Couchbase with UserEvidence

THE STATE OF ENTERPRISE AI DEVELOPMENT:

Implementation Insights & Architectural Realities

*&

Contents

Executive Summary

Enterprise adoption of artificial intelligence has reached an inflection point. **Deloitte** found that most advanced generative AI (GenAI) projects are delivering ROI that meets or exceeds the expectations of 74% of business leaders. This positive return is helping to drive rapid adoption, as organizations across industries increasingly deploy GenAI solutions. Reflecting this surge in adoption, **Gartner** predicts that global GenAI spend will hit $644 billion in 2025, an increase of 76.4% from 2024. Similarly, **IDC** expects AI investment will reach $632 billion by 2028.

As investment grows, many are actively developing agentic AI solutions, which have the potential to dramatically improve productivity and create a foundation for more complex technical projects. Gartner expects the technology will be integrated into 33% of enterprise software applications by 2028—an increase from under 1% in 2024—allowing 15% of daily work decisions to be made without human intervention. Yet serious issues like managing hallucination concerns and standardizing data architecture currently compromise productivity.

This report explores how organizations today are approaching AI implementation and agentic applications. The survey data reveals that many prioritize speed and gains in the short term, which may create substantial challenges as AI implementation scales.

The findings reveal a pattern of high confidence coupled with immature implementation practices. This disconnect between stated expertise, real concerns, and actual architectural foundations suggests that many enterprises are building AI applications on unstable ground.

The implications extend beyond technical implementation to business strategy. As organizations race to deploy autonomous agents and agentic systems, those that fail to address underlying data architecture may find their AI initiatives delivering diminishing returns rather than the transformative outcomes they expect.

“We’re at a pivotal moment in the evolution of enterprise AI. The rapid adoption of generative AI signals a wave of transformation already underway. But the real breakthroughs will come when organizations address the data complexity issues behind these systems. Retrieval augmented generation (RAG) offers a clear path to safer, more reliable AI, but its effectiveness depends on supporting the entire text-heavy, RAG data lifecycle—at millisecond speed. This is most easily done with JSON data as the common denominator through a unified developer database platform. The future is agentic, but realizing that future requires us to get serious about data complexity, speed, architecture, and trust.”

Mohan Varthakavi,
VP of Software Development, AI, and Edge at Couchbase

Key Takeaways

Chapter 1

- AI adoption is accelerating rapidly, and optimism rules the day. 38% of organizations are already deploying GenAI in production, with software development emerging as the primary use case driving immediate value.
- Concerns about data privacy and large language model (LLM) hallucinations are high. Addressing them requires understanding and deploying retrieval augmented generation (RAG) workflows and their data utilization in each stage of RAG.

Chapter 2

- Coding assistants make developers more productive, while enterprises are widely deploying GenAI chatbots. Although organizations are deploying fewer agentic systems, confidence is strong that this will change by 2027.
- Data architecture complexity is clearly an impediment to facilitating RAG and GenAI usage for both analytics and operational AI-powered applications and agentic systems.

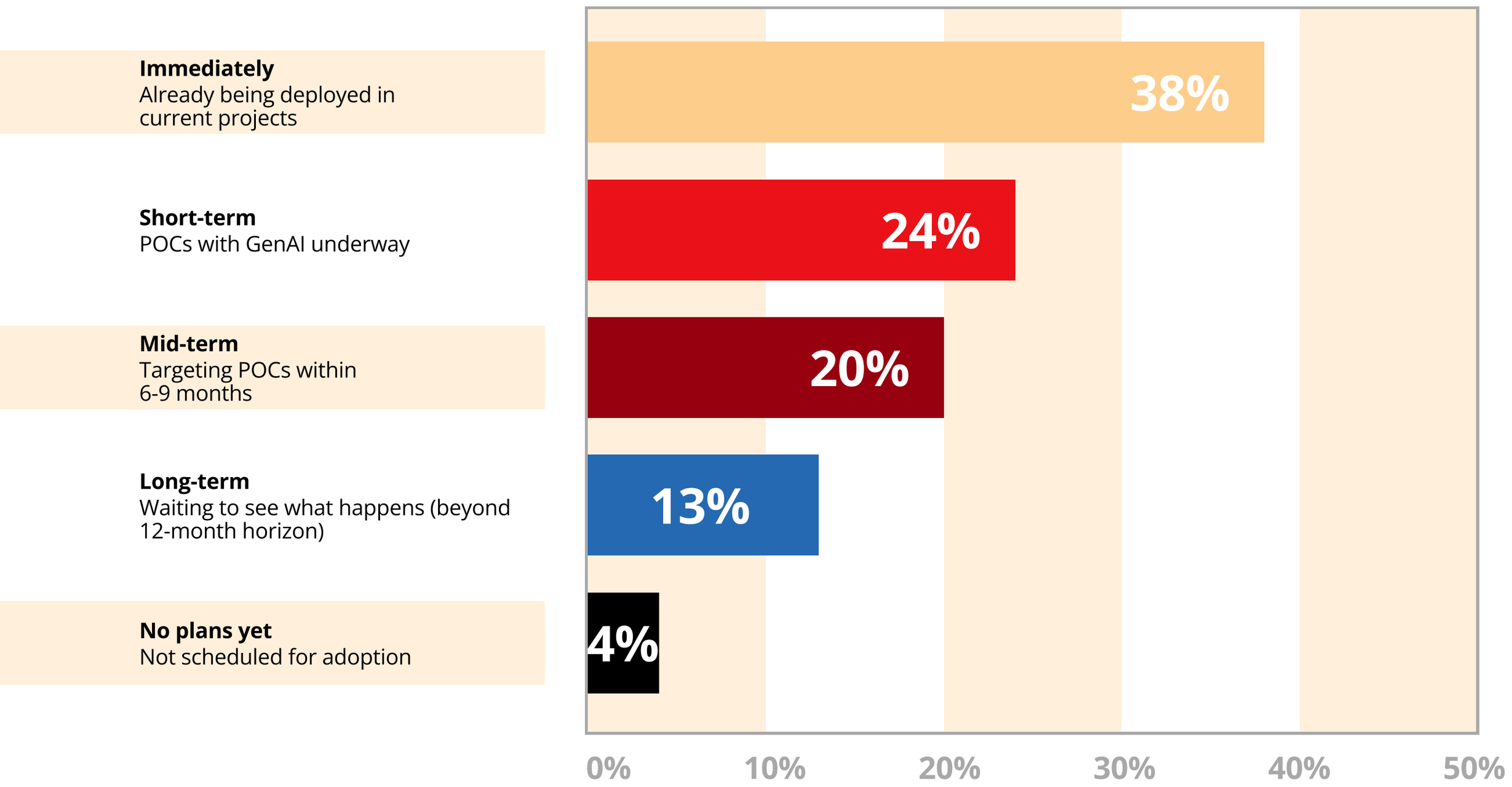
Chapter 3

- RAG can resolve hallucination fears without compromising enterprise data, but managing the process is complex and currently unruly.
- Confidence in RAG is high, but experience with LLMs and RAG is immature. 49% of respondents are only using ChatGPT models in their work, for example. Only 29% of respondents are using a unified data platform to manage AI data requirements.
- Managing data consistency within RAG workflows is inconsistent, not unified, and therefore not ready to facilitate changes as RAG-powered systems evolve.
- Guardrails are in place, but only 35% consider themselves to have comprehensive guardrails for managing drifting behavior of agents.

Chapter 1:
GenAI Motivations and Concerns

The data is clear: GenAI implementation is moving forward rapidly. 38% of respondents are already deploying this technology in current projects, and an additional 24% are executing proofs of concept (POCs). That equals 62% of organizations already engaged with GenAI.

How quickly are you adding GenAI into your application delivery plans?

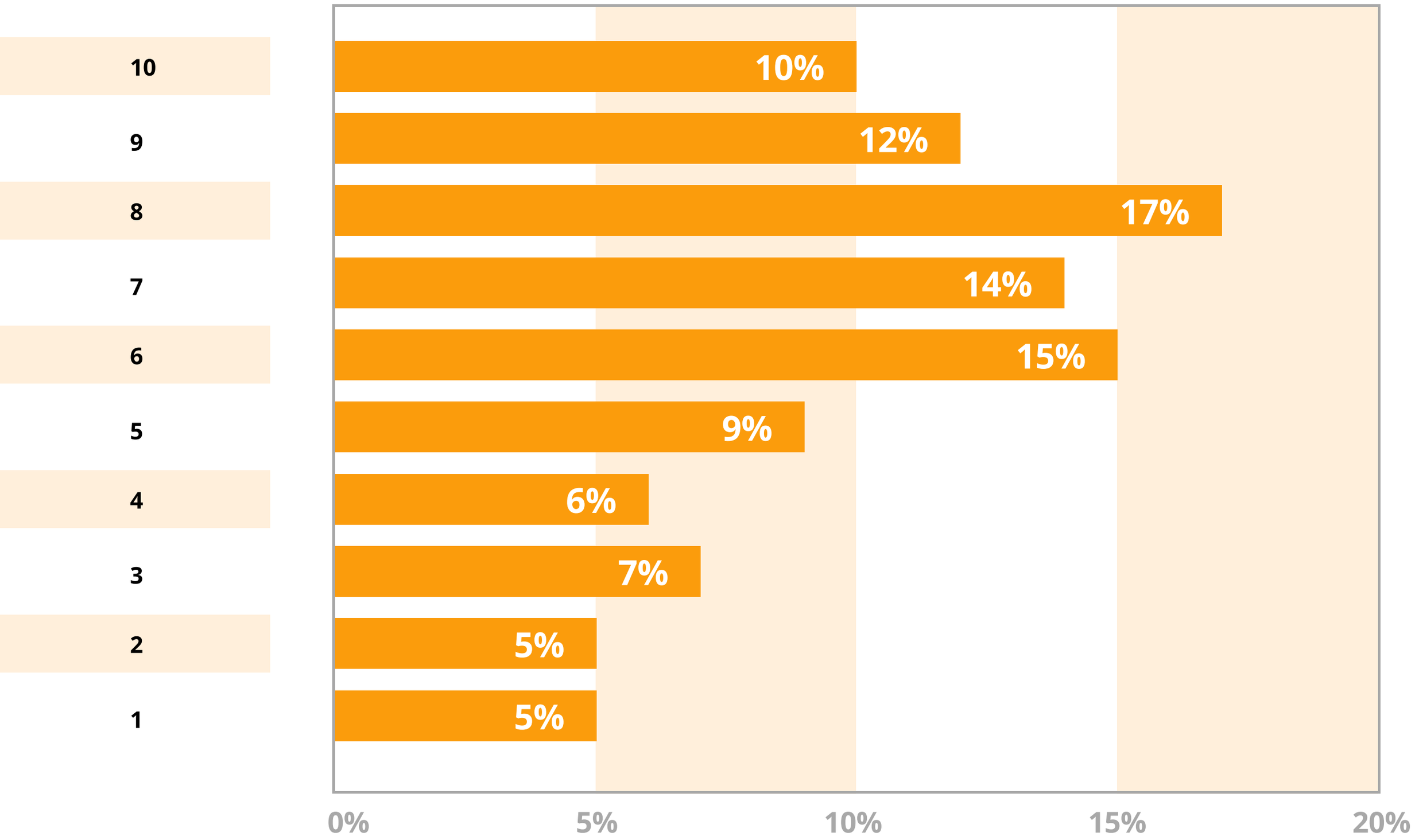


The remainder are moving more slowly, with 20% targeting POCs within six to nine months and the remaining 17% anticipating their adoption timeline extending to over a year or not at all.

Overall nearly two-thirds of respondents have work underway with GenAI. But the question is: what kind of work does this involve—and can we identify a maturity model for enterprise AI adoption as teams learn the ins and outs of real-world implementation?

Many organizations are concerned about falling behind in terms of AI adoption, expressing a fear of missing out (FOMO). 39% rate this concern high (eight or higher on a 10-point scale), while 38% rate their worry about being left behind as moderate (five, six, or seven).

On a scale of 1 to 10, how worried are you about falling behind the AI adoption curve?



Only 23% are barely worried (one, two, three, or four). This means more than three-quarters of respondents are worried about falling behind the AI adoption curve, creating a heightened sense of urgency to gain experience and expertise quickly.

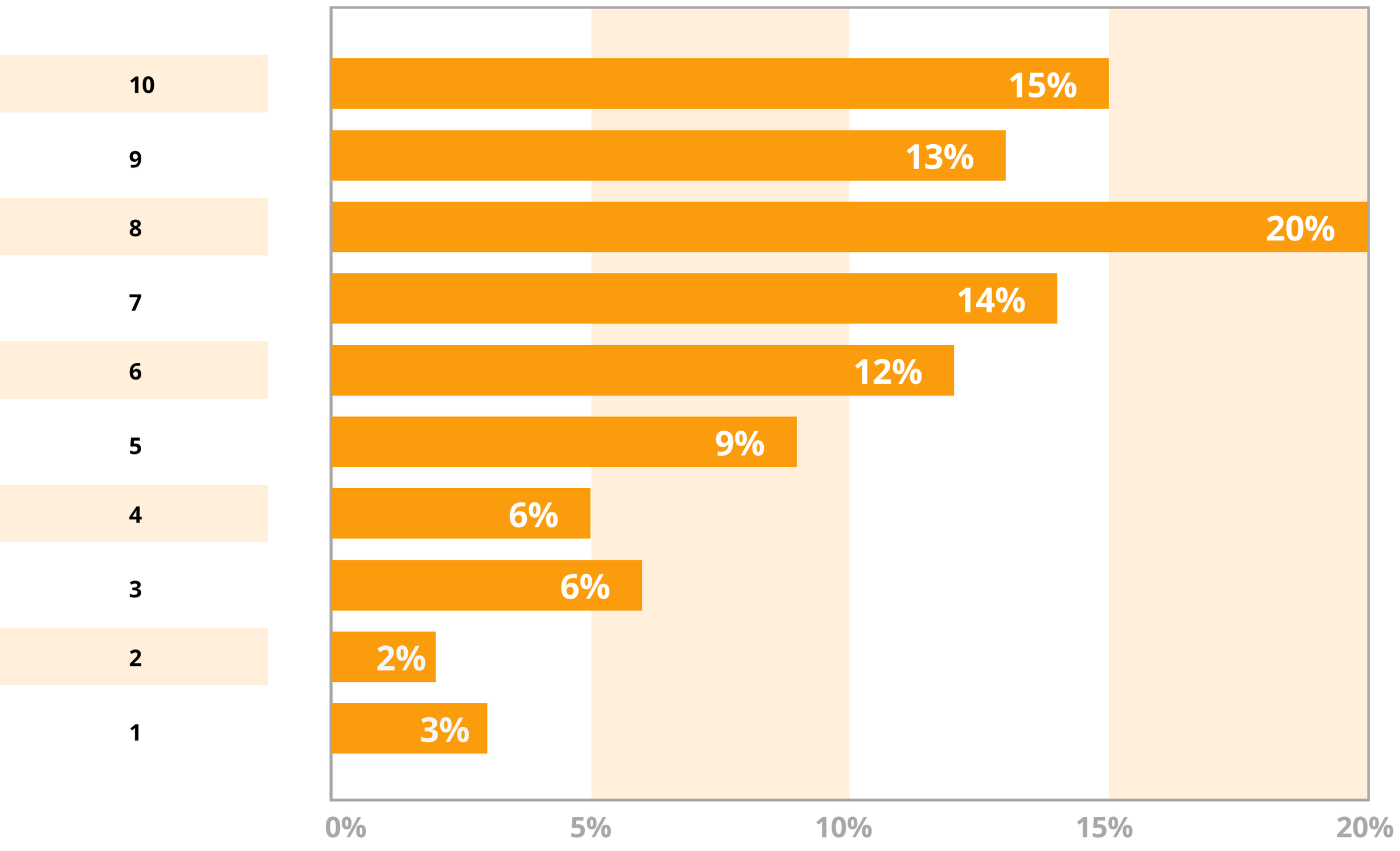
The idea of taking no action regarding GenAI is nearly unthinkable. But knowing what to do and how to do it is still a work in progress. Most organizations are wading into using GenAI for simple things like implementing coding assistants to increase development team productivity or building chatbots to better serve customers or employees.

As organizations work on these projects, they encounter challenges with their own architectural readiness, their support of RAG processes, and their overall early experience with using GenAI. Despite these challenges, however, optimism is still extremely high for what is yet to come.

Missing out on AI early adoption is one thing, but survey respondents also hold real concerns regarding corporate data and GenAI. Most responders are highly concerned about the risks of sharing proprietary data with LLMs, which could trigger countless disclosure issues.

48% say they're extremely worried (eight or higher on a 10-point scale), 35% are moderately worried (five, six, or seven), and 17% have modest concerns (one, two, three, or four) about sharing proprietary data with LLMs. Devising a strategy for including corporate data in GenAI activities, without compromising it, appears to be a key to GenAI's success in the enterprise.

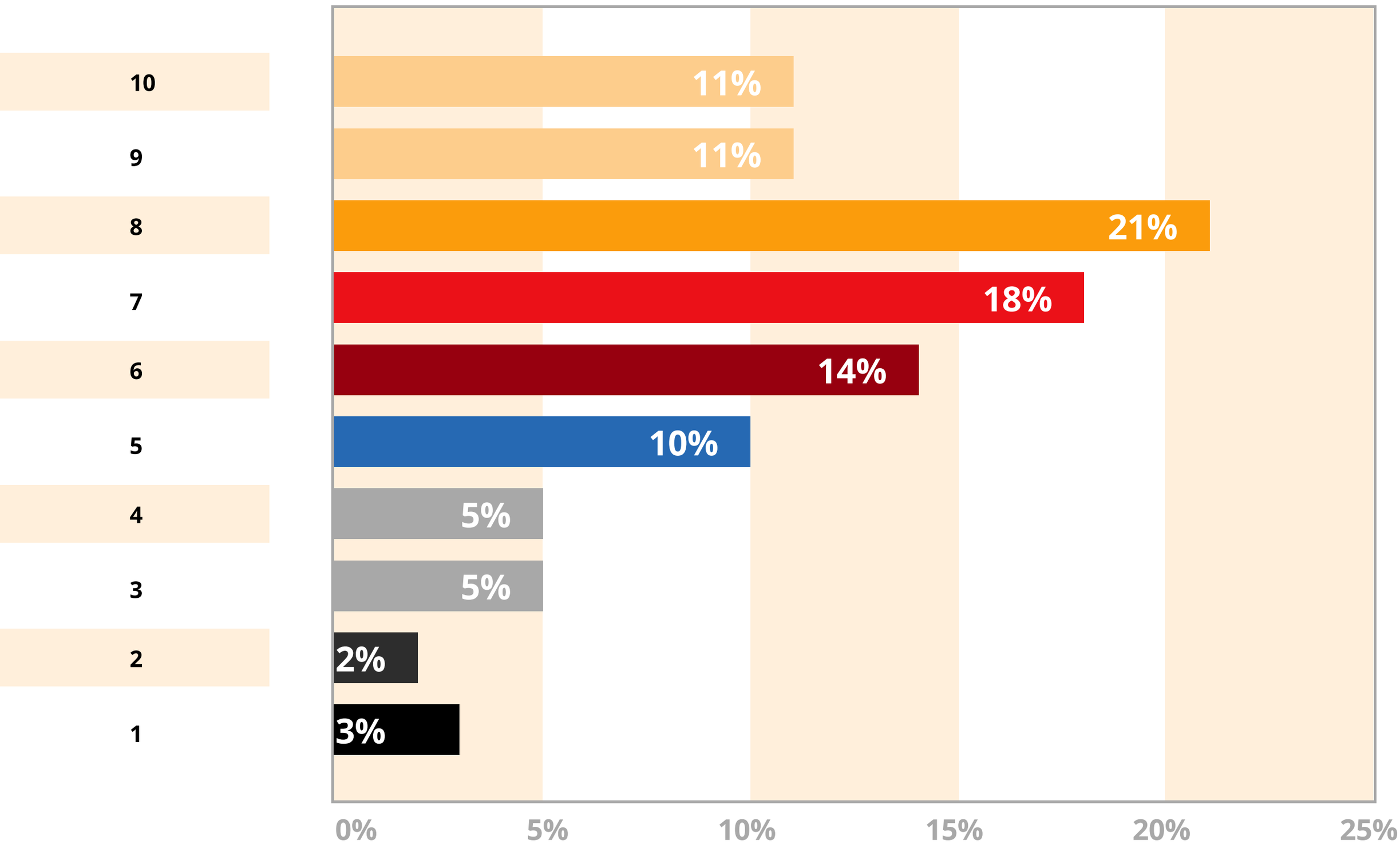
On a scale of 1 to 10, how worried are you about sharing proprietary data with LLMs?



