bright data

Data for AI 2026

The rise of web data infrastructure

Executive summary

Organizations developing AI systems are operating in a high-velocity, high-stakes environment where real-time access to public web data has shifted from an advantage to a necessity. A survey of 500 AI practitioners from companies that are building AI systems was conducted by Vanson Bourne in February of 2026, focusing on current AI applications, tools, and the near future.

Consistent with previous years' surveys on public web data for AI, nearly all organizations say real-time data is necessary for their AI, and data consumption continues to increase. This year's survey found a 132% average increase in real-time data usage.

The increase in data consumption for AI aligns with the necessary foundational web data infrastructure layer that powers all AI. The old web must connect with the new web, agents must be able to interact with and retrieve information, fresh data must be accessible to predictive or foundation models, and available for use to train robotics. Every point of data retrieval, relies on the web data infrastructure.

The critical web data layer is becoming more difficult to access and materially limiting AI initiatives. With the current challenges and expectations for restrictions to worsen in the next year, a reliable data partner is a competitive advantage to succeed.

Table of Contents

Agentic Adoption

Agentic Deployment is Facing an AI data Infrastructure Problem

Moving from development to production requires more reliable data delivered faster.

Biggest challenges in scaling AI systems

- Cost of infrastructure and operations
- Access to reliable, up-to-date external data
- Internal integration and orchestration complexity
- Talent or skills gap

Top data requirements in next 12 months

- **59%** Greater reliability and data quality
- **58%** Faster data collection or retrieval speeds
- **52%** More real-time data
- **50%** Stronger compliance, governance, or legal safeguards
- **46%** Access to a greater volume (quantity) of data
- **45%** Broader (comprehensive) access to web-based data sources

97% of Organizations Use a Combination of AI Agents Types to Connect to the Real-time Web

The top 3 benefits of different types of agents

Benefits of Using Enrichment Agents

Improved accuracy or data quality
61%

Better customer, supplier, or market insights
57%

Faster decision-making
56%

Benefits of Using Research Agents

Improved accuracy or quality of research findings
58%

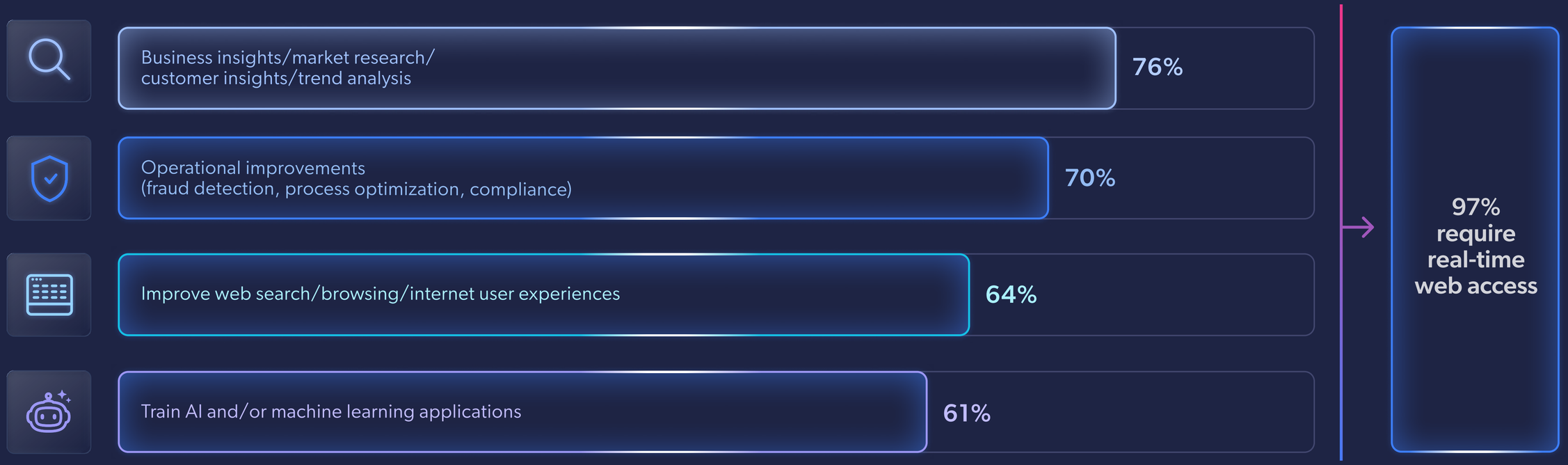
Better strategic, market, or competitive insights
55%

