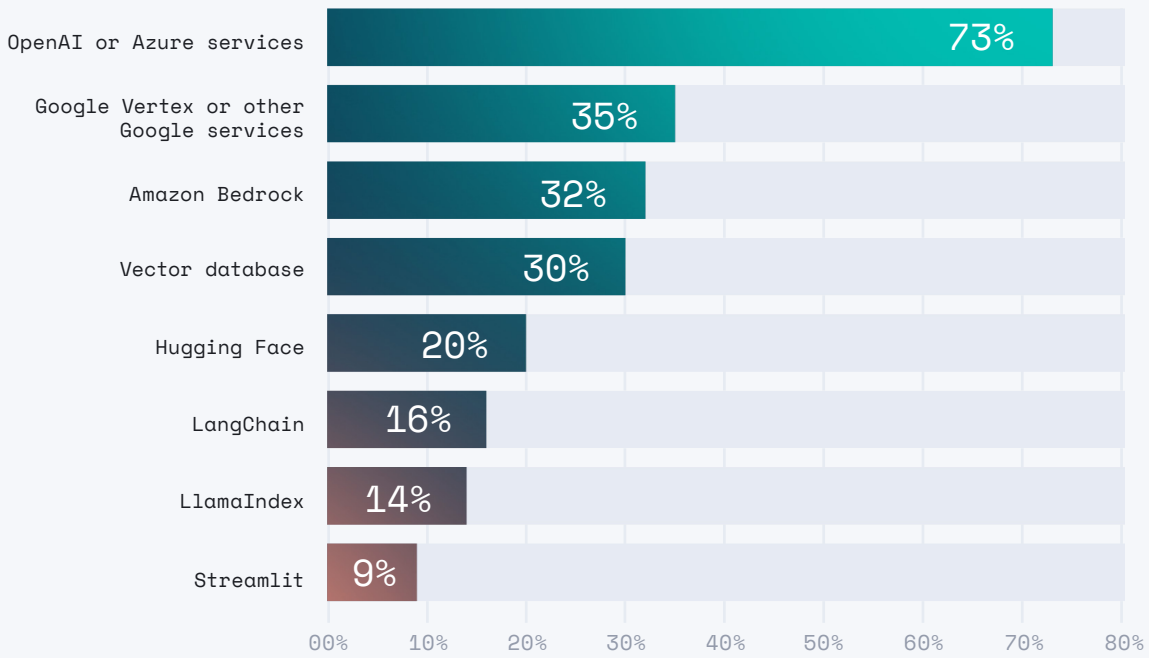
(Participants were asked to choose their top 4 considerations)



 | © 2024 Elasticsearch B.V. All rights reserved.

In addition to the model providers, frameworks gaining popularity

Which of these technologies are part of your organization's current or planned generative AI search application stack?



elastic search labs

GenAI blogs and tutorials for developers

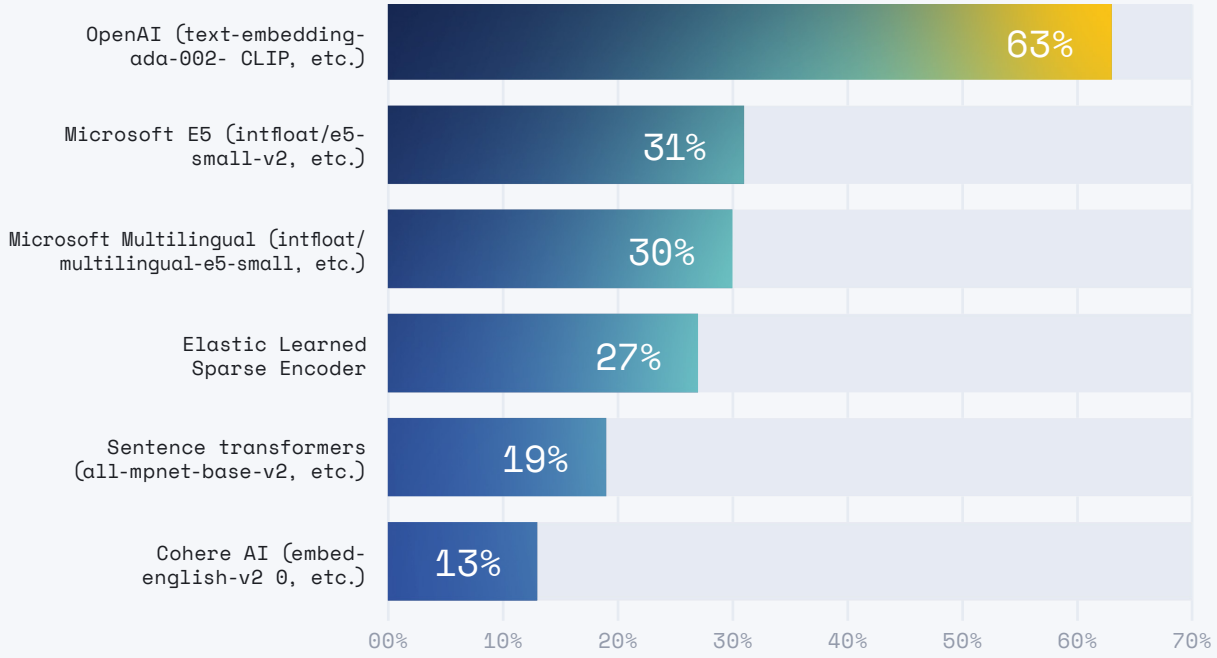
Elasticsearch works with Amazon Bedrock, Azure OpenAI, Cohere, Hugging Face, Google Gemini, LangChain, LlamaIndex, Mistral, and more.