Many respondents also express concern with how accurately LLMs answer questions. Are they truthful and trustworthy, or are they responding with hallucinations and falsehoods?

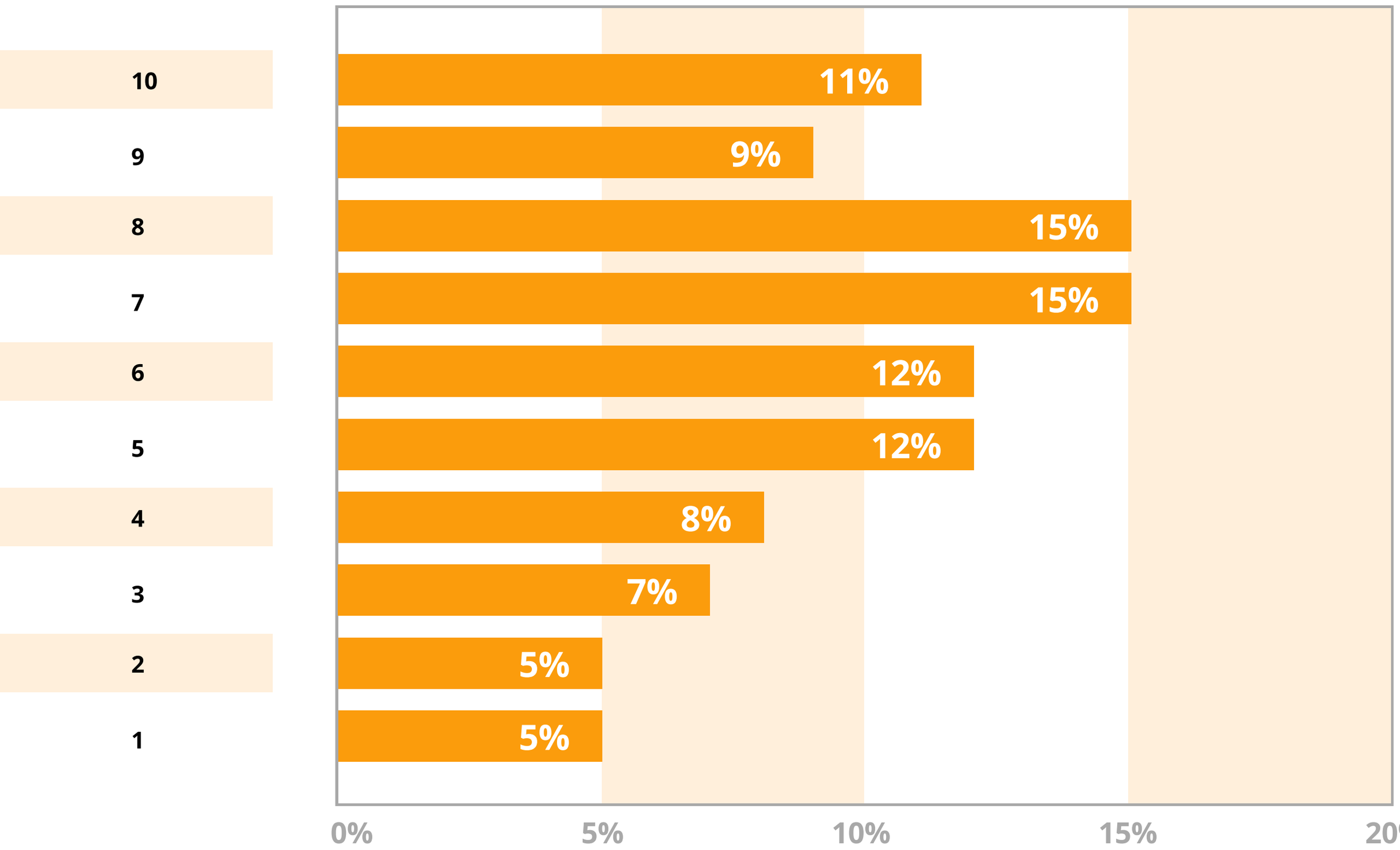
43% of respondents say they're very worried about GenAI hallucinations (eight or higher on a 10-point scale) and another 42% are moderately worried (five, six, or seven). With 85% of respondents expressing concern about LLM hallucinations at scores above four, it appears to be true that most people don't yet trust GenAI.

On a scale of 1 to 10, how worried are you about LLM hallucinations?



When it comes to building autonomous AI agents, confidence improves somewhat. 25% of respondents have little to no concern (one, two, three, or four on a 10-point scale) about developing AI agents. But the majority are still worried, with 39% holding moderate concerns (five, six, or seven), and 36% expressing a high level of concern (eight or higher).

On a scale of 1 to 10, how worried are you in building autonomous AI agents?



These responses appear to reinforce the importance of taking preliminary steps to handle data correctly and avoid hallucinations before proceeding to create autonomous AI agents.

“Hallucinations are a real problem in AI. In the beginning, it was easy to spot hallucinations because AI was not as sophisticated as it is today, but as it advanced it has become more difficult to spot them. For example, our early experiments with AI have shown that AI would invent non-existent businesses around our location, or — in the case of AI-powered merchant statement analysis with Fee Navigator — they would invent numbers that don't exist.”

Adrian Talapan, CEO and Founder at Qreli

Overriding concerns that enterprises must address before implementing GenAI break down into two related areas. Worry over sharing proprietary data with LLMs not only forfeits the entire value of that data, it also exposes the organization to significant security and legal risks. In addition, these organizations must devise a method of trusting LLM responses to be factual and not hallucinations.

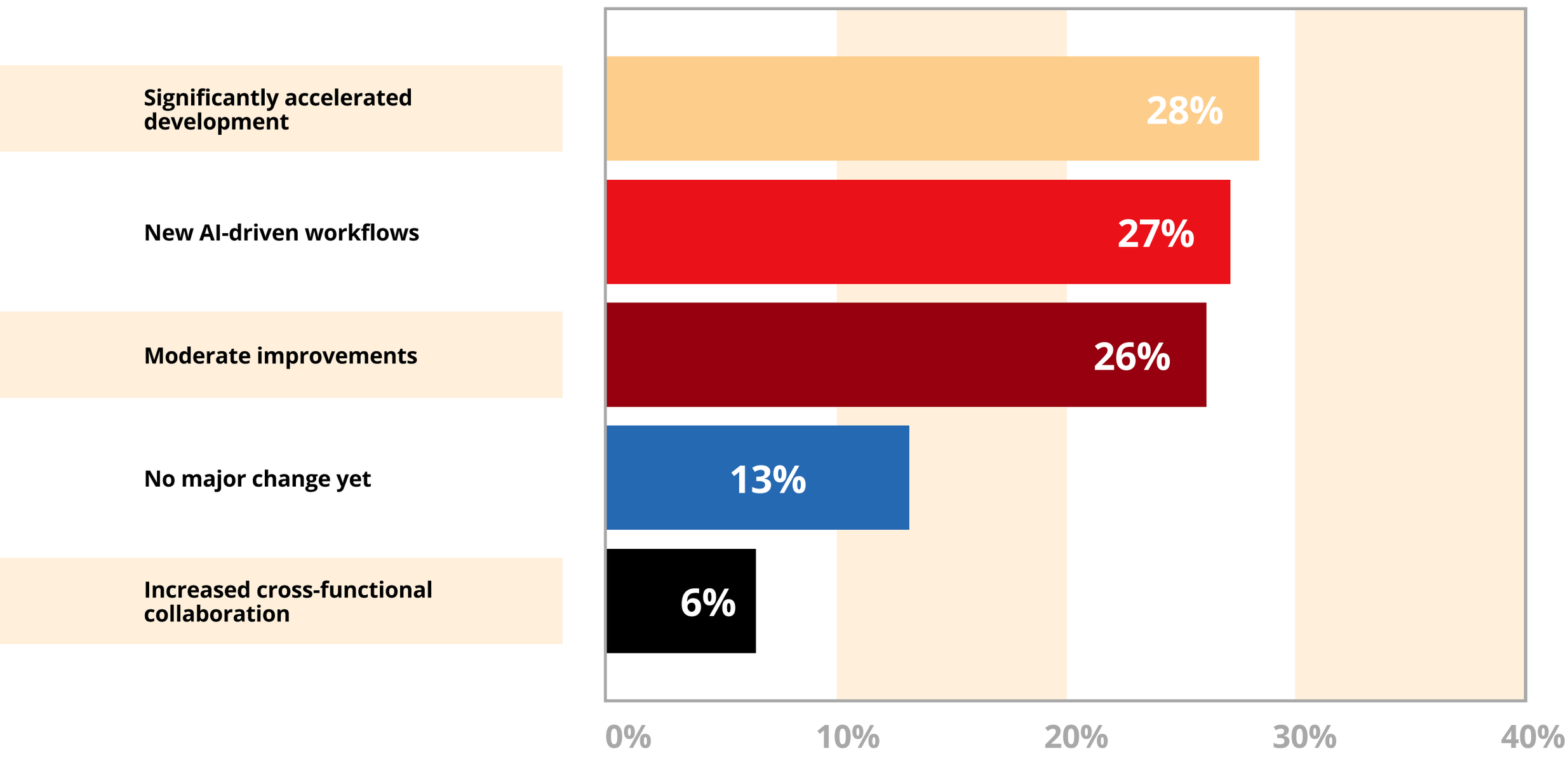
These two concerns go hand in hand. In order to fight hallucinations and falsehoods, organizations must contribute stronger and more specific contextual information when conversing with GenAI. This context originates from their internal data.

The paradox is that, in order to solve for issues around response accuracy, organizations must first resolve concerns over sharing their data. However, a safe, reliable solution for that hasn't yet been deployed at scale.

Chapter 2:
Early AI Project Successes

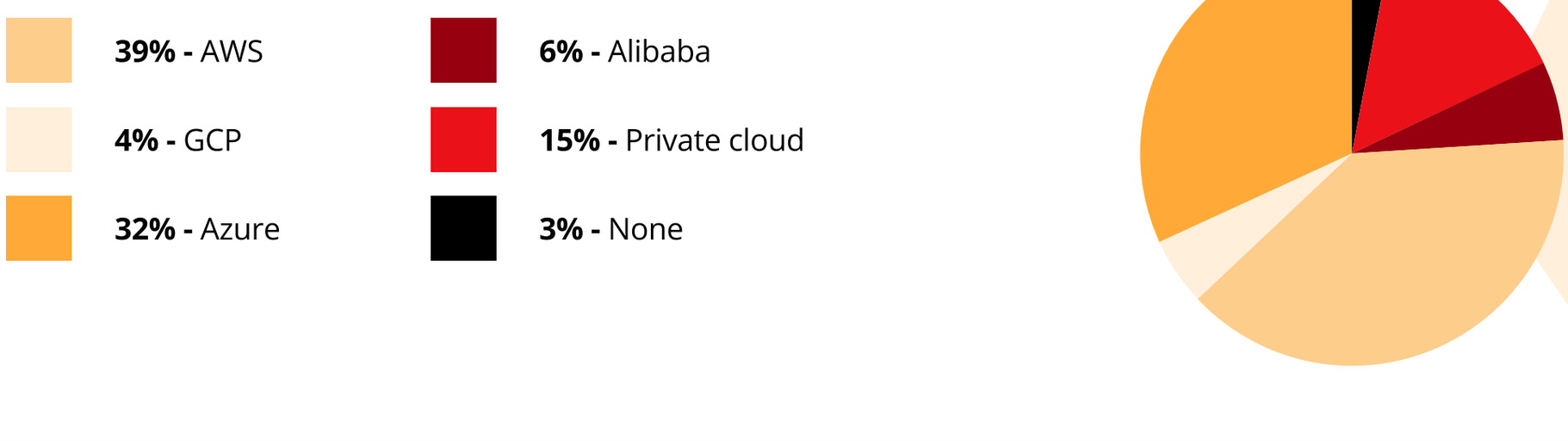
AI has had a positive effect on development teams, including significant (28%) and moderate (26%) app development acceleration. 27% of respondents have added new AI-driven workflows, and 6% have increased cross-functional collaboration, all of which reinforce the internal productivity impact of using copilots. Only 13% report having seen no major changes yet.

How has AI transformed your app development process in the past year?



For their operating environments, most respondents use AWS (39%) or Azure (32%) as their cloud provider. Use of GCP (4%) and Alibaba (6%) is much less common. 18% report using a private cloud provider or none at all.

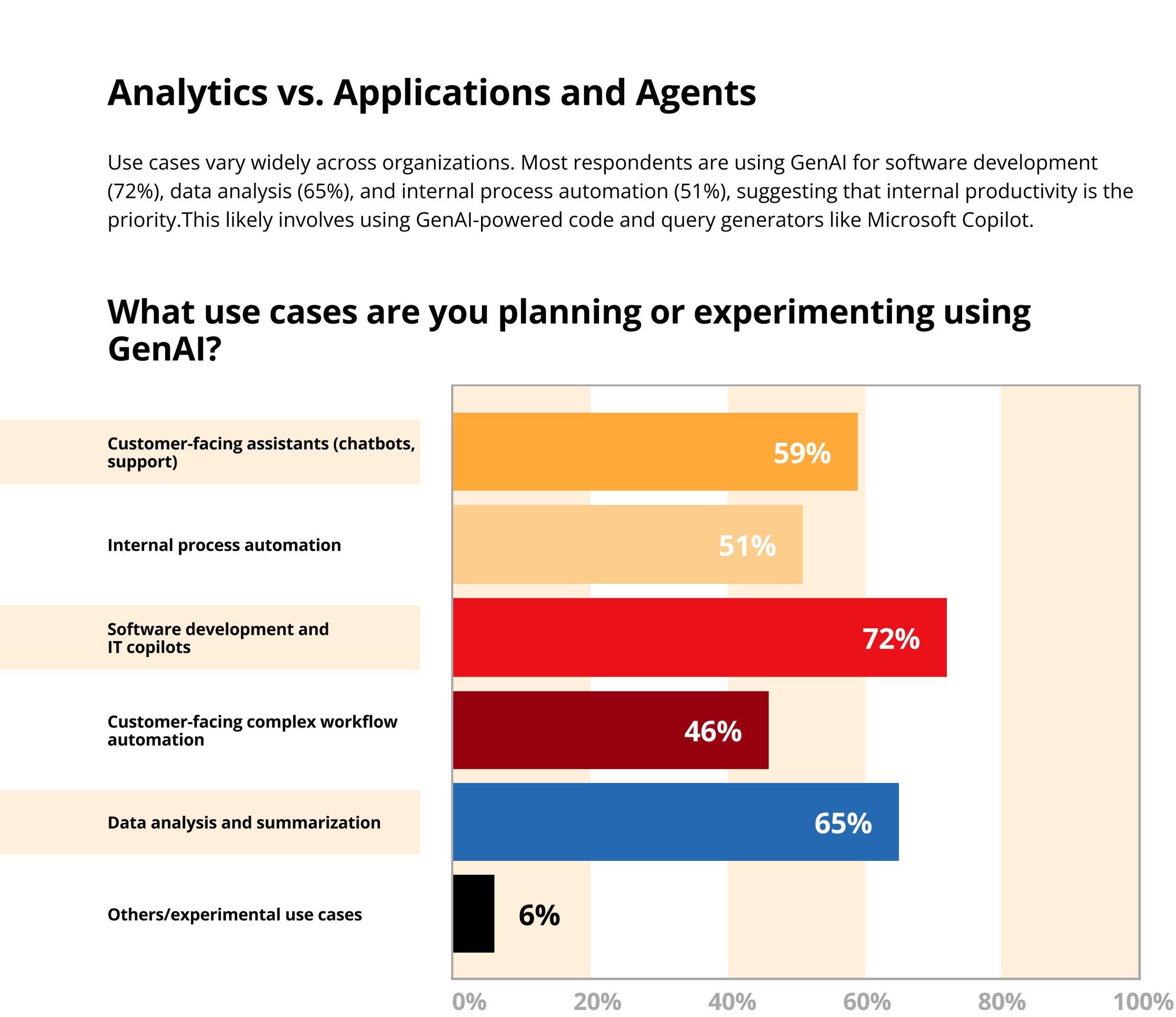
Which cloud provider (if any) are you using?



Analytics vs. Applications and Agents

Use cases vary widely across organizations. Most respondents are using GenAI for software development (72%), data analysis (65%), and internal process automation (51%), suggesting that internal productivity is the priority.This likely involves using GenAI-powered code and query generators like Microsoft Copilot.

What use cases are you planning or experimenting using GenAI?