Faster time from research question to actionable insight
52%

AI Deployment Dependent on Real-Time Access to Web Data

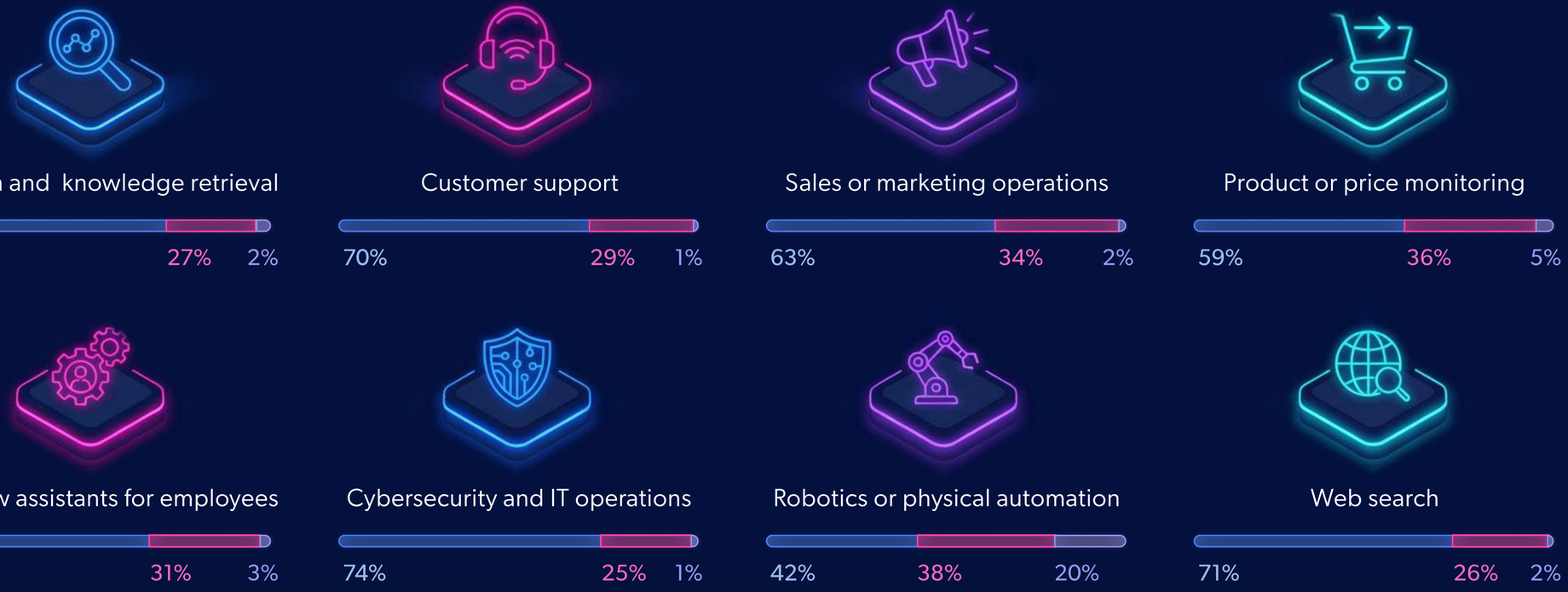
Most respondents have implemented AI in multiple ways for different functions and often they support each other. Agents may be used to retrieve information for the foundation or predictive models which serves one of the five use cases that rely on real-time access to the public web.