Visit elastic.co/search-labs/integrations

 © 2024 Elasticsearch B.V. All rights reserved.

83% have used or plan to use embedding models

Which of these embedding models has your organization used, evaluated, or has plans to evaluate?



elastic search labs

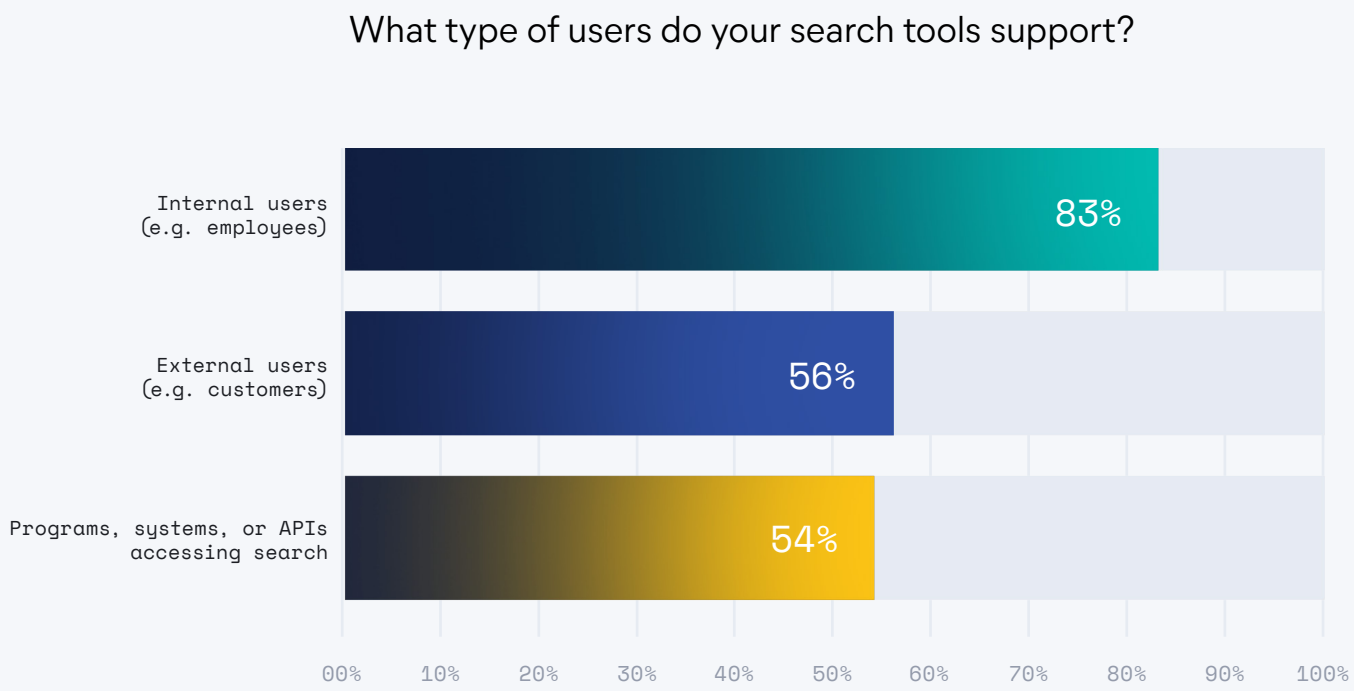
GenAI blogs and tutorials for developers

Read our latest ML research and how we trained our sparse encoder model — ELSER.