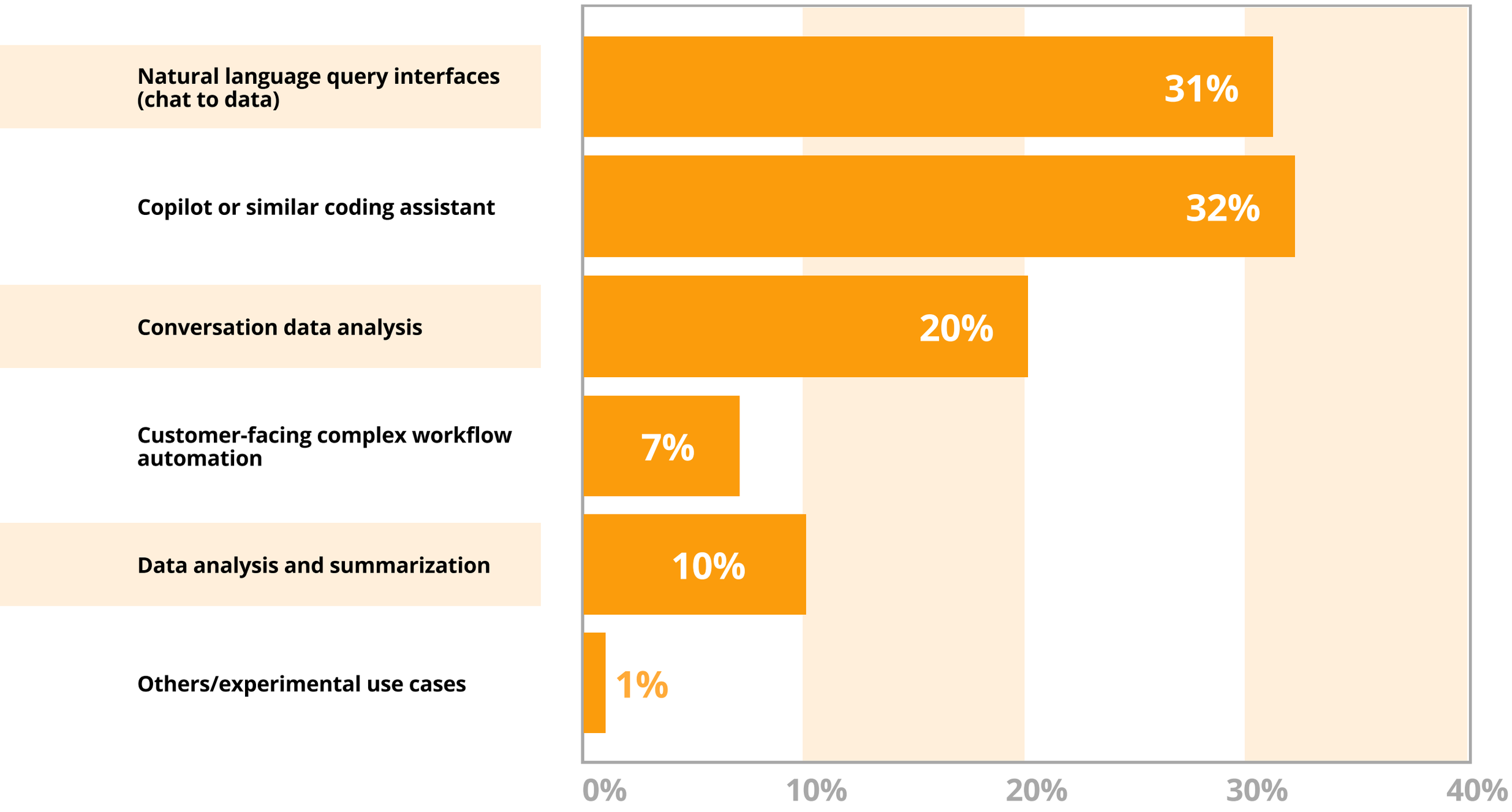
However, many of their AI applications are external and customer-facing, such as chatbot assistants for support (59%) and automating complex customer-facing workflows (46%). In addition, 6% are working on experimental use cases.

Given the overwhelming implementation of copilot assistants, their use has essentially become mainstream. Chatbot development is likely to be the first real-world GenAI application. Internal and external automations powered by GenAI appear to be next in line as organizations move forward with their agentic application plans.

Analytics Focus

When implementing conversational analytics in AI applications, organizations report a relatively even split between using built-in natural language query interfaces (31%) and copilot tools (32%). In comparison, more sophisticated conversation data analysis (20%) and conversational BI tools (7%) are less common.

How are you implementing conversational analytics in your AI applications?

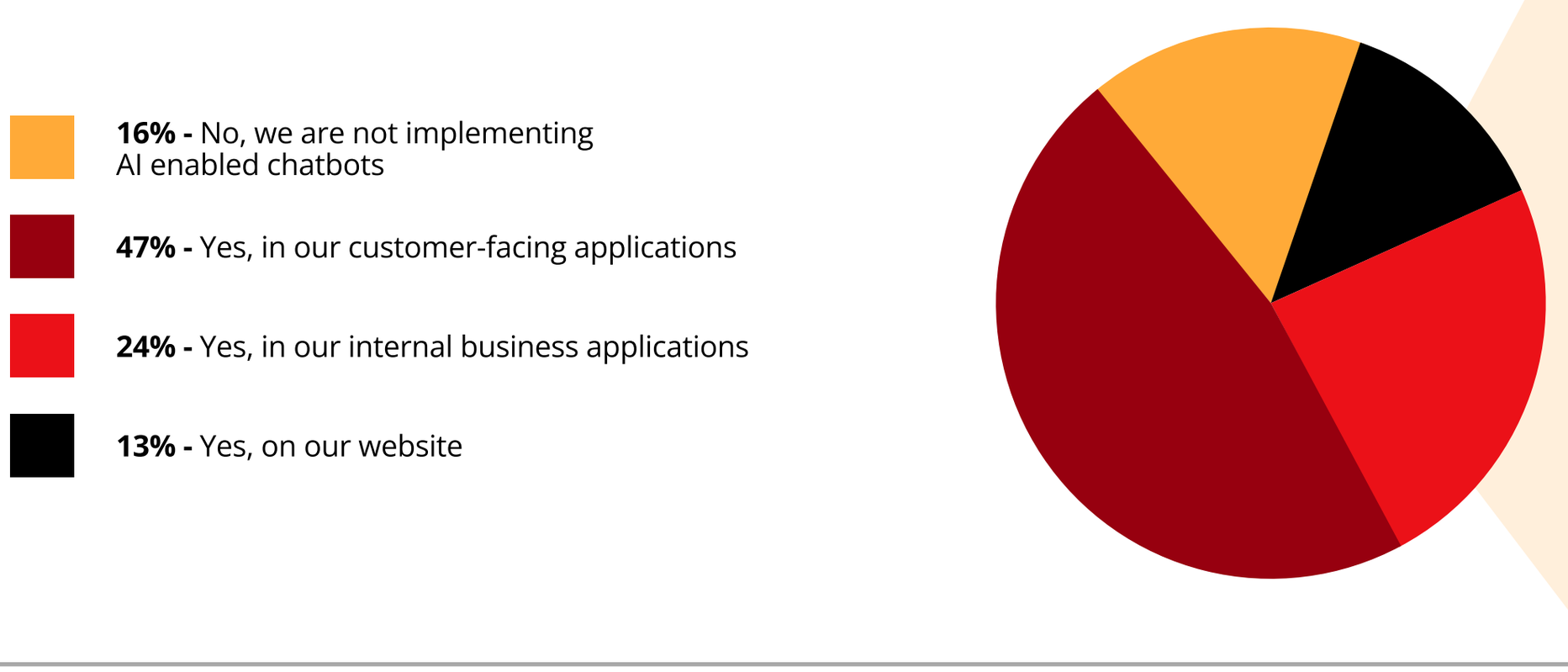


These responses indicate that individuals are benefiting from using AI-assisted tools. However, AI-powered analytics inside applications is still relatively rare.

Chatbots

AI-enabled chatbot applications are popular for both internal and external usage. The majority of respondents report implementing either customer-facing applications (47%), web chatbots (13%) or internal business application chatbots (24%). Only 16% of respondents have yet to deliver AI-enabled chatbots.

Are you implementing AI enabled chatbots?

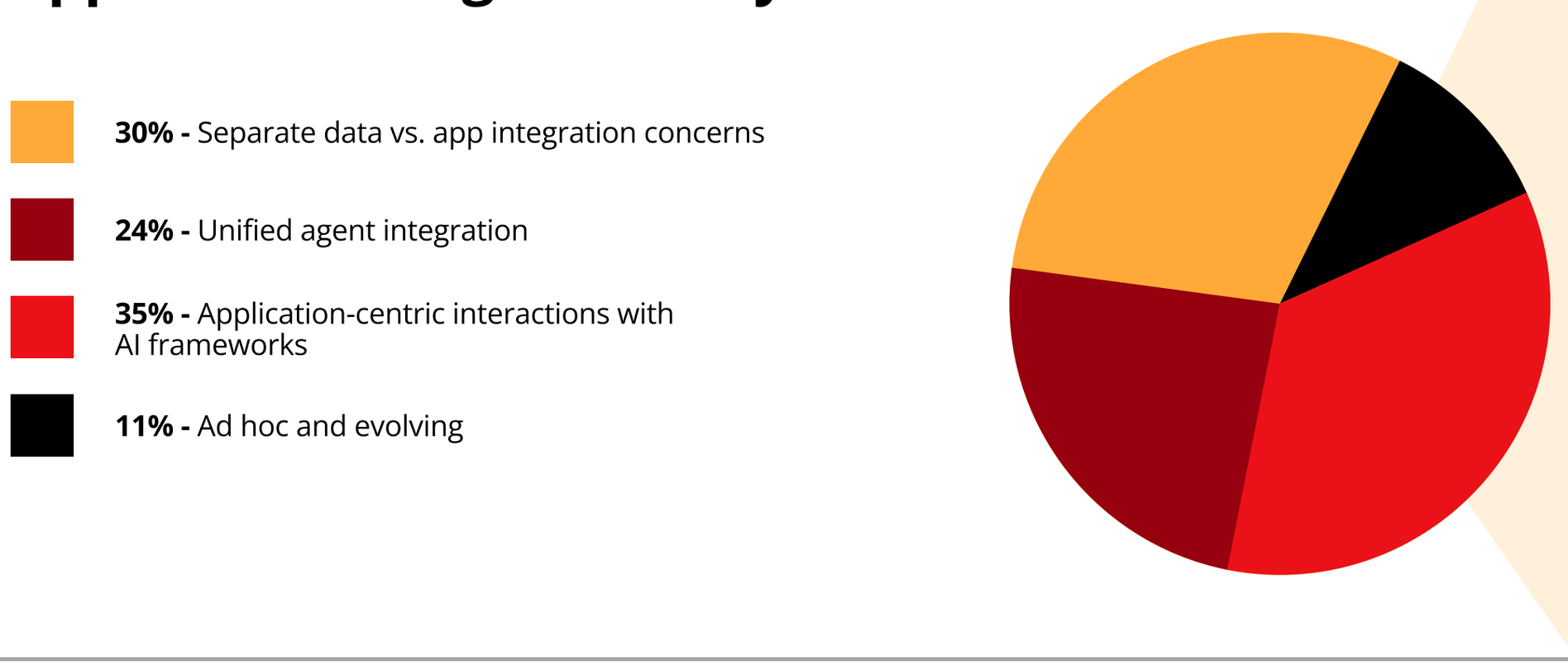


Applications and Agents

Most organizations are taking an application development-centric approach when determining how data and application integrations work with AI frameworks (models). 35% report using application-centric interactions with AI frameworks, and 30% separate their data and application concerns from one another.

However, 11% take an ad-hoc and evolving approach, which indicates they're working in an application-first manner and addressing data management as they progress.

How do you handle data integration versus application integration in your AI architecture?

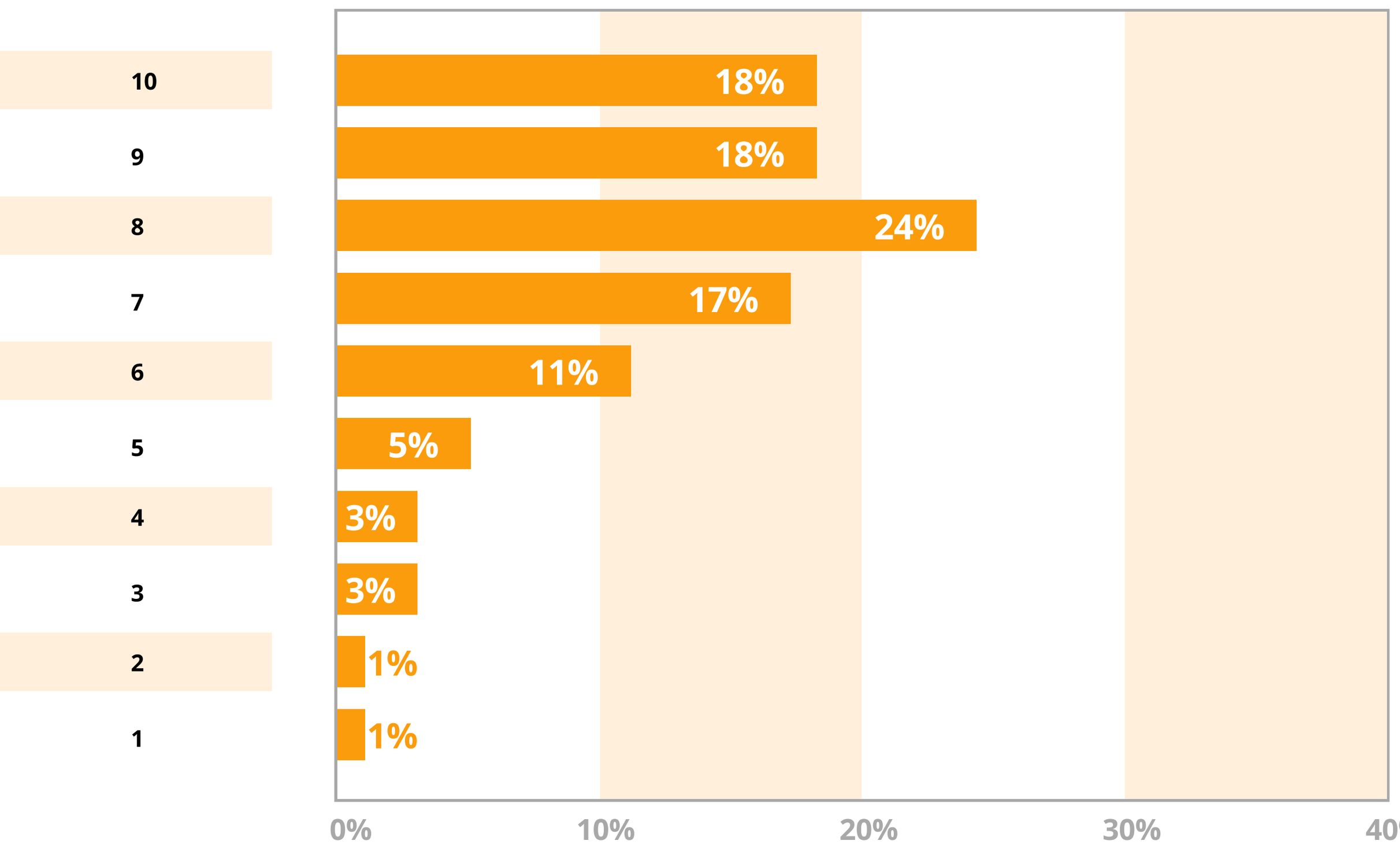


Only 24% have unified agent integrations between data and application functionality, recognizing the intertwined nature of data, AI models, and application functionality.

This low level of unified agent integration is one reason why confidence levels in addressing the most risky issues with data and AI (security and hallucinations) may be too high.

Despite concerns with AI autonomy, however, most respondents are confident in their ability to create and deploy autonomous agents. 60% report high confidence (eight or higher on a 10-point scale) that they will be able to deploy autonomous, decision-making, and action-taking agents and agentic applications.

In a scale from 1 to 10, how confident are you that you will be able to create and deploy autonomous, decision-making and action-taking agents and agentic applications?



"While technical challenges certainly impact AI implementation timelines, we find that AI looks a bit like a hammer nowadays, seeking nails everywhere. We believe a more robust approach is to use AI where it shines, and to control experiences and risk in a deterministic, non-AI fashion."

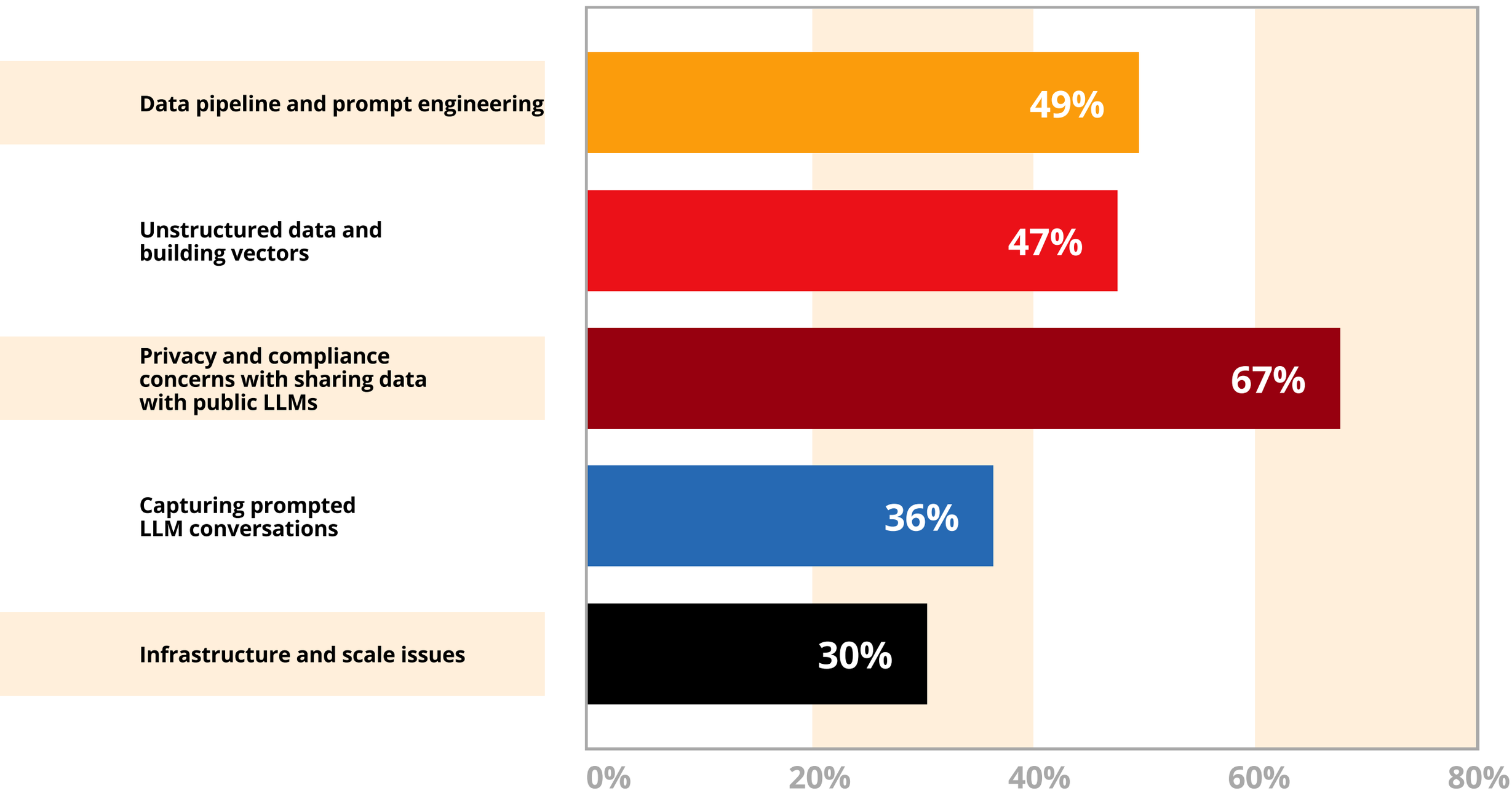
Adrian Talapan, CEO and Founder at Qreli

Chapter 3: RAG Implementation: Confidence vs Capability

Respondents report a range of challenges with incorporating data in AI implementations. Privacy and compliance concerns related to sharing data with public LLMs stand out as the biggest challenge (67%), validating the risks of oversharing and losing control of important enterprise data.