Top uses cases by function



Agents Relying on a Real-time Web Connection to Support All Areas of Business

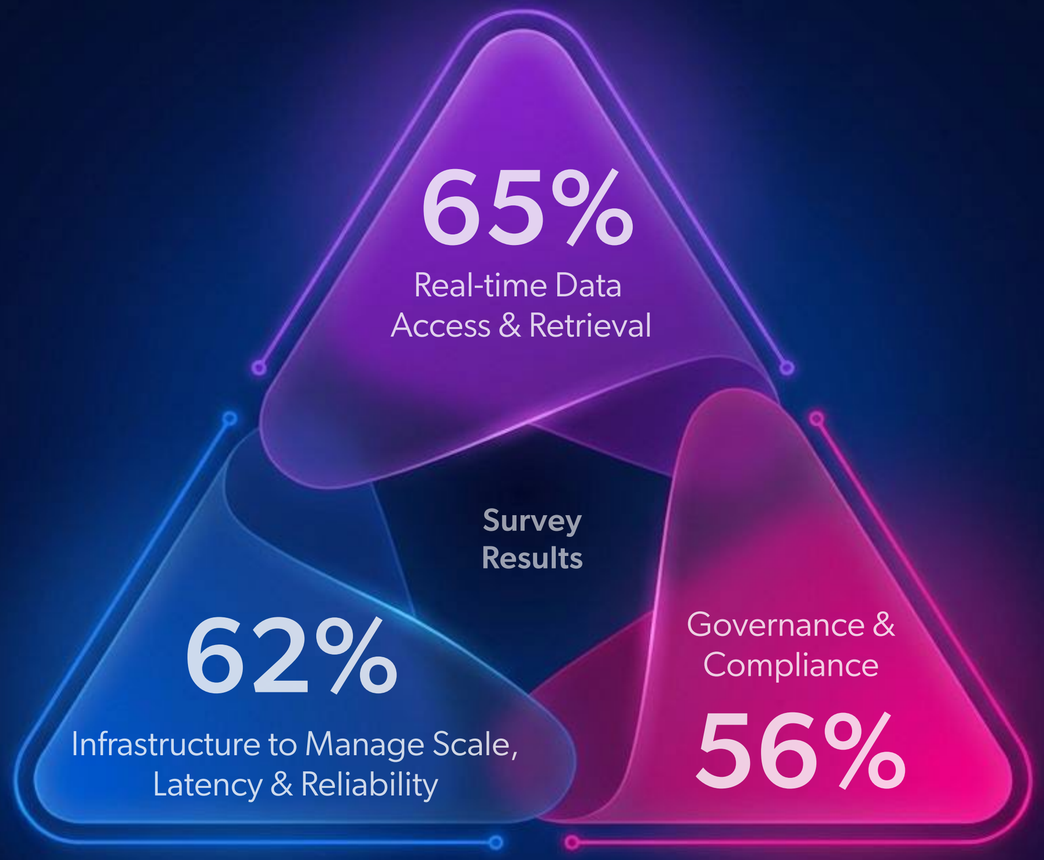
60% of AI leaders that work across verticals and business functions say their primary output is agents connected to the real-time web, which is required for the majority of tasks. On average, AI leaders report they deploy agents connected to the real-time web for a total of 5 business functions.



Web Execution

Building for the Future Web: Critical Infrastructure Needs

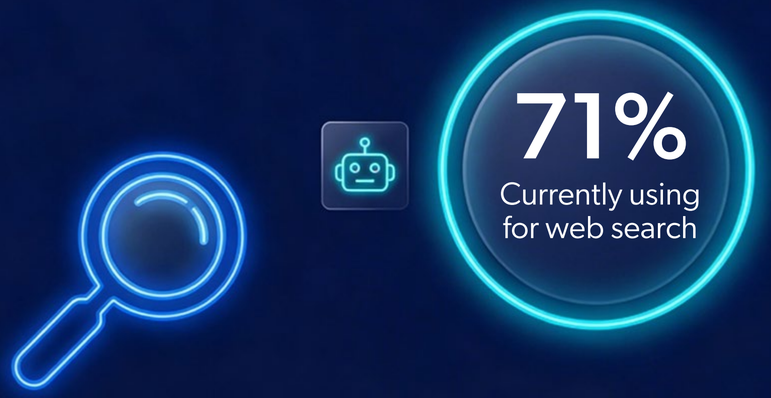
There are three non-negotiable pillars of AI web infrastructure that are necessary to support the future web. AI leaders recognize real-time data access and retrieval as the most important, scoring above infrastructure to manage scale, latency, and reliability and governance and compliance.



The Emerging Two-Tier Internet

Infrastructure that can reliably and compliantly navigate the open web becomes mission critical.

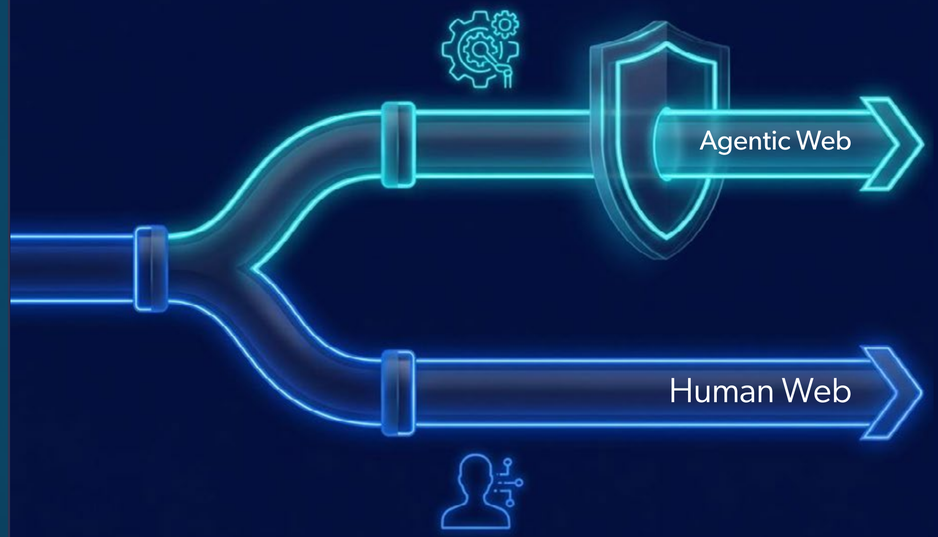
AI Agents for Web Search



71%

Currently using for web search

Two-Tier Internet Emergence



87%

Organizations that AGREE a "two-tier internet" is emerging

The Shift to an Agentic Web: Timeline Predictions

The web is changing from human to agentic, and competition is leveraging this for efficiency.
Here's how quickly AI leaders see this change happening.