Visit elastic.co/search-labs/blog/category/ml-research

 © 2024 Elasticsearch B.V. All rights reserved.

Use cases support a *variety of user types*



 | © 2024 Elasticsearch B.V. All rights reserved.

IV.

Search with Elasticsearch

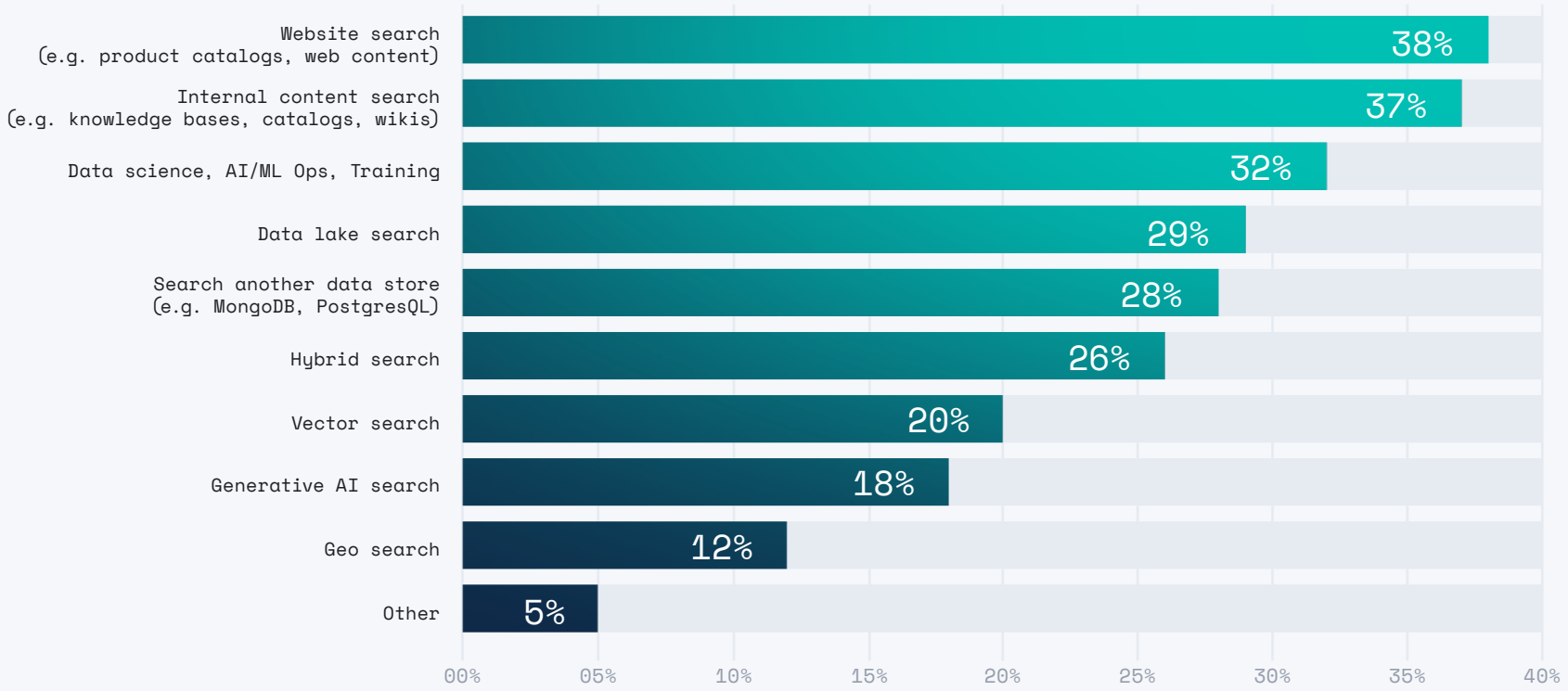
Key takeaways:

- In a short span of time, vector search and hybrid search have become important use cases for Elasticsearch users.
- Elastic's wide range of ingest options have equally wide adoption. Are there ingest tools on that list you haven't considered before?

 | © 2024 Elasticsearch B.V. All rights reserved.