Many respondents also indicate concerns with how to handle that data safely. 49% report issues with data pipeline and prompt engineering, indicating that gathering and feeding data to LLMs is a challenge. Among those using RAG to protect enterprise data and still use it, 47% struggle with unstructured data and building vectors.

What challenges have you encountered with data incorporation in your AI implementations?

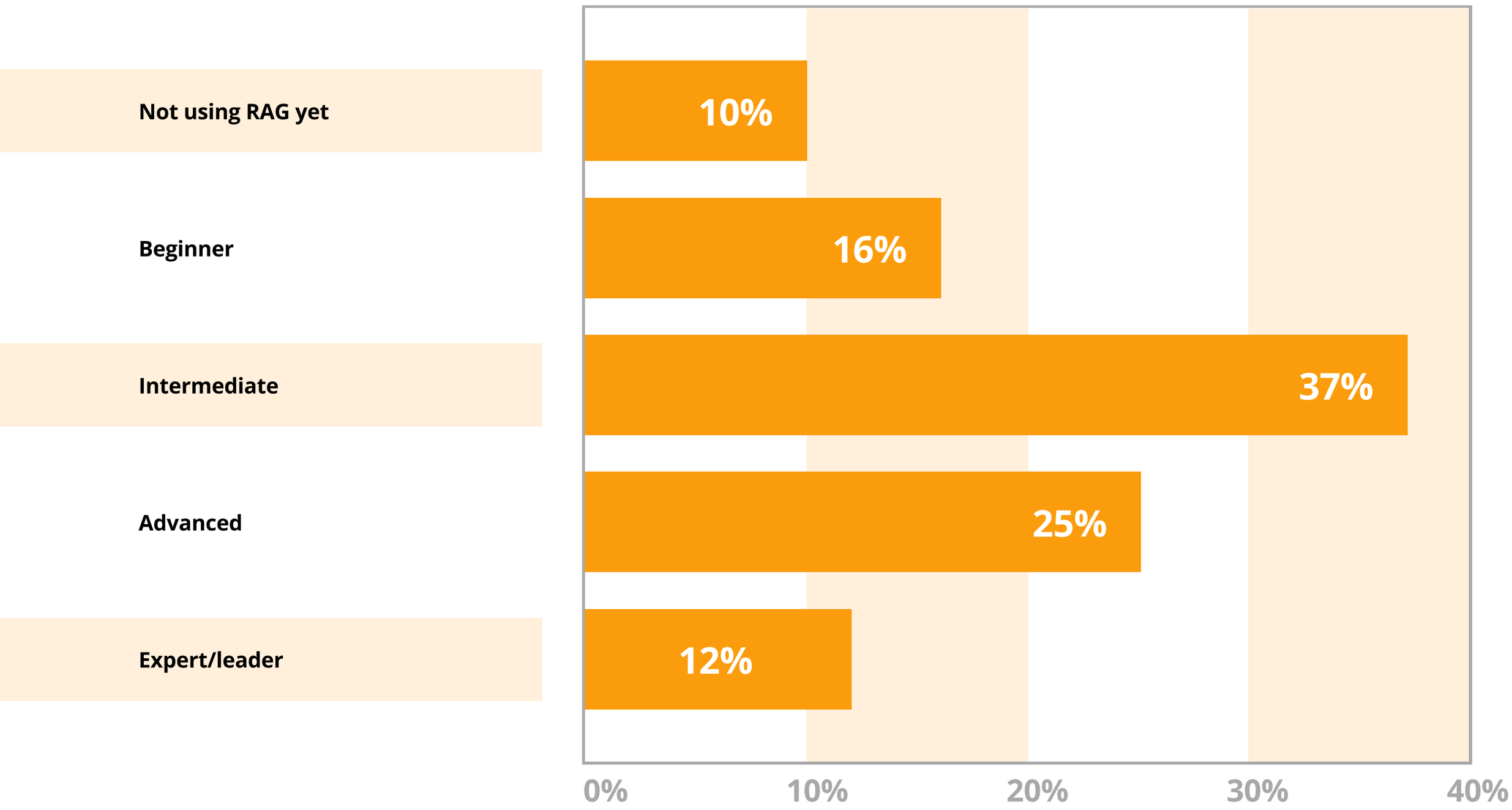


All of these concerns align with the deliberate pace of GenAI implementation that may be restraining AI adoption and contributing to the FOMO mentioned in Chapter 1.

Overall, development teams are relatively confident in their understanding and adoption of RAG techniques and workflows. However, in some cases, they may be overconfident.

Nearly three-quarters (74%) of respondents consider their development teams intermediate, advanced, or expert regarding where and why to adopt RAG techniques and workflows. Just 16% consider their teams beginner, while 10% aren't yet using RAG.

How advanced are your development teams regarding where and why they need to adopt RAG techniques and workflows into their everyday activities?



If 37% of teams are advanced or expert, why have enterprises deployed so few agentic applications today? Where do organizations struggle in supporting RAG?

RAG is a multi-step process designed to improve LLM response accuracy to battle back hallucinatory answers. It helps protect proprietary corporate information from becoming training data for LLMs while still using that data to inform LLMs about where to look within its knowledge base for accurate information.

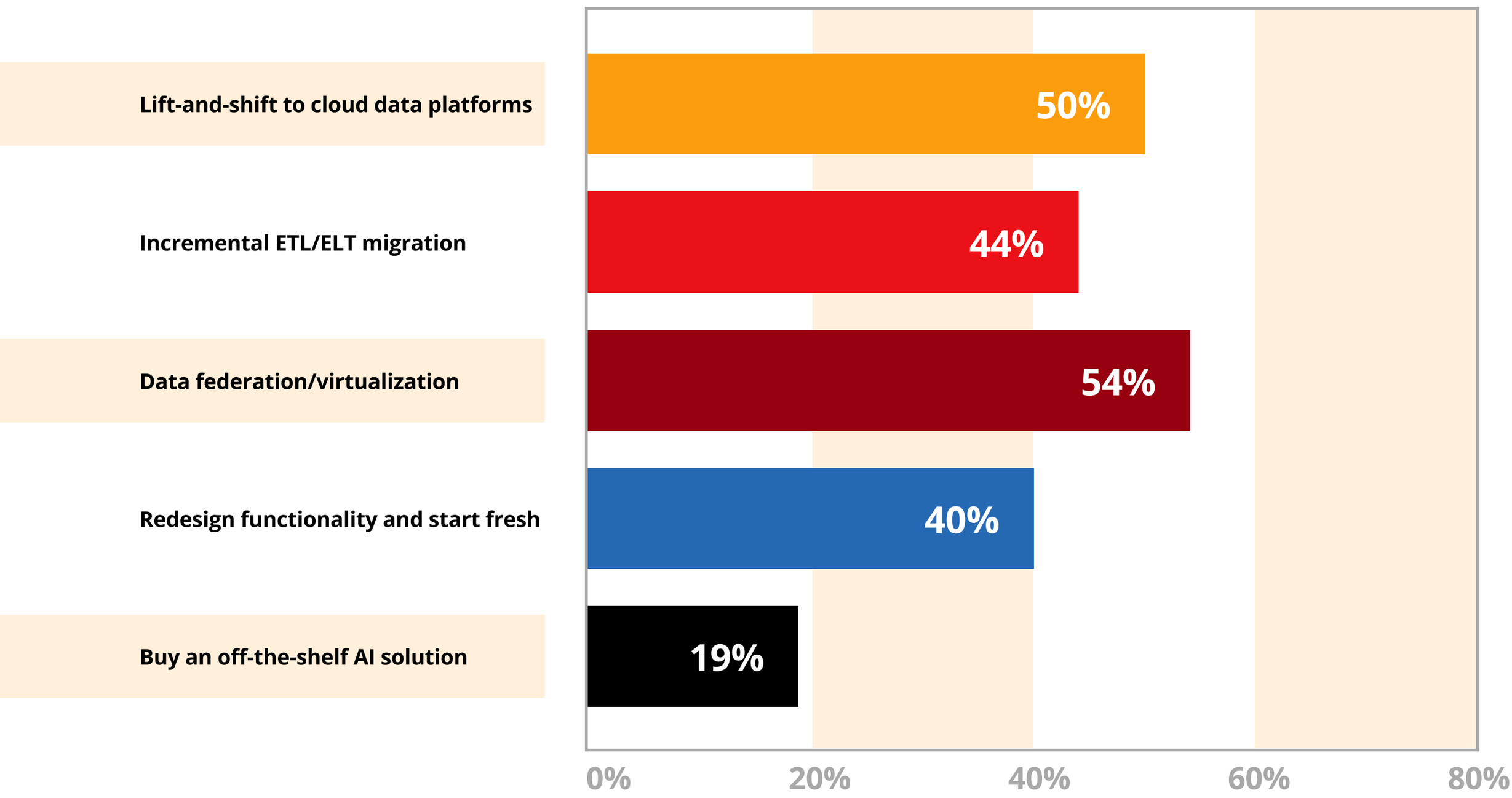
The sequence stages of the RAG workflow include data preparation, vectorization, prompt engineering, LLM model management and conversations, response validation, and post-response actions. RAG uses language models in multiple stages of the sequence, first to build vector indexes from corporate data, then to converse with the main LLM, and possibly to assist in response validation or post-response actions.

Data Preparation

When migrating and utilizing data during AI adoption, organizations actively modify their data architectures in multiple ways. A majority of respondents take a federated and virtualized approach (54%) or a lift-and-shift to the cloud strategy (50%).

An additional 44% use incremental extract, transform, load (ETL) and extract, load, transform (ELT) techniques to move and prepare their data. Only 19% have purchased off-the-shelf AI solutions, while twice as many (40%) redesign and start fresh with new projects.

What strategies do you employ for data migration and utilization when adopting AI capabilities?



All of these activities signal dramatic changes related to preparing enterprise data for use with AI. This appears to indicate that AI is not data-ready. GenAI expects text-based, natural language interactions, but most enterprise data is stored in structured formats that aren't optimized for these interactions.

More than 70% of respondents don't use a unified multi-model database to feed AI, which means only **29% have a unified database feeding AI.** Rather, these organizations appear to take a piecemeal approach to delivering data. 35% have one primary data access model, 18% have a polyglot approach of using many purpose-built databases at once, and 18% are undecided.

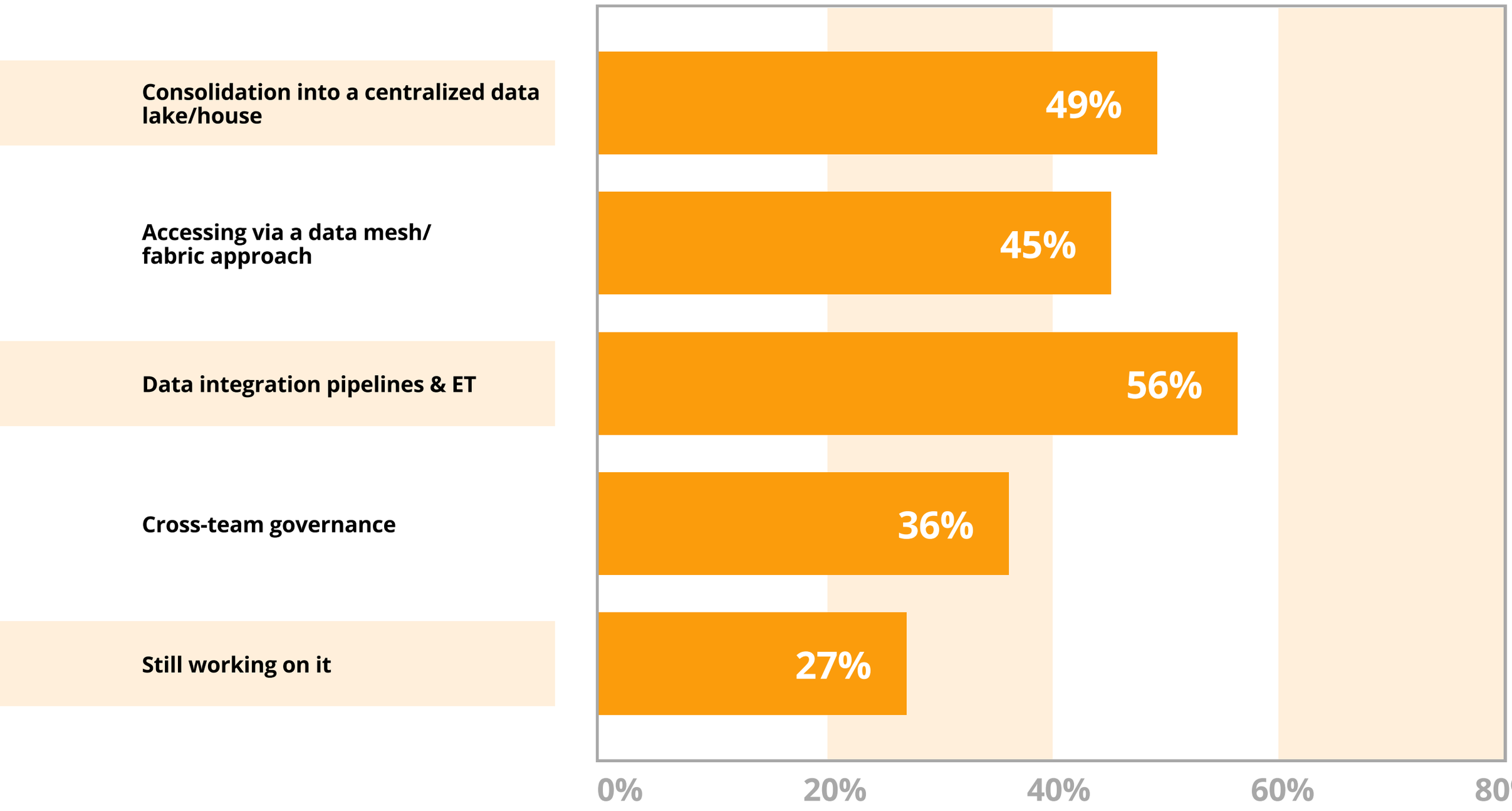
How are you utilizing multi-model database capabilities in your AI implementation?

- **29% -** Unified multi-model DB feeding AI
- **35% -** Limited use (one data access model as primary)
- **18% -** Polyglot approach (many purpose-built databases at once)
- **18% -** Exploring/undecided

This illustrates a complicated approach to gathering and using enterprise data with GenAI, which could become even more entangled downstream. These entangled complications manifest themselves later in the inability to debug and back trace what data, provided within a prompt, induced a troublesome hallucination response from an LLM.

Enterprise data silos create complications for organizations as they develop AI-enabled applications. Respondents address data silos with data integration pipelines and ETL (56%), consolidation into central data lakes and warehouses (49%), access via data fabrics or meshes (45%), and cross-team governance (36%).

How are you addressing data silos when developing AI-enabled applications?



27% are still exploring how to incorporate data silos for use with AI. This spotlights the complexity, difficulties, and diverse techniques involved in wrangling existing data sources from across the enterprise to work with AI. Addressing data utilization complexities is a prerequisite to becoming proficient with RAG and building AI-enabled applications. It appears that complexity levels may be shifting from data architectures to GenAI and agentic design.

Using a text-based data management format like JSON may be appropriate for capturing and using data with GenAI. Implementing a unified data management platform with multipurpose data access capabilities for eliminating gaps across point solutions would make preparing data for use within RAG workflows significantly easier, faster, and more accurate. A JSON-based, developer-friendly, multipurpose database platform could dramatically reduce this complexity.

Vectorization

RAG uses vector indexes, created as abstracted coordinates from corporate data, to provide stronger context for LLMs alongside natural language prompts. Prompts include instructions for the desired information, while vectors provide information about where the LLM should look to find it.

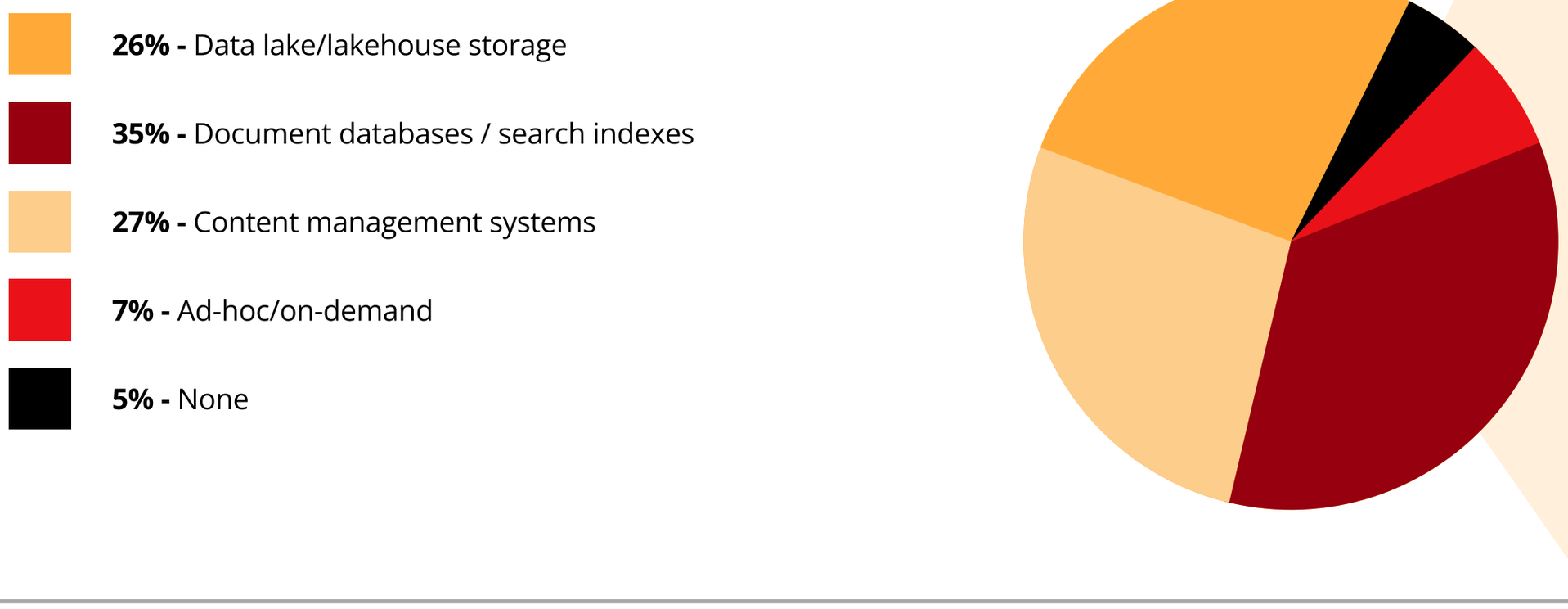
Vectors are built by feeding contextual enterprise data to an independent embedding model (a second language model). The embedding model is allowed to read but not keep this enterprise data, eliminating the risks of sharing this data publicly.