Web Access is Mission Critical for Agentic Operations

Every organization recognizes multiple factors driving the necessity of real-time data.

Top 6 reasons businesses need real-time web access

56%
Improved trust in AI outputs

54%
Competitive pressure to respond to real-time market changes

51%
Increasing customer expectations

49%
Information changes too quickly for static training data

42%
Reduced reliance on frequent retraining cycles

39%
Need for up-to-date signals from the public web

Foundation Models

Data Volume Growth is Outpacing Internal Infrastructure Maturity

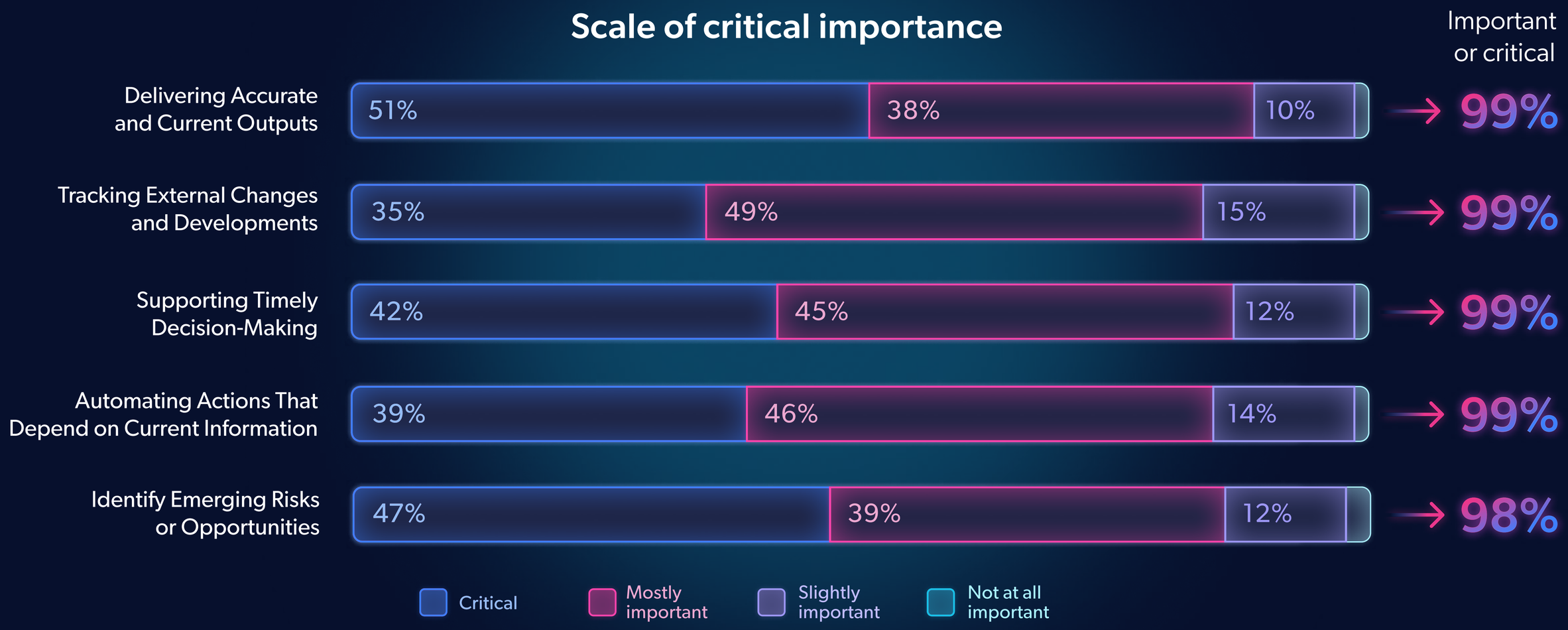
In the past 12 months, organizations used on average 132% more data to train their models than the previous 12 months.

Percentage of respondents who face these challenges in finding, cleaning and processing public web data for AI

85%	Ensuring data quality and completeness
81%	Scaling data collection across multiple regions
80%	Managing evolving data privacy regulations
79%	Consistency and reliability of data sources
79%	Navigating legal restrictions
78%	Integrating unstructured data
75%	Compliance
75%	Timeliness of data acquisition

Relying Solely on Training Data is No Longer an Option to Operate AI