How our customers use search: *A variety of search use cases* with Elastic

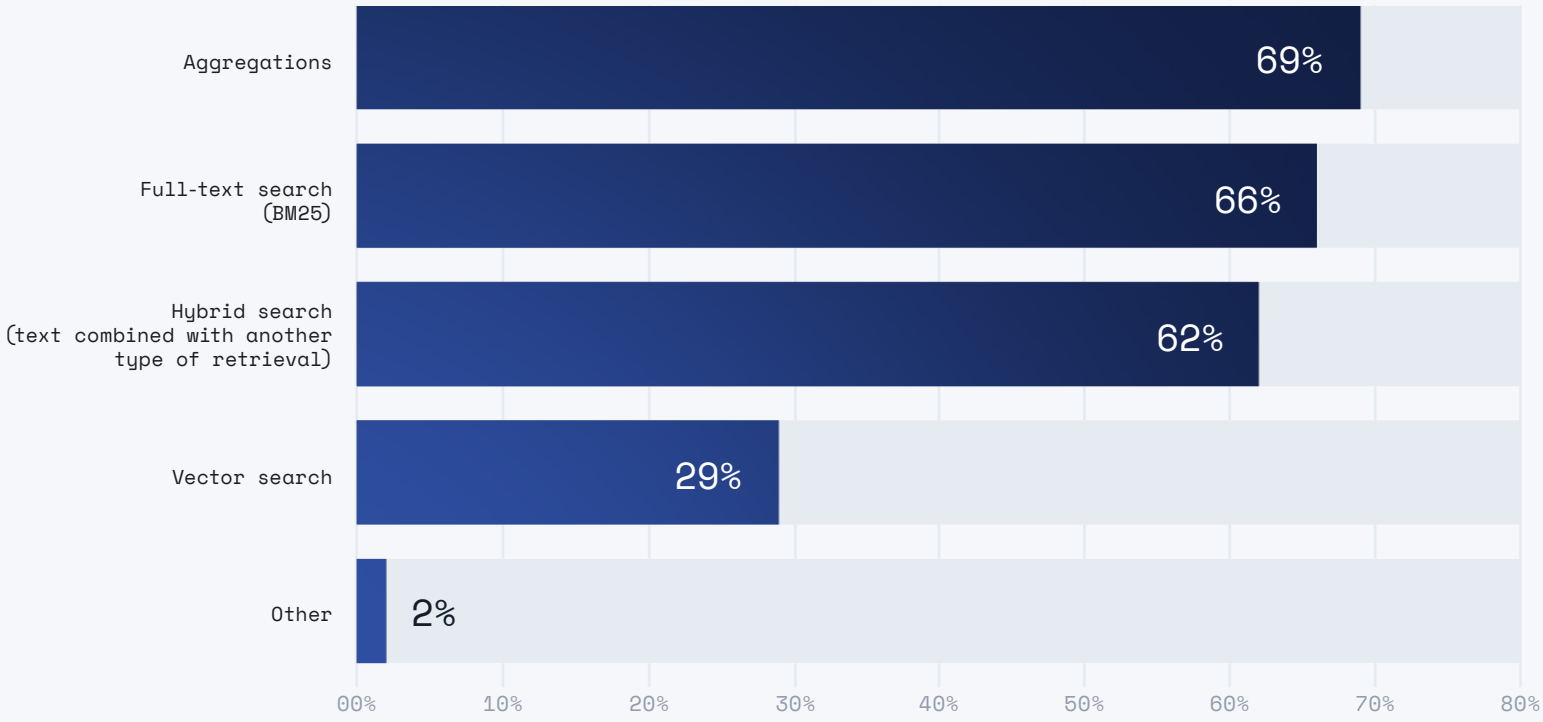
Fast-emerging use cases: Hybrid search, vector search, GenAI search, data science/MLOps/training



 | © 2024 Elasticsearch B.V. All rights reserved.