The embedding model uses nearest-neighbor algorithms to derive (or vectorize) the numerical coordinates of similar information that an LLM might understand without compromising the original source data. Databases store these coordinates and facilitate queries and lookups for the vector values prior to each LLM conversation.

Corporate data comes in many formats via many channels, including structured data from applications and databases and unstructured data in the form of documents, images or other media. All formats can be vectorized, stored, and used while conversing with LLMs.

Respondents use a wide range of approaches to unstructured data management in AI applications. 35% use document databases and search indexes, 27% use content management systems, and 26% use data lakes. The remaining 12% take an ad-hoc approach or no approach at all.

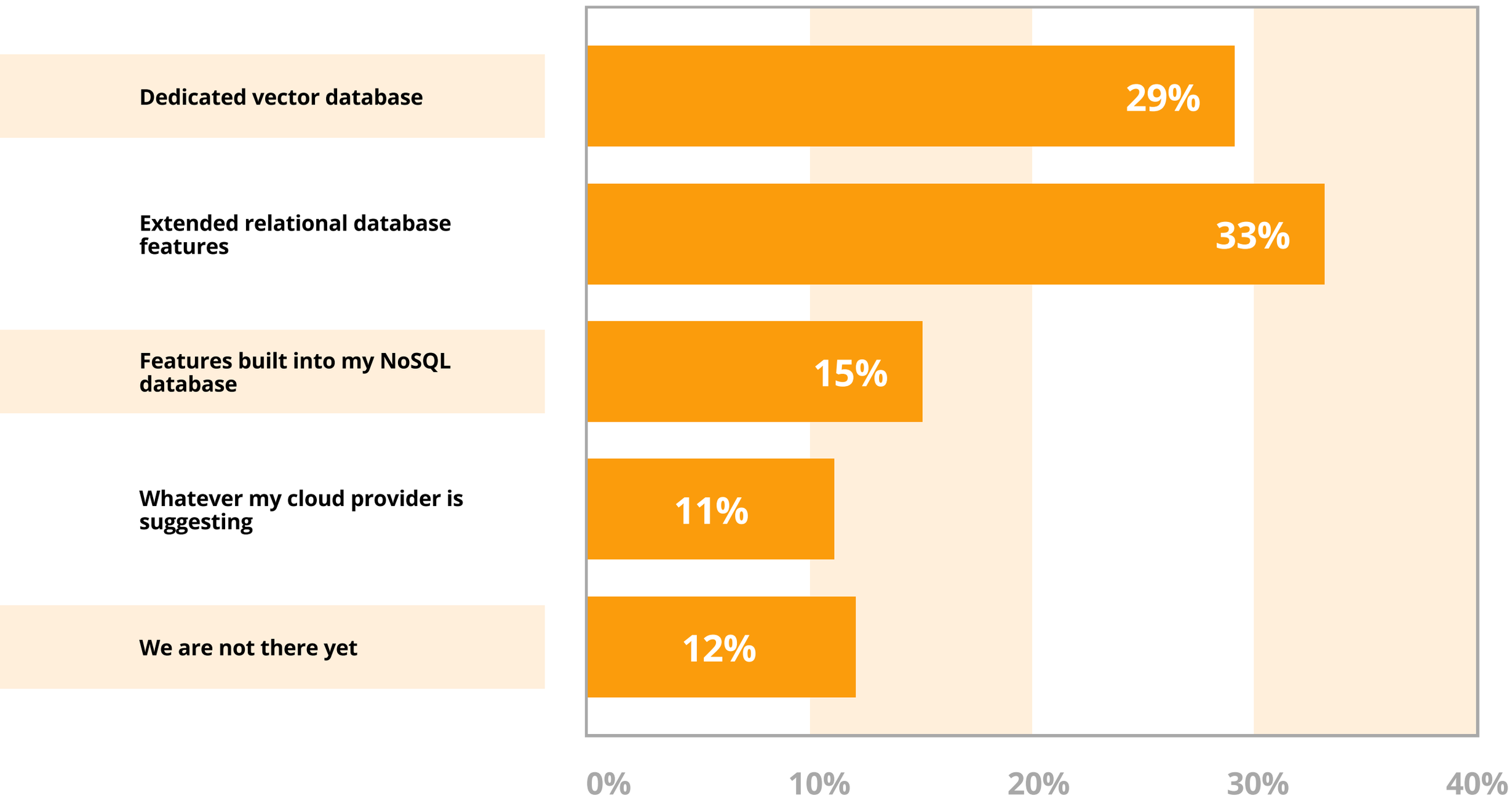
What is your primary approach to unstructured data management in AI applications?



Unstructured data such as PDFs, graphics, presentations, videos, and audio as well as semi-structured formats like JSON, XML, or even HTML contain massive amounts of vital information to inform AI. Yet as the data shows, enterprises struggle to manage unstructured data for AI applications.

To integrate embeddings (aka vectors) in database architecture, respondents are split between extended relational database features (33%) and dedicated vector databases (29%). An additional 15% use built-in NoSQL features.

How do you currently manage the integration of embeddings (aka vectors) in your database architecture?

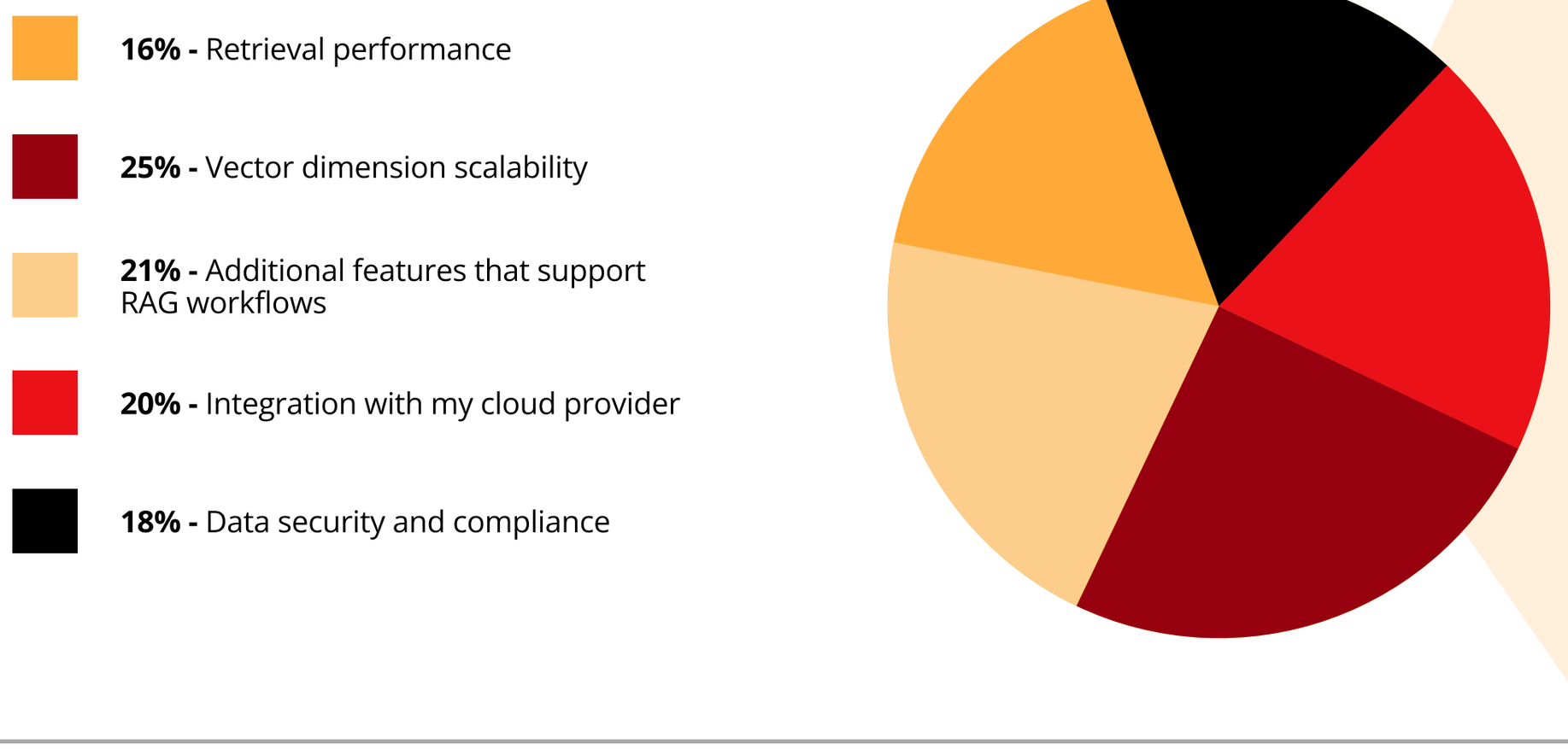


No current best practice appears to exist yet, as 23% don't yet have a dedicated solution or depend on suggestions from their cloud provider.

This variety and uncertainty may cause some organizations to invest in suboptimal solutions they'll need to revisit when their RAG workflows become more advanced.

When deciding on a vector database, 25% of respondents prioritize vector dimension scalability, 21% consider additional features that support RAG workflows, and 20% prioritize integrations with their existing cloud provider. An additional 18% favor data security and compliance, while retrieval performance is a priority for only 16%.

What factors influenced your choice between vector databases?



This focus on scalability suggests an appetite for future-proofing vector infrastructure over immediate feature needs for RAG support, cloud interoperability, real-time responsiveness, or security. It also indicates that organizations may prioritize the biggest container for vectors over a database designed for specific use cases.

As they explore the operational and development benefits when mastering RAG, it may be worthwhile for organizations to investigate the relationship between storing vectors alongside their source data within a scalable unified data platform.

Prompt Engineering

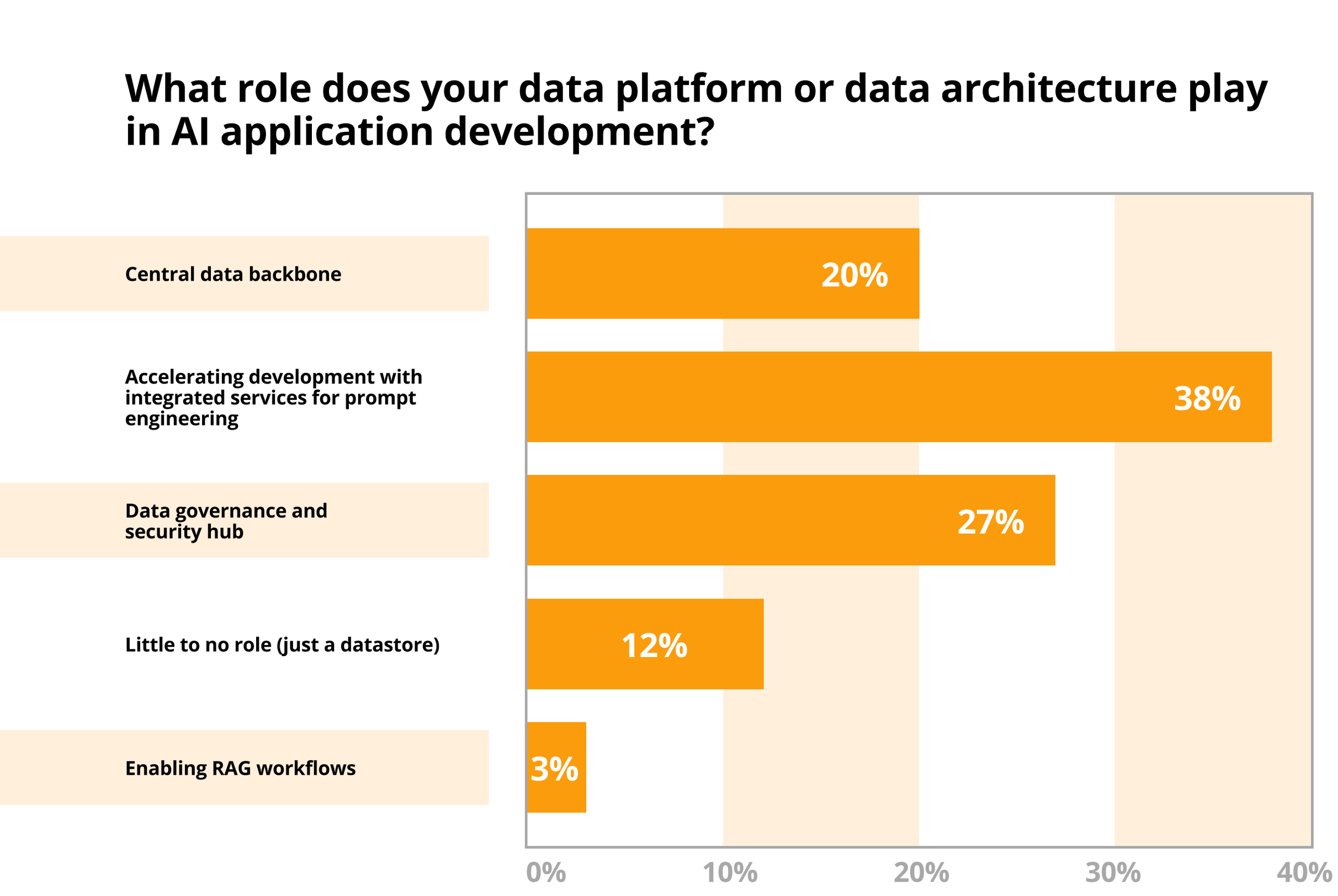
Prompt engineering is the exercise of programmatically building requests and instructions for LLMs like ChatGPT within software programs like chatbots or agents. It involves question-and-answer sessions with an LLM, similar to building a MadLib where different variables are inserted at runtime within the software program.

These MadLib-style prompts also include the reference points provided by vector coordinates. The cooperating LLM reads and interprets the prompts, uses vectors to pinpoint its own knowledge, and assembles and generates a text-based natural language response. During a session, these conversations can go back and forth, based on how the software program is designed to operate.

This back-and-forth transcript is a valuable artifact. Organizations can evaluate them for accuracy and appropriateness during a post-conversation validation step in the RAG workflow that may also include other analysis or additional metadata, all of which should use a JSON format.

Many organizations have specific, dedicated designs for data architecture in AI application development. 38% use their data platform to accelerate development with integrated services for prompt engineering, indicating that they may be the copilot users. An additional 27% use their data platform for data governance and security. Only 20% use architecture as a central data backbone.

What role does your data platform or data architecture play in AI application development?



A mere 3% use their data architecture to enable RAG workflows, which is four times less than those who don't use their data architecture with AI (12%). This suggests that organizations don't know what kind of role a unified data architecture can play in streamlining RAG workflows, or they're unaware of what RAG is.

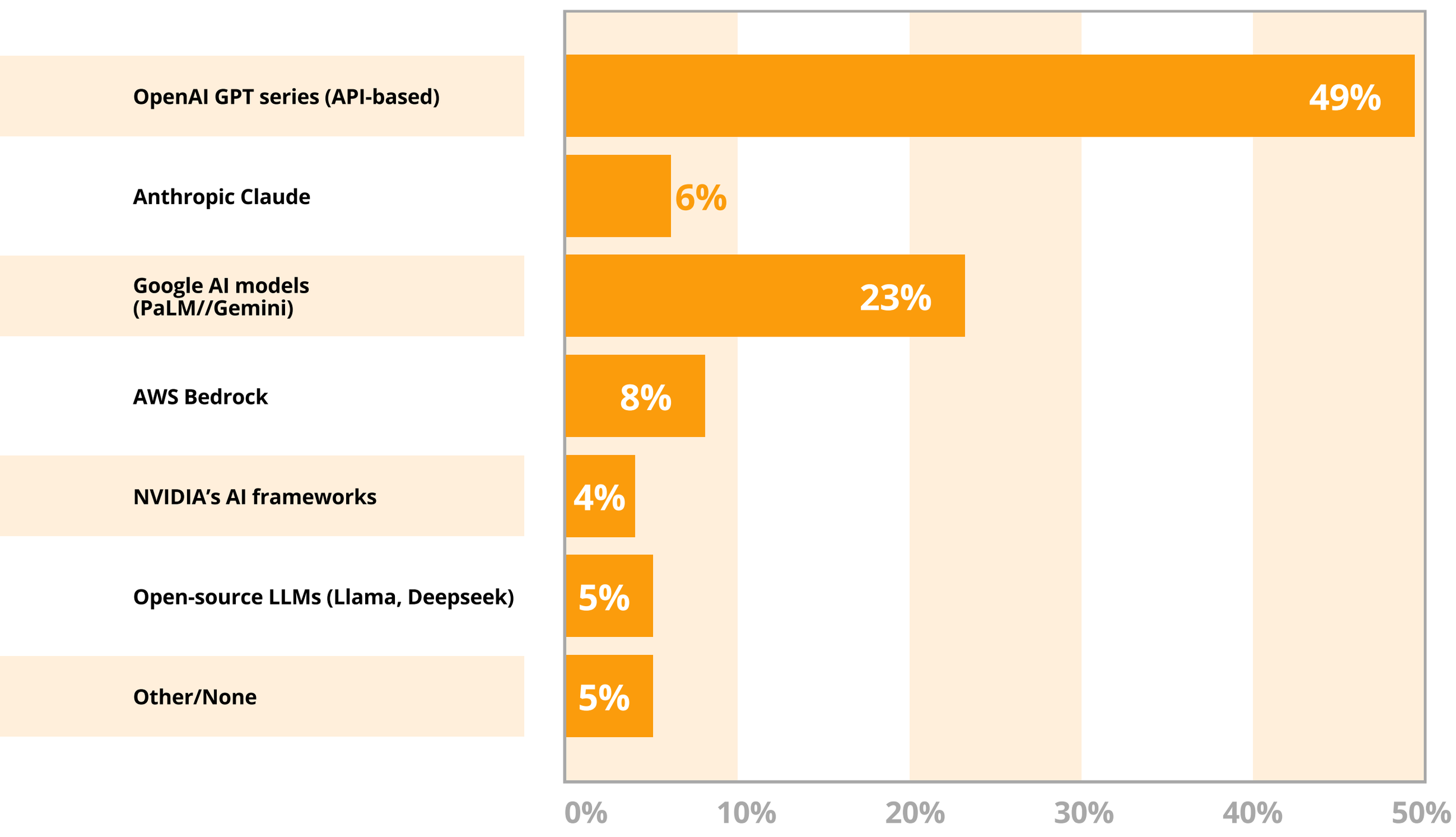
Managing Models and Applications

GenAI development centers around choosing and using LLMs and frameworks to access them. Therefore, deciding where to host models, which models to consider, and which roles to use them for (e.g., embedding, primary inquiry, or response validation) is critical.

An all-knowing model like ChatGPT versus a specialty model trained in medicine would dramatically change the result from a primary inquiry. The models' latency, responsiveness, and lack of session-to-session memory would be another implementation concern for scaling AI interactions. In addition, the way organizations manage the evolution of knowledge within models creates a potential area of concern, as it may change how they respond to inquiries or drift off topic. It appears as though these issues haven't yet become apparent to respondents.

Among all GenAI model frameworks, OpenAI GPT series is the most popular, with 49% of respondents using it. 23% use Google AI models like PaLM and Gemini, but much smaller groups (less than 10%) use AWS Bedrock, Anthropic Claude, NVIDIA, Facebook, and others.

What GenAI model frameworks have you tried using?

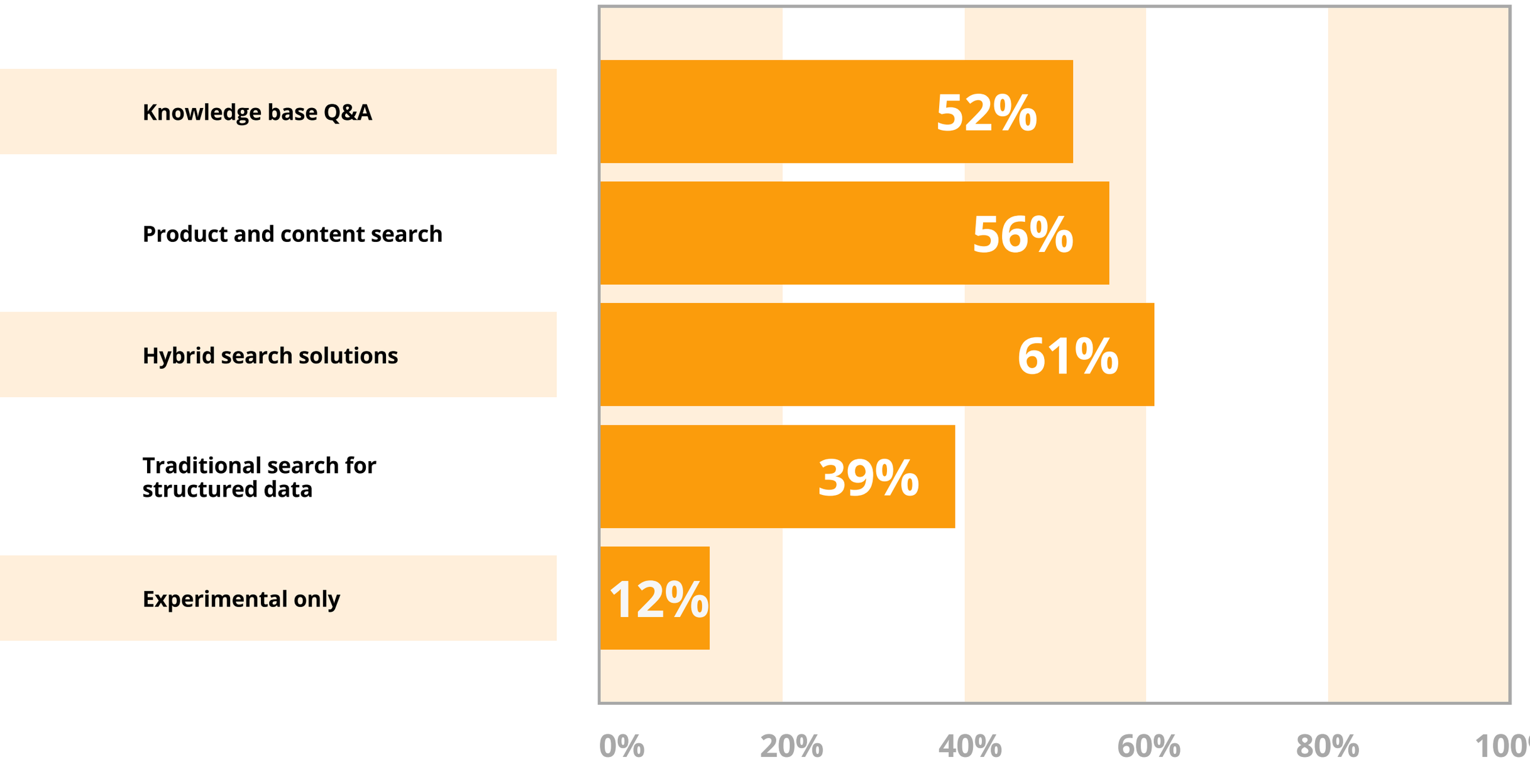


The predominance of OpenAI GPT series is surprising, as the expectation is that model framework popularity would be spread out more broadly and more aligned to the clouds in which they run. This may suggest that model specialization isn't yet necessary, the most accessible models and frameworks are winning, or that projects are still so new they don't yet need the variety of available LLMs. In addition, models may be moving that much faster than the enterprises that use them.

Using GenAI as an interpreter of intention via semantic search has become common across organizations, especially for chatbot-style applications.

Organizations have adopted semantic search for a range of use cases. 61% of respondents rely on semantic search that derives meaning and intention within hybrid queries rather than simple keyword scanning. 56% use this method for product and content search, and 52% use it for knowledge base Q&A.

What are your primary use cases for semantic search versus traditional keyword search?



To compare semantic search versus keyword search performance, most respondents prioritize user engagement metrics (30%) and user satisfaction and feedback (25%). Just 24% measure ranking quality.

This indicates that organizations appear to prioritize real-world impact over technical metrics.

What metrics do you use to compare semantic search versus keyword search performance?

14% - Precision and recall

24% - Ranking quality (NDCG, MRR)

30% - User engagement metrics

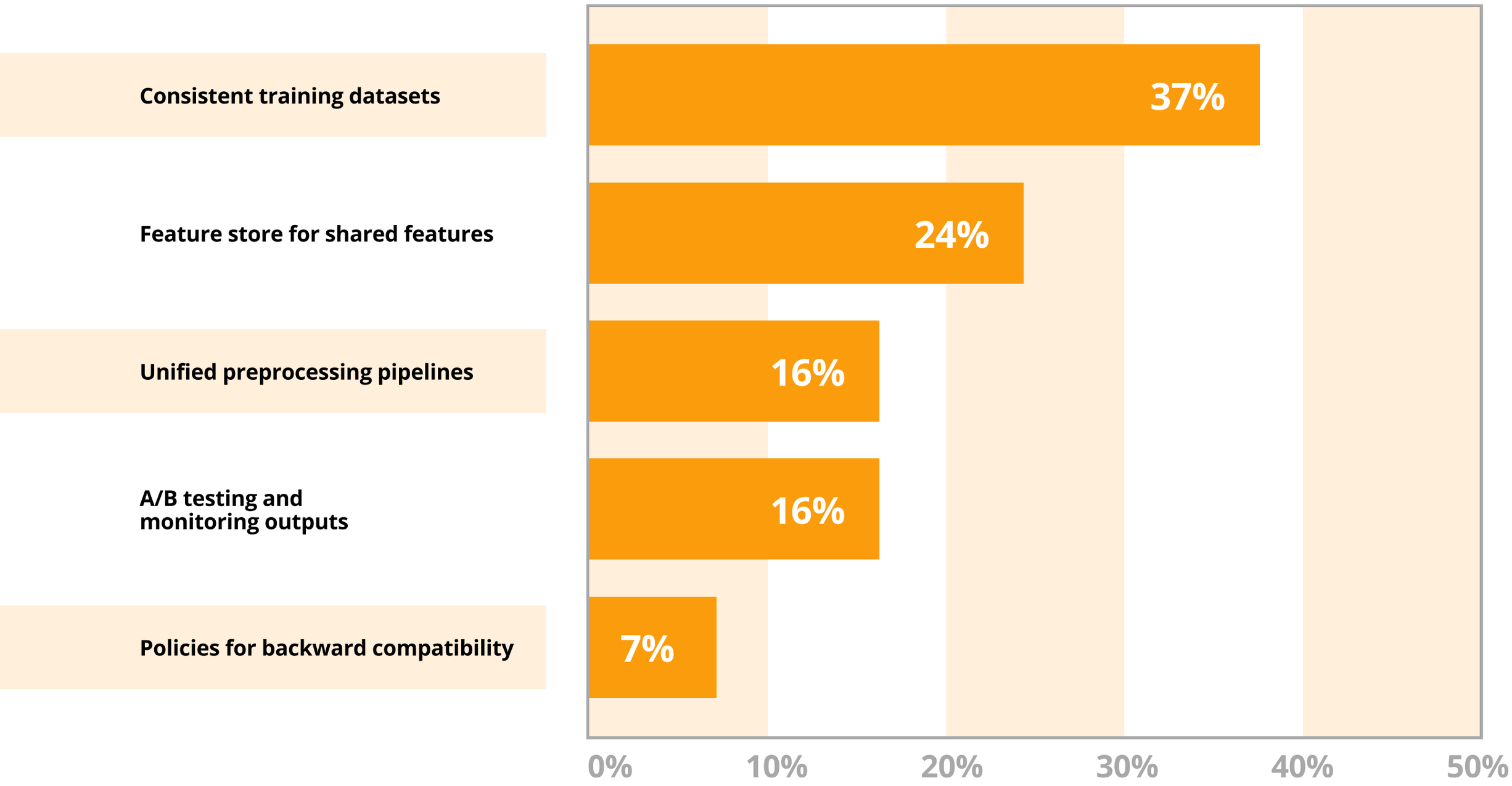
25% - User satisfaction and feedback

7% - A/B testing outcomes

However, these are relatively simple use cases. How will semantic search and semantic caching requirements evolve as applications, models and agents become more complex? How will organizations keep track of system integrity when inputs constantly evolve? Models that become smarter may adjust their opinions, and data inputs to prompts and vectors will change as data utilization increases, while expectations for accurate outcomes will become more strict.

Just 37% of respondents report using consistent training datasets across different AI model versions. The rest use a range of other approaches, including feature store for shared features (24%), unified preprocessing pipelines (16%), and A/B testing and monitoring outputs (16%). This may compromise agentic system reliability as AI models evolve and change their opinions (known as drift) when deployments scale and interaction rates intensify.

How do you maintain data consistency across different AI model versions?



This may also indicate a lack of foresight or immaturity as to why organizations should keep, parse, and store AI interactions and conversational history. With this data, organizations can facilitate session-to-session conversational context and memory, as LLMs can't recollect past interactions independently. Organizations can also preserve conversation transcripts in JSON, using them to validate LLM responses prior to invoking a subsequent action as a result of an agentic interaction.

Validation and Guardrails

Validations and guardrails keep GenAI activities within their intended scope. This post-conversation evaluation step checks for hallucinations (i.e., misleading, unintended responses). If the conversation passes this fail-safe stage, then the software program can proceed to complete the next action. Otherwise, an error or correction activity may occur.

The notion that LLM conversations may drift off topic is real, given that LLM knowledge and prompt data and instructions evolve over time. This necessitates the ongoing validation stage of RAG workflows. There's a growing need to simplify the monitoring process so organizations can easily identify where AI drifts and determine the root cause.

To address these concerns, most organizations have implemented guardrails for GenAI execution. 35% of respondents report comprehensive guardrails, and 33% report basic guardrails. The remaining 31% have no guardrails implemented.

Are you implementing guardrails for GenAI execution?

- **35% -** Yes, comprehensive guardrails
- **33% -** Yes, basic guardrails
- **17% -** Planning to implement
- **12% -** No, not yet
- **2% -** Not applicable

We can read this a couple of ways: These responses may indicate that enterprises have a goal to balance AI safety with continued development or that they lack knowledge about establishing appropriate guardrails, which often require constant attention and evaluation after every LLM interaction. The latter is a key, underserved step in RAG workflows.

“Use AI to solve creative and out-of-band problems, but then integrate it into logic flows whereby you check AI's output before moving to the next step. Sometimes, you may need to send AI back to the drawing board to try the solution again. Sometimes, it may never work. Other times, it will produce the expected result in a fraction of traditional approaches.”

Adrian Talapan, CEO and Founder at Qreli

RAG operation is cyclical. It involves multiple interaction stages that include the assembly, preparation, and vectorization of prompts and input data at the front end; the capture and utilization of the conversation transcripts generated as a result of LLM interactions; the post-conversation development and use of response validation techniques; and the creation of durable guardrails to keep agents from drifting off task.

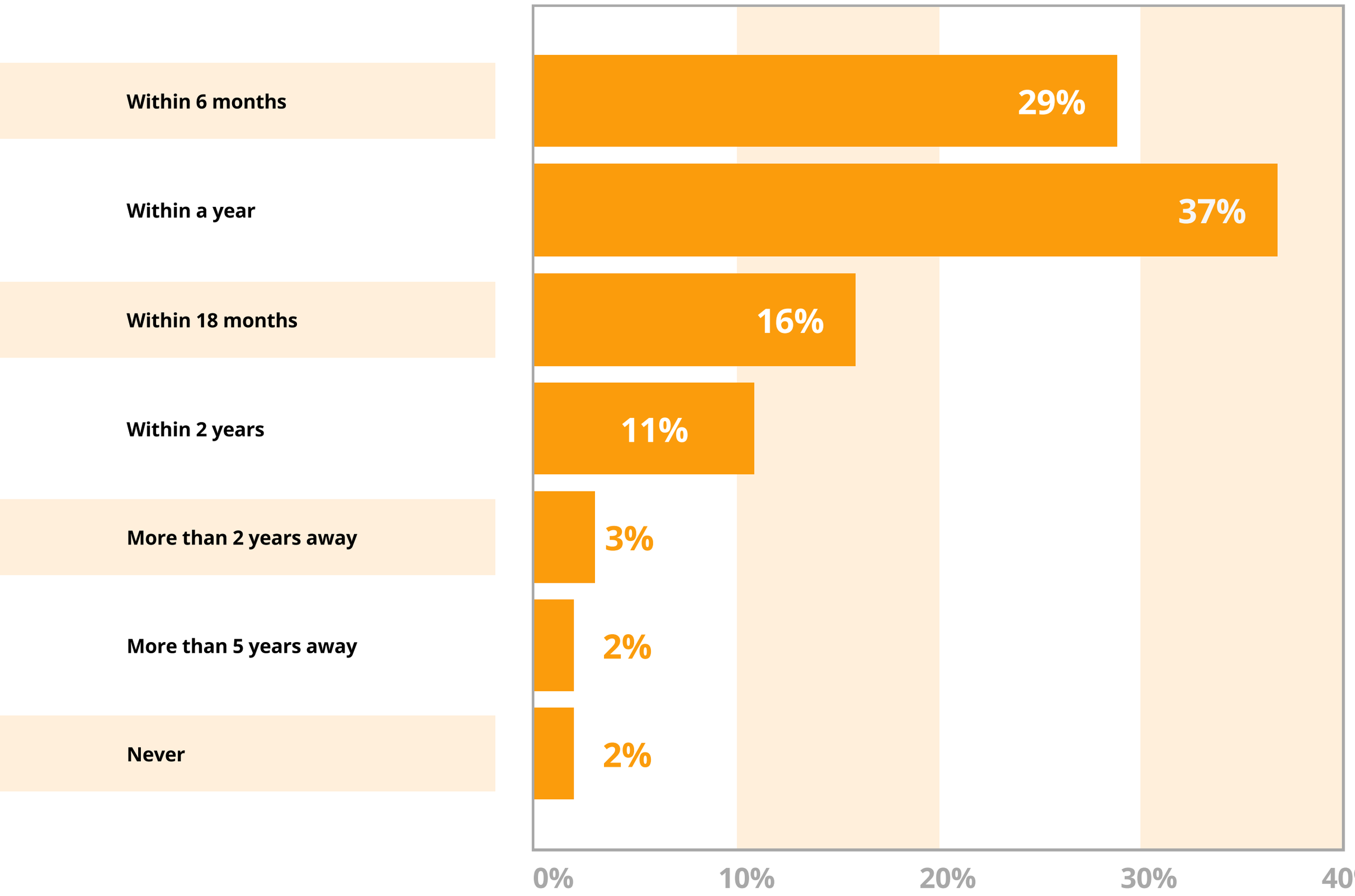
Once these RAG applications and agents are tested, and deployed, how do they meet the performance expectations of their constituents?

Agent Catalogs and Managing the AI Data Lifecycle

Agentic systems are built from collections of (semi) autonomous agents that perform predefined tasks and workflows while interacting with GenAI models for guidance, wisdom, and suggested next steps. Some consider AI agents to be the new microservices.

Most organizations are already moving forward with autonomous AI plans. 66% of respondents intend to deploy AI agents within a year or less, and 29% plan to do so within six months.

When will you begin deploying agents?