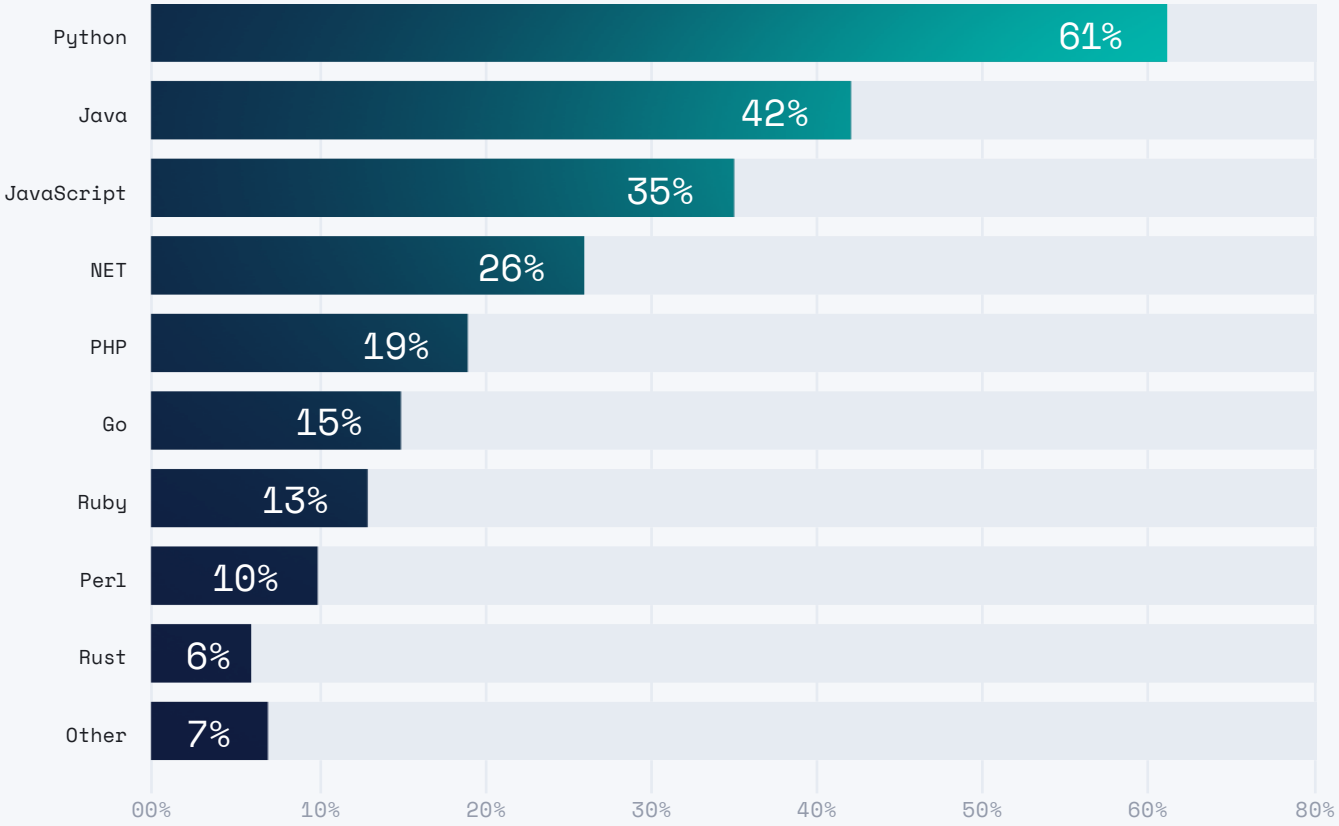
Elasticsearch is used for a *wide range of queries*

What type of queries do you use Elasticsearch for?
(Participants were asked to choose all applicable considerations)