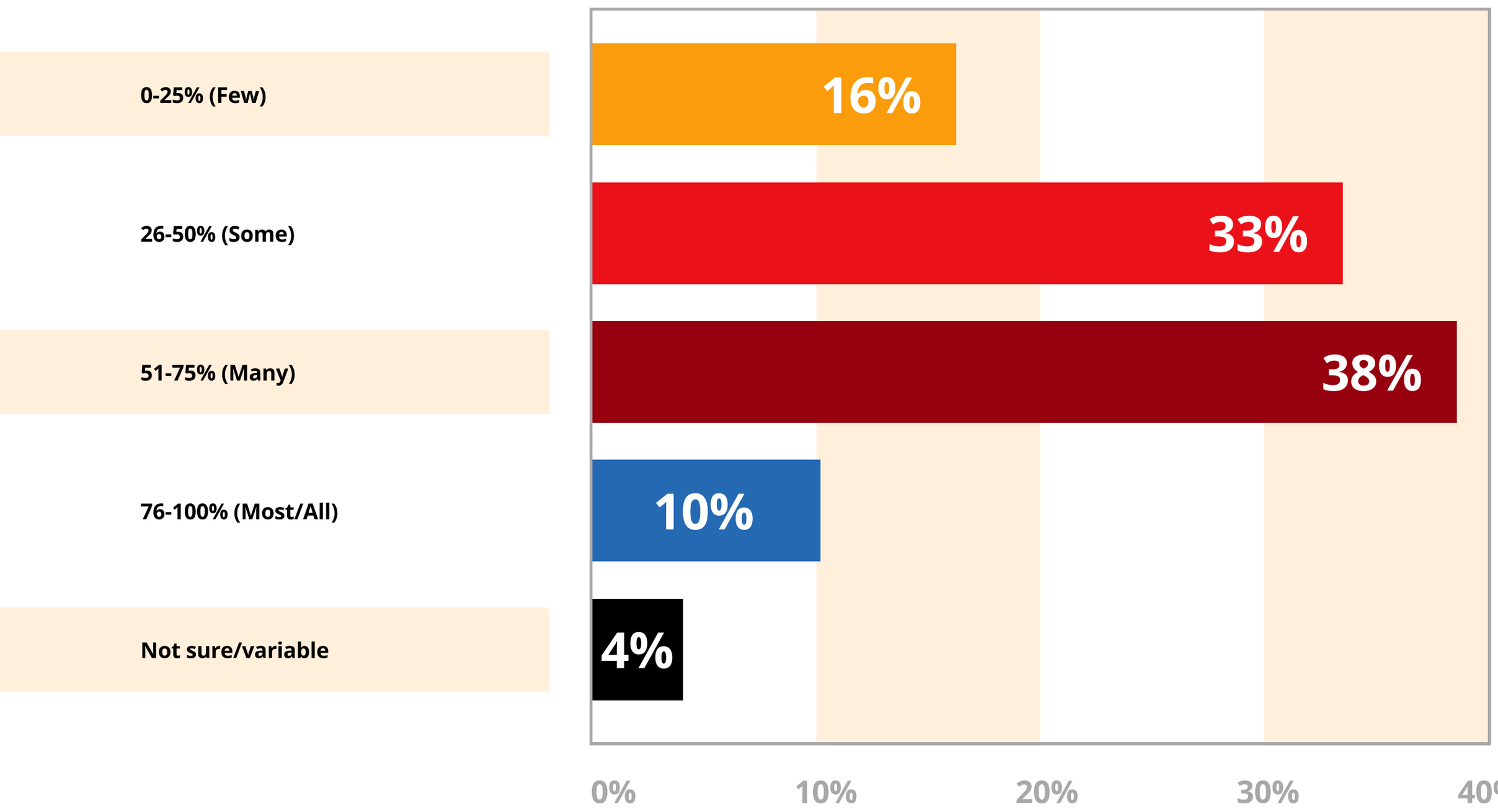


FOMO and competitive pressure appear to outweigh concerns about security or hallucinations. Although expectations and confidence are high, reconciling issues around using corporate data safely and securely and trusting conversational accuracy consumes the conscience of most respondents.

In addition to managing issues with the data supply chain for RAG, enterprises must consider performance and scaling expectations for agentic systems once they're deployed. ***RAG needs to move at lightspeed.*** *However, slow, unruly data can become an impediment to successful AI applications and agentic systems.*

Most respondents have significant AI-related performance demands. 48% say more than half of their applications require real-time database capabilities for AI functions, while only 16% say less than a quarter of their AI functions need a real-time database.

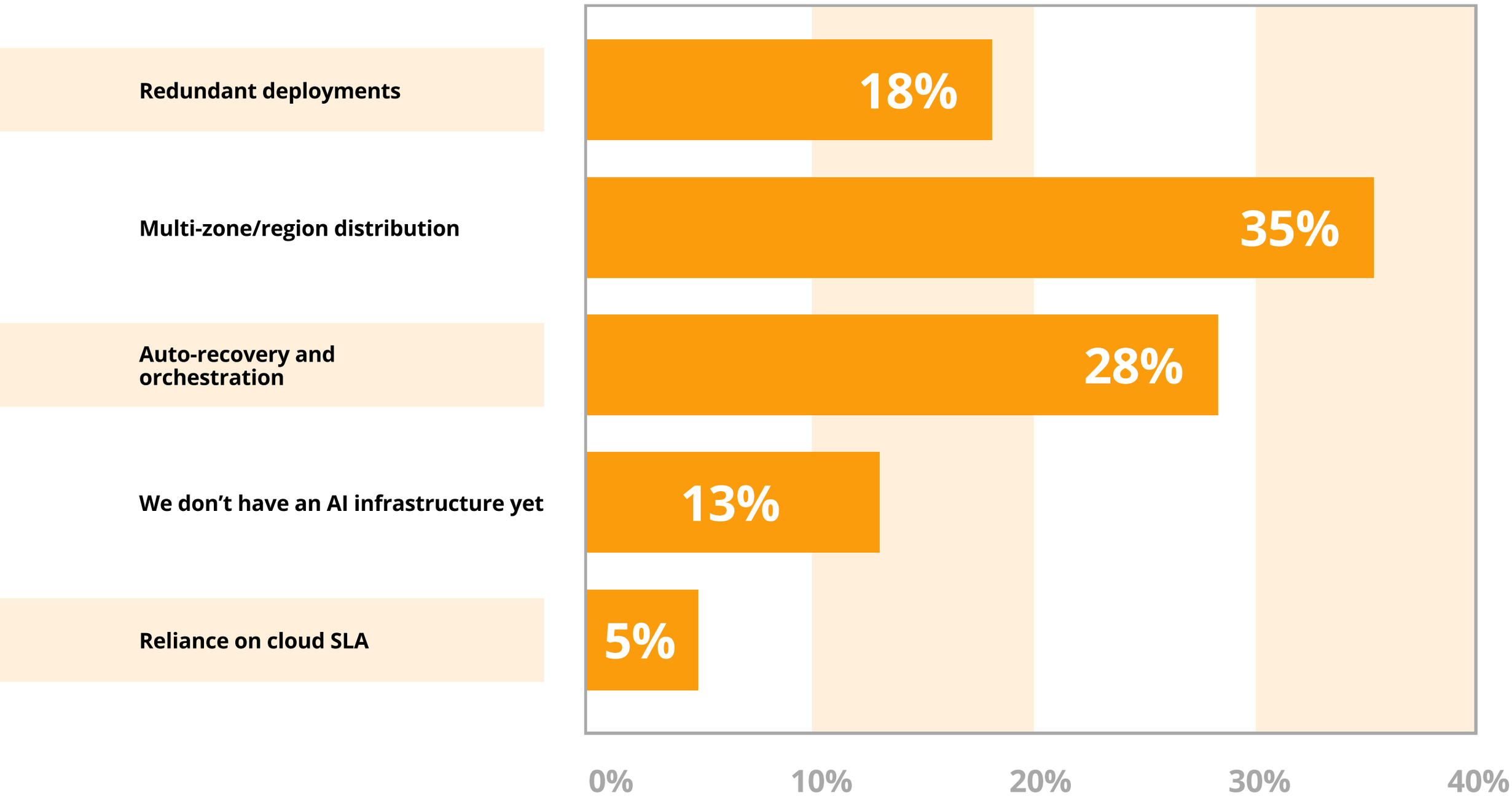
What percentage of your applications require real-time database capabilities for AI functions?



AI response cycles are time-sensitive, which leaves little room for slow data processing within the organization. To make AI work effectively, every millisecond counts. If AI is real time, organizations must design for that throughout the RAG workflow.

When considering how to balance high availability with fault tolerance in their AI infrastructure, organizations tend to prioritize geographic distribution (providing availability) over redundancy (favoring fault tolerance). 35% of respondents rely on multi-zone or region distribution. 28% use auto-recovery and orchestration tools, and 5% rely on cloud provider service-level agreements (SLAs), indicating that the primary expectation is ensuring AI systems are available — likely to end user audiences.

How do you balance high availability and fault tolerance in your AI infrastructure?



The lower need for redundant deployments (18%) indicates that these organizations don’t view AI interactions as relevant to persist (store) or as important to provide durable financial transactions, for example.

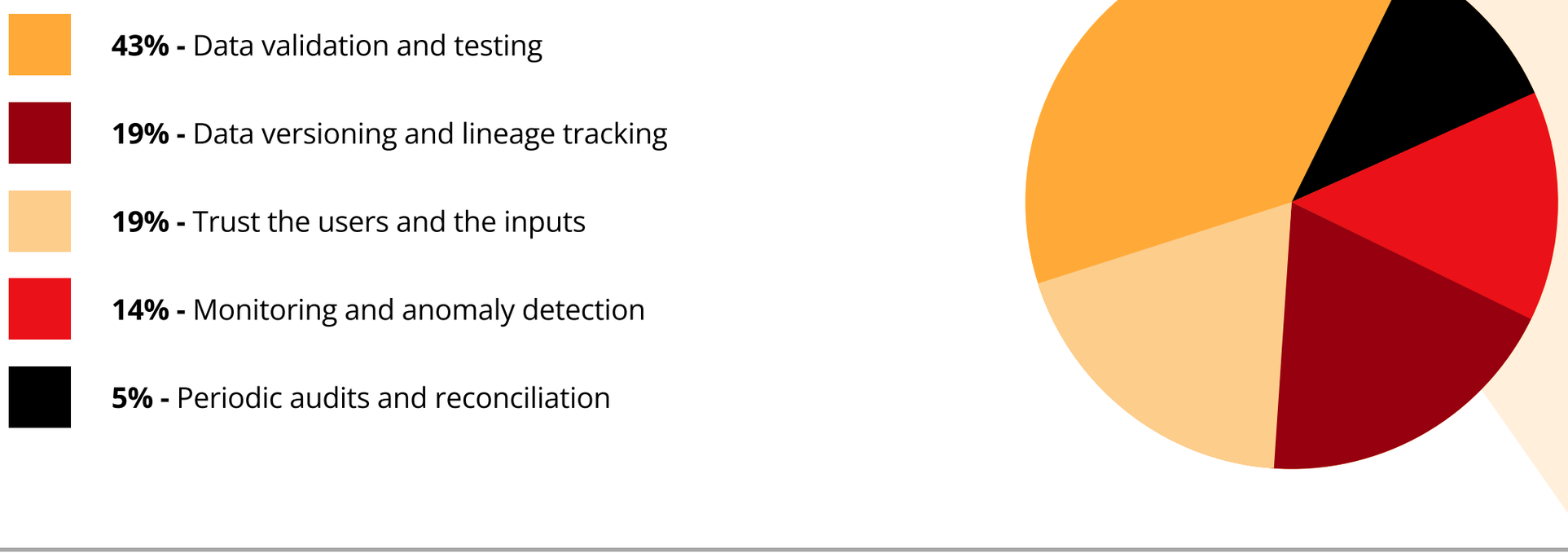
This suggests AI applications are becoming an always-on, real-time activity that may be transient. But maintaining durable, permanent records of AI interactions is less vital. However, organizations must consider how they’ll run high-performance, trustworthy agents when they have strong fears about their actual activities.

The ability to ensure data integrity across the AI application ecosystem becomes important once teams understand the multi-stage nature of building and running AI-enabled applications or agents. This includes how their software programs consume, use, repurpose, and recycle data and how agents must talk to other agents independently and at scale.

For most respondents, data validation and testing (43%) are the means to ensure stability and avoid contradictions and errors in data. Only 19% of organizations are versioning their data while tracking its lineage, and another 19% are (potentially dangerously) trusting users and inputs alone.

Only 14% have implemented monitoring and anomaly detection, and only 5% use audits to reconcile issues. These varied responses illustrate the complexities of data management, including trusting its contents and values and assuring its ongoing correctness while running AI.

How do you ensure data integrity across your AI application ecosystem?



Capturing and preserving software code along with relevant agentic correspondence could dramatically improve developer productivity and provide active visibility into the ongoing behavior of an agent or agentic system. But a catalog for agent code and related content can do much more.

It could use its own AI to evaluate agentic behavior while facilitating and suggesting proper guardrails to prevent drift. It could also evaluate performance latency or interruptions and facilitate code reuse for subsequent agents. In addition, it could facilitate interactions and publishing to other agent catalogs using model control protocol (MCP). MCP is the interface agents use to identify, register, and work with other agents, including those from other domains and catalogs.

While there are many uses for an agent catalog, it becomes even more valuable when it includes the software code as well as artifacts like conversation transcripts, version numbers, vectors, and other metadata that can improve the efficiency of agentic RAG workflows.

Conclusion: Building AI-Ready Data Architecture

When we apply our data-architectural lens to the process of building agentic application systems that embed RAG workflows, we see a disorganized mess throughout the process. Our observations are as follows:

- The focus of both analytics and operational AI-powered applications and agentic systems is clearly shifting from data management to facilitating GenAI usage. But right now, complexity simultaneously lies in both areas, which may be the reason organizations have deployed so few agentic systems.
- Enterprise organizations split activities almost evenly between analytic projects and internal or external application projects. This indicates that organizations don't have a clear priority for GenAI use cases.
- Data safety and security are primary concerns, while a close second is avoiding and managing hallucinations. The RAG technique and data workflow is designed to address these issues simultaneously.
- The convenience of working with brand-name LLMs and AI frameworks is the current state of the art. However, it's likely that less popular and more specialized models will evolve as applications become more sophisticated.
- Coding assistants make developers more productive, and GenAI chatbots are being widely deployed. Organizations are deploying fewer agentic systems, but they're optimistic this will change by 2027.
- Data for AI-enabled use cases flows through many different systems, preparations, transformations, AI interactions, and response validations before any RAG-enabled workflows perform any actual work.
- Organizations are less aware of how the lack of a unified data management platform is slowing their ability to produce breakthrough agentic applications while they work their way up the RAG learning curve.
- Text represents a significant portion of RAG workflow data. Text is best saved, stored, and parsed as JSON, which is AI's data format.
- Application performance is a lesser concern because most applications that require scale and performance are operating at "good enough" levels. These expectations will grow with adoption.
- RAG workflows are cyclical and must operate at millisecond speed. This is an undiscovered issue that organizations could correct by supporting the entire RAG data lifecycle from a unified data platform.
- Agentic systems will bring significant value to enterprises large and small by automating both internal and external workflows.

"The next step in building AI-ready data architecture is unifying siloed systems, securing data throughout its lifecycle, and optimizing for GenAI friendly formats like JSON while supporting advanced workflows like RAG. This approach enables organizations to scale AI agents and unlock deeper, more actionable insights."

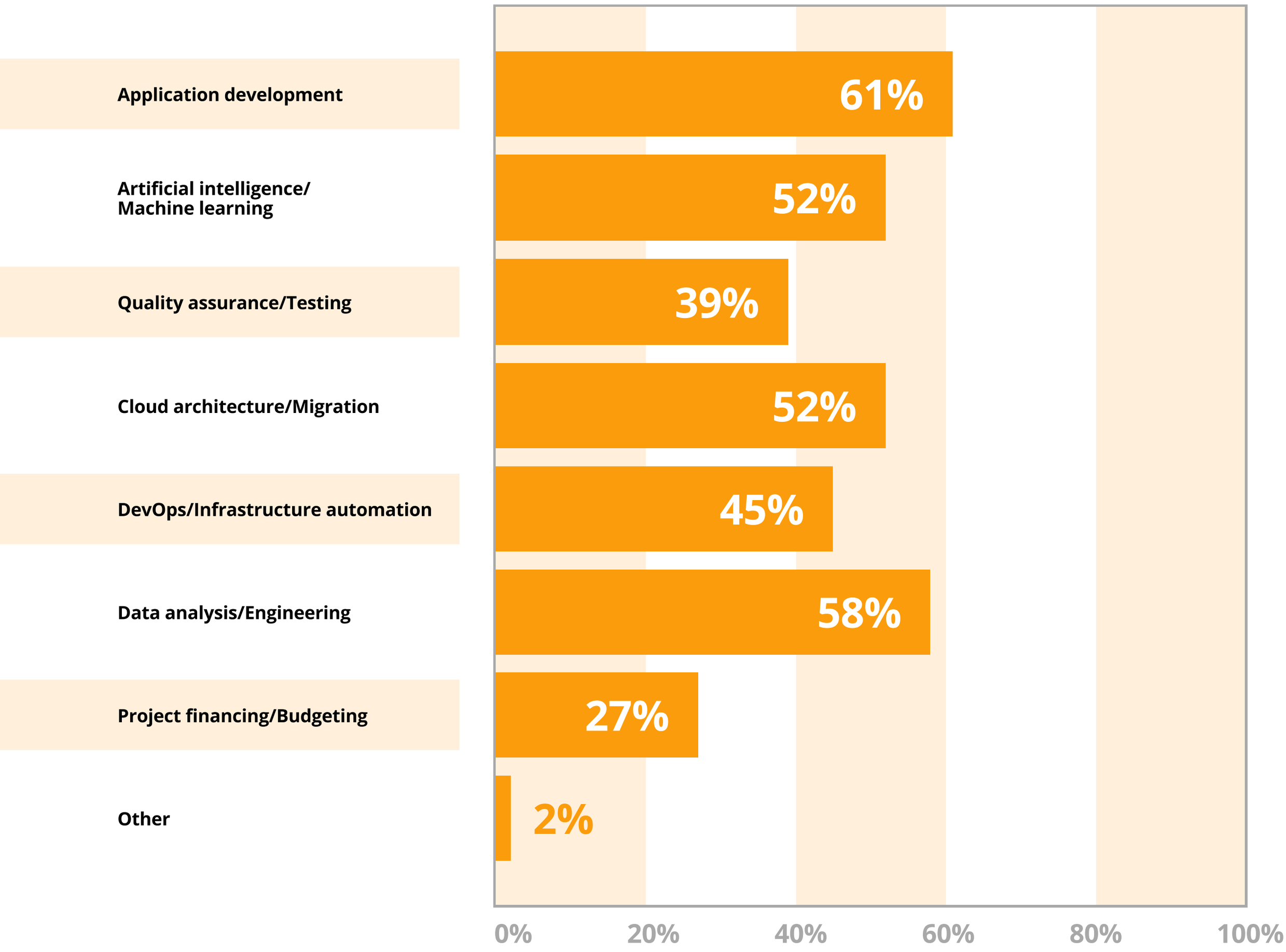
Mohan Varthakavi, VP of Software Development, AI, and Edge at Couchbase

Methodology and Demographics

Couchbase commissioned an independent market survey from UserEvidence of 619 product, engineering, data, and AI professionals. The research sample was vendor-neutral and did not target Couchbase or UserEvidence customers, although they weren't excluded from participating.

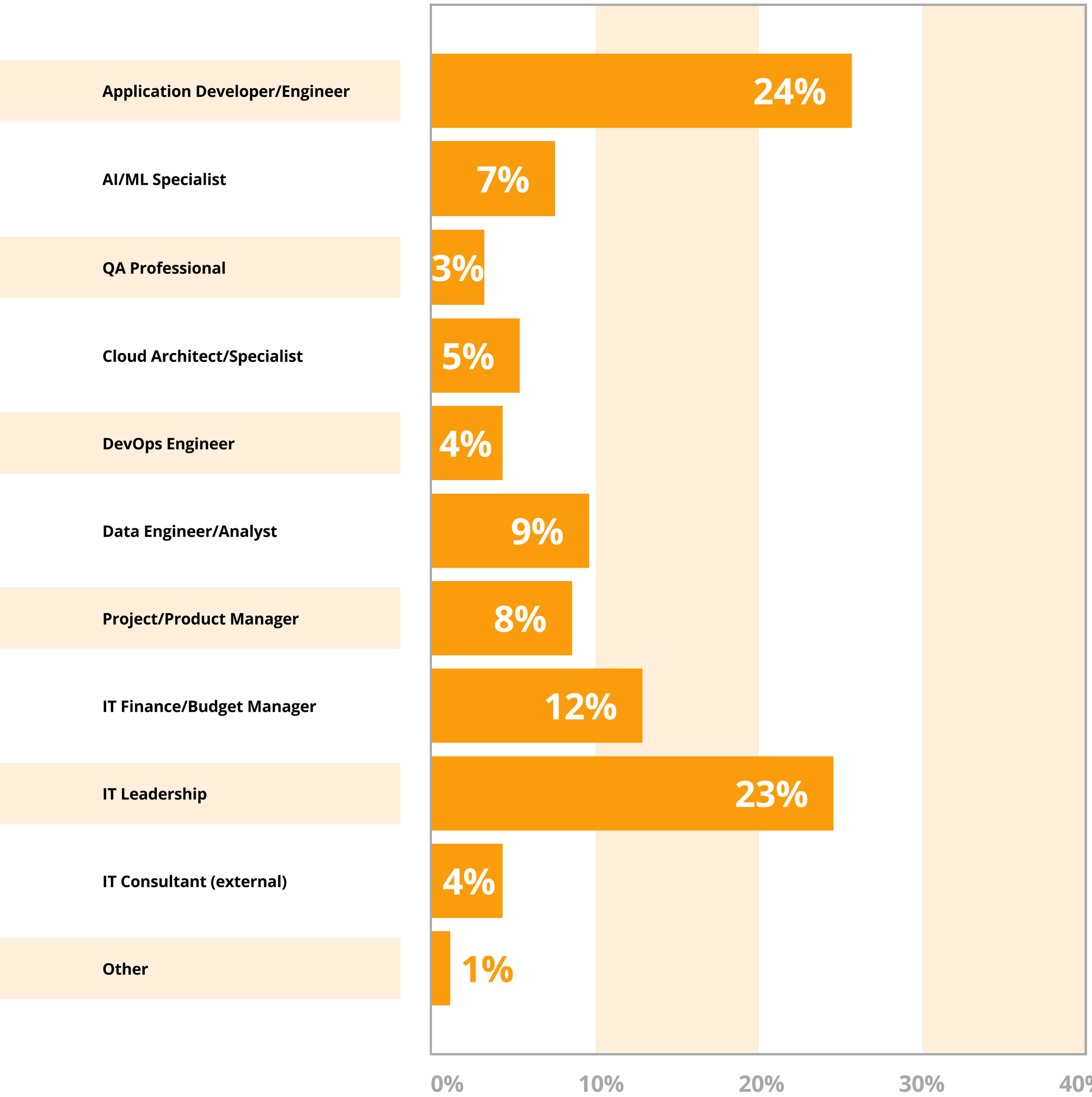
Respondents had experience with application development (61%), data analysis and engineering (58%), cloud architecture and migration (52%), and devops and infrastructure automation (45%) among other professional areas.

Which of the following areas do you have professional experience with?



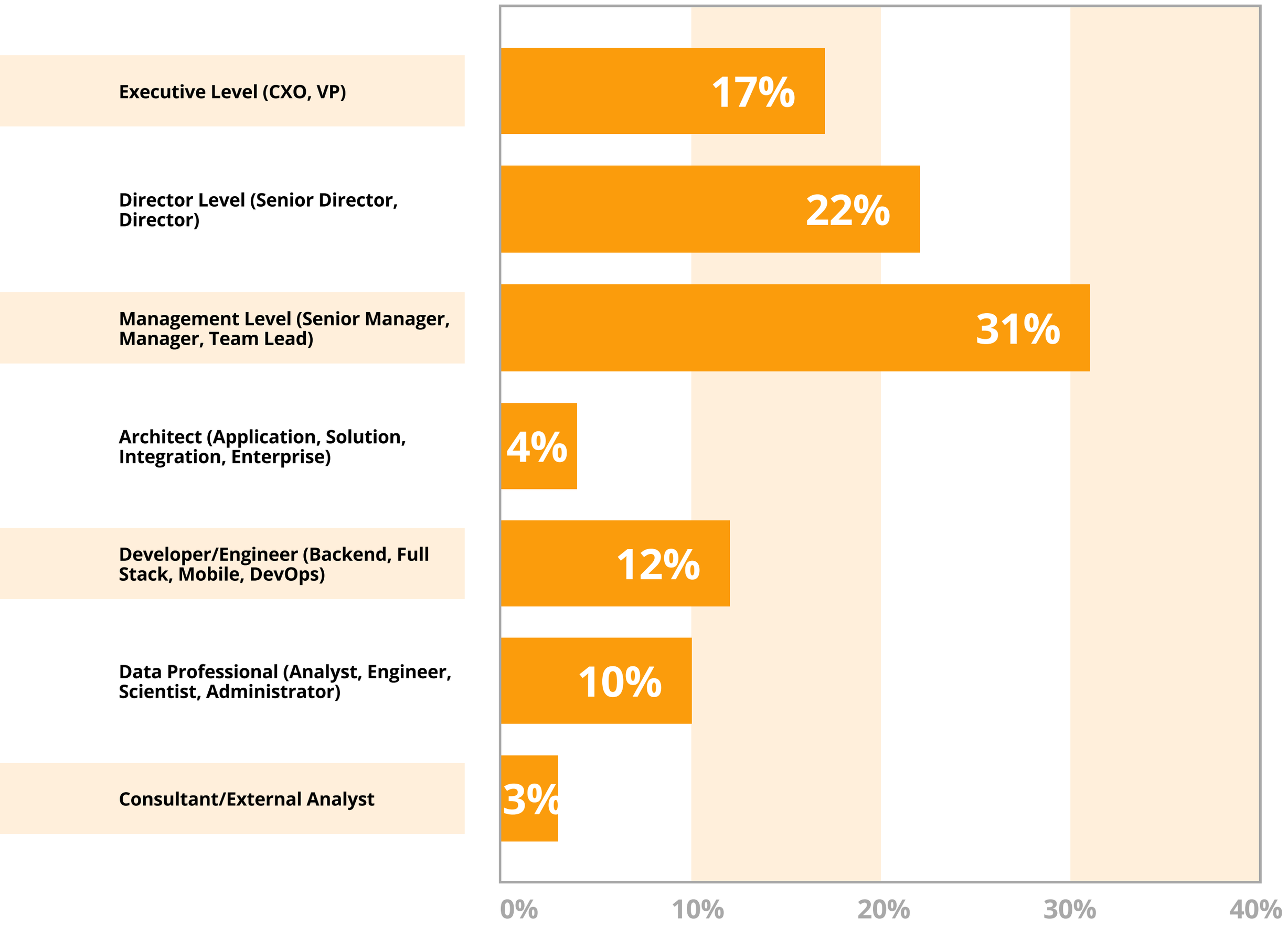
Most respondents held primary roles as application developers or engineers (24%), IT leadership (23%), or IT finance or budget managers (12%). Other roles included data engineer or analyst (9%), project or product manager (8%), and AI or ML specialist (7%).

Which of the following best describes your primary role in IT?



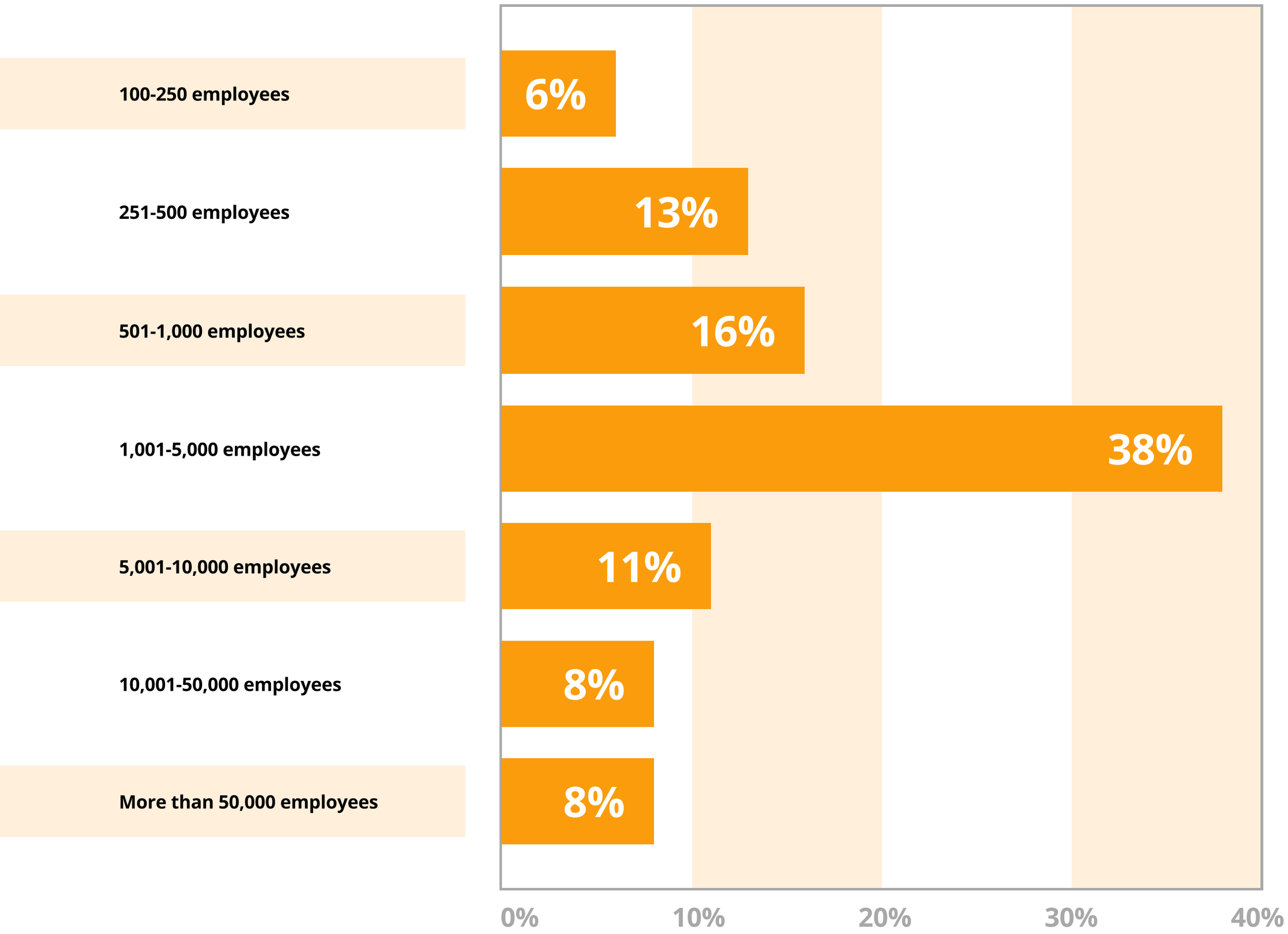
The majority of respondents held leadership positions, with 31% at management level, 22% at director level, and 17% at executive level. The remaining 29% were practitioners and consultants.

Which of the following best describes your current position?



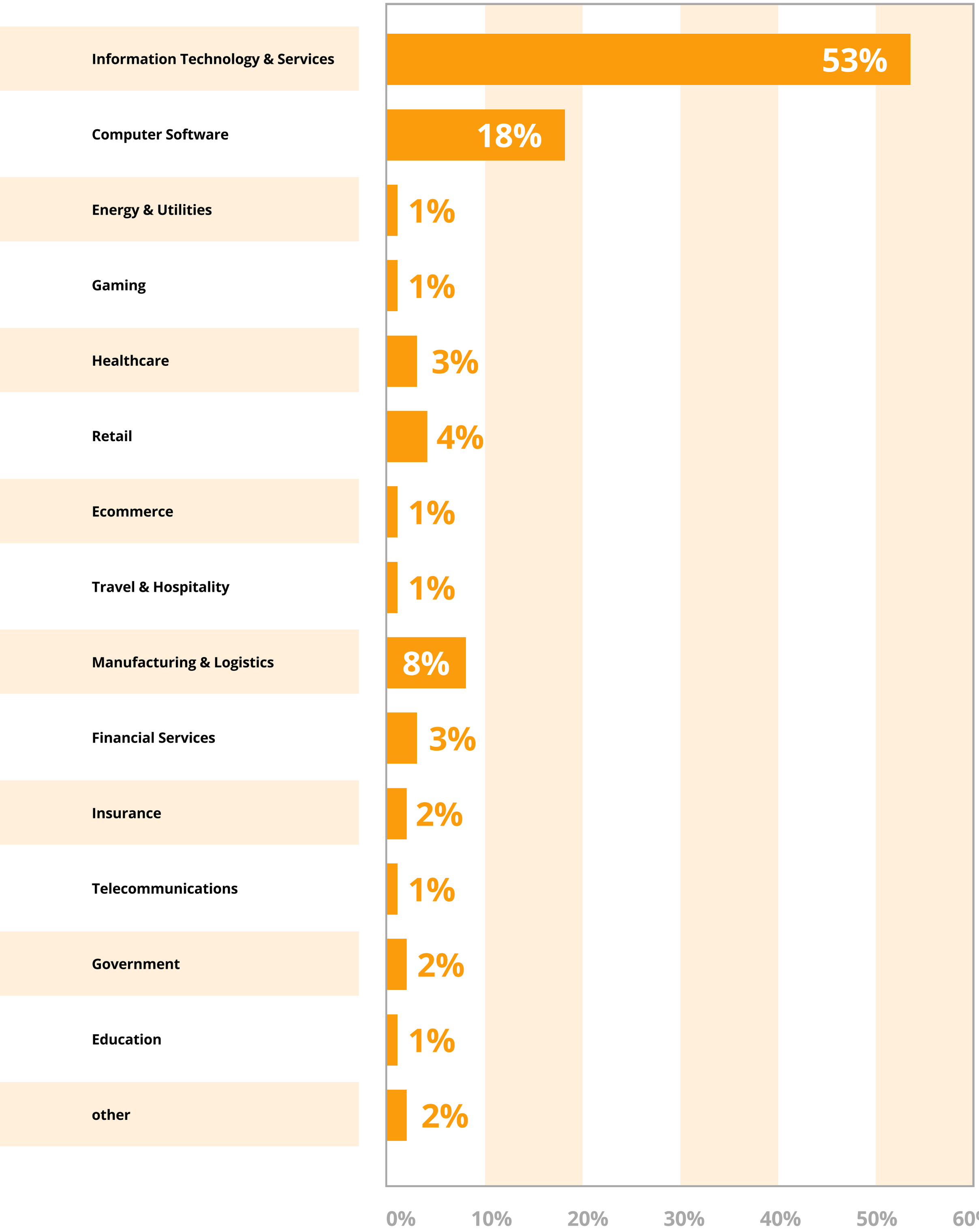
All respondents represented organizations with at least 100 employees. Nearly two-thirds (65%) of respondents worked for organizations with more than 1,000 employees, while 27% represented organizations with more than 5,000 employees.

What is the approximate size of your organization (by number of employees)?



The vast majority of respondents' organizations operated in the IT and services (53%) and computer software (18%) industries. The remaining operated in manufacturing and logistics (8%), retail (4%), healthcare (3%), financial services (3%), and other industries.

In which industry does your organization primarily operate?



About UserEvidence

UserEvidence is a software company and independent research partner that helps B2B technology companies produce original research content from practitioners in their industry. All research completed by UserEvidence is verified and authentic according to their research principles: Identity verification, significance and representation, quality and independence, and transparency. All UserEvidence research is based on real user feedback without interference, bias, or spin from our clients.

UserEvidence Research Principles

UserEvidence is a software company and independent research partner that helps B2B technology companies produce original research content from practitioners in their industry. All research completed by UserEvidence is verified and authentic according to their research principles: Identity verification, significance and representation, quality and independence, and transparency. All UserEvidence research is based on real user feedback without interference, bias, or spin from our clients.

1. Identity Verification

In every study we conduct, UserEvidence independently verifies that a participant in our research study is a real user of a vendor (in the case of Customer Evidence) or an industry practitioner (in the case of Research Content). We use a variety of human and algorithmic verification mechanisms, including corporate email domain verification (i.e., so a vendor can't just create 17 Gmail addresses that all give positive reviews), and pattern-based bot and AI deflection.

2. Significance and Representation

UserEvidence believes trust is built by showing an honest and complete representation of the success (or lack thereof) of users. We pursue statistical significance in our research, and substantiate our findings with a large and representative set of user responses to create more confidence in our analysis. We aim to canvas a diverse swatch of users across industries, seniorities, personas—to provide the whole picture of usage, and allow buyers to find relevant data from other users in their segment, not just a handful of vendor-curated happy customers.

3. Quality and Independence

UserEvidence is committed to producing quality and independent research at all times. This starts at the beginning of the research process with survey and questionnaire design to drive accurate and substantive responses. We aim to reduce bias in our study design, and use large sample sizes of respondents where possible. While UserEvidence is compensated by the vendor for conducting the research, trust is our business and our priority, and we do not allow vendors to change, influence, or misrepresent the results (even if they are unfavorable) at any time

4. Transparency

We believe research should not be done in a black box. For transparency, all UserEvidence research includes the statistical N (number of respondents), and buyers can explore the underlying blinded (de-identified) raw data and responses associated with any statistic, chart, or study. UserEvidence provides clear citation guidelines for clients when leveraging research that includes guidelines on sharing research methodology and sample size.

About Couchbase

As industries race to embrace AI, traditional database solutions fall short of rising demands for versatility, performance and affordability. Couchbase is seizing the opportunity to lead with Capella, the developer data platform architected for critical applications in our AI world. By uniting transactional, analytical, mobile and AI workloads into a seamless, fully managed solution, Couchbase empowers developers and enterprises to build and scale applications and AI agents with complete flexibility—delivering exceptional performance, scalability and cost-efficiency from cloud to edge and everything in between. Couchbase enables organizations to unlock innovation, accelerate AI transformation and redefine customer experiences wherever they happen. Discover why Couchbase is the foundation of critical everyday applications by visiting **www.couchbase.com** and following us on **LinkedIn** and **X.**

Couchbase