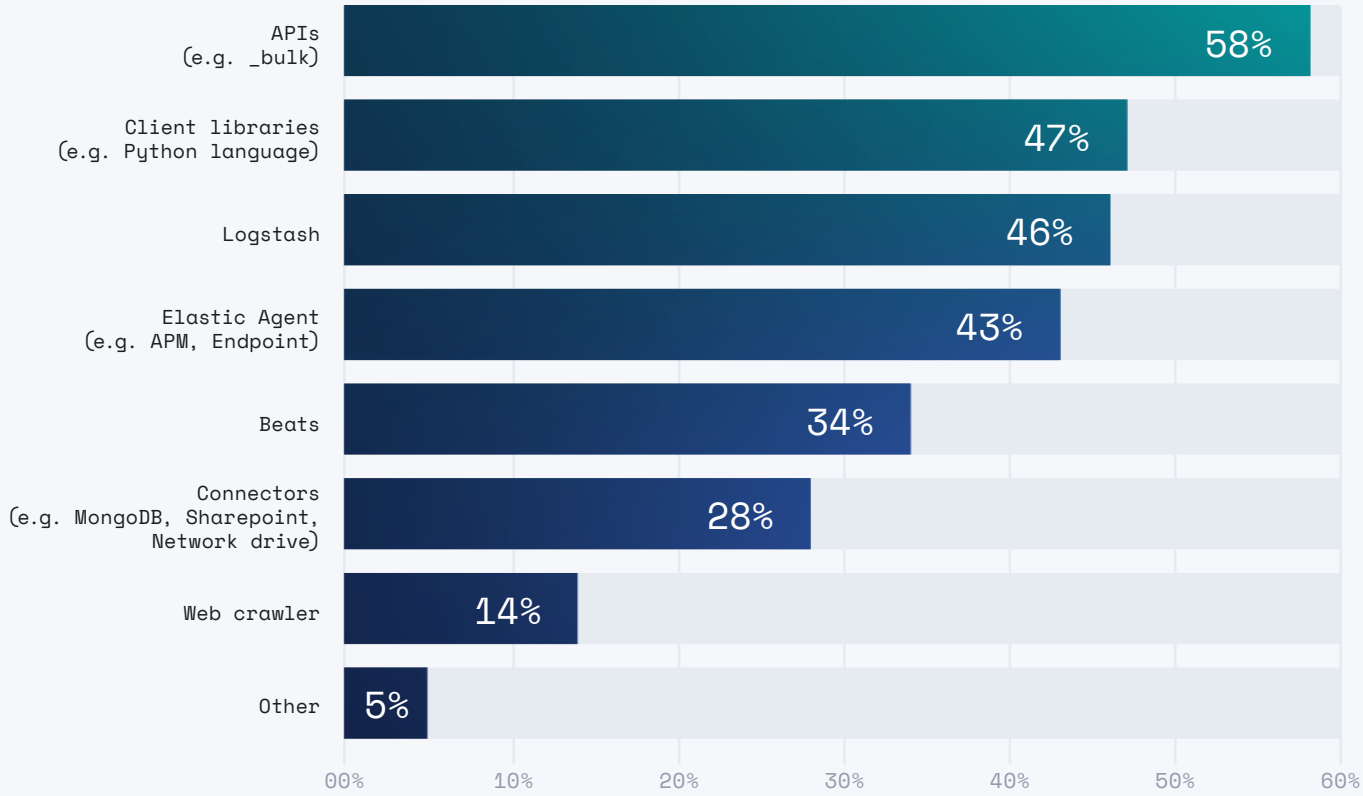
 | © 2024 Elasticsearch B.V. All rights reserved.

Which programming language do you use with Elasticsearch?



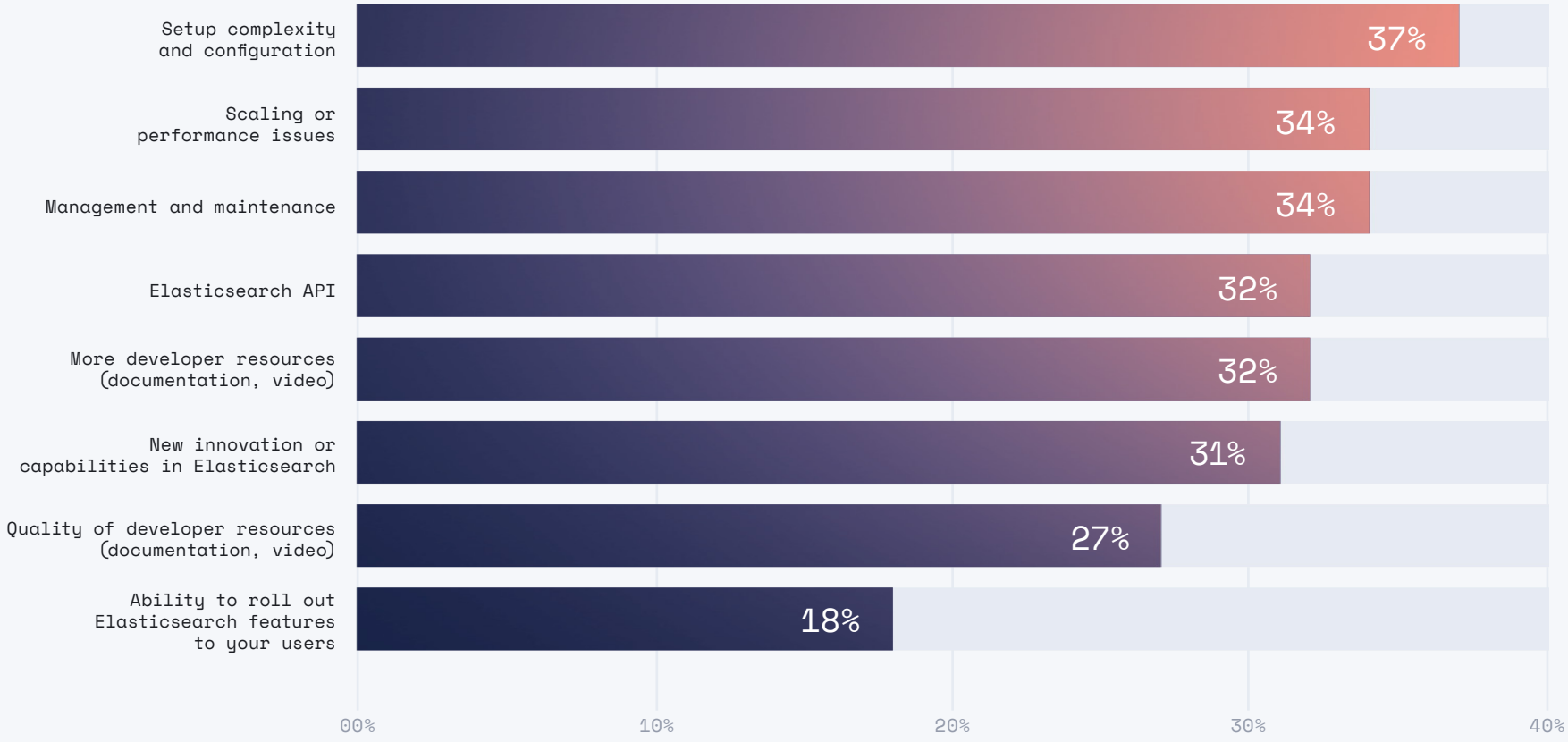
 | © 2024 Elasticsearch B.V. All rights reserved.

How do you ingest data into Elasticsearch?



 | © 2024 Elasticsearch B.V. All rights reserved.

What areas of Elasticsearch would you most like to see improved?



 | © 2024 Elasticsearch B.V. All rights reserved.

We're acting on community feedback!

Elastic Cloud Serverless

Developers want an Elastic experience that simplifies setup complexity, reduces performance issues, and management overhead.

Elastic Cloud Serverless provides the same Elastic experience you expect, without managing the underlying clusters, nodes, data tiers, or scaling. Serverless instances are:

- fully-managed
- autoscaled
- automatically upgraded by Elastic

...so you can focus more on gaining value and insight.

Visit elastic.co/elasticsearch/serverless

elastic search labs

Developers want more (and better) resources and latest innovations in search technology.

Search Labs is a blog and a GitHub repo from the Elastic engineering team, where we write about upcoming features and provide best practices on building modern search solutions.

Visit Search Labs for:

- design rationale and details of new innovation from the Elastic team
- best practices, how-tos, and guides
- integrating tools from Cohere, Azure OpenAI, and more into your RAG workflows

Visit elastic.co/search-labs

 © 2024 Elasticsearch B.V. All rights reserved.

elastic search labs

What next?

Daily breakthroughs in LLMs and generative AI have put developers at the forefront of the movement, influencing its direction and possibilities.

Elasticsearch enables developers to build modern search applications for RAG workflows. It includes a vector database and open inference capabilities for leveraging embeddings, reranking, and completions to help developers accelerate and scale generative AI experiences.

Ready to start testing and building?

Learn how to create advanced search experiences with generative AI, embedding models, reranking capabilities, and more.

Search Labs has guides, integrations, notebooks, and example apps — everything you need to research and start building with Elasticsearch.

Visit Search Labs

 | © 2024 Elasticsearch B.V. All rights reserved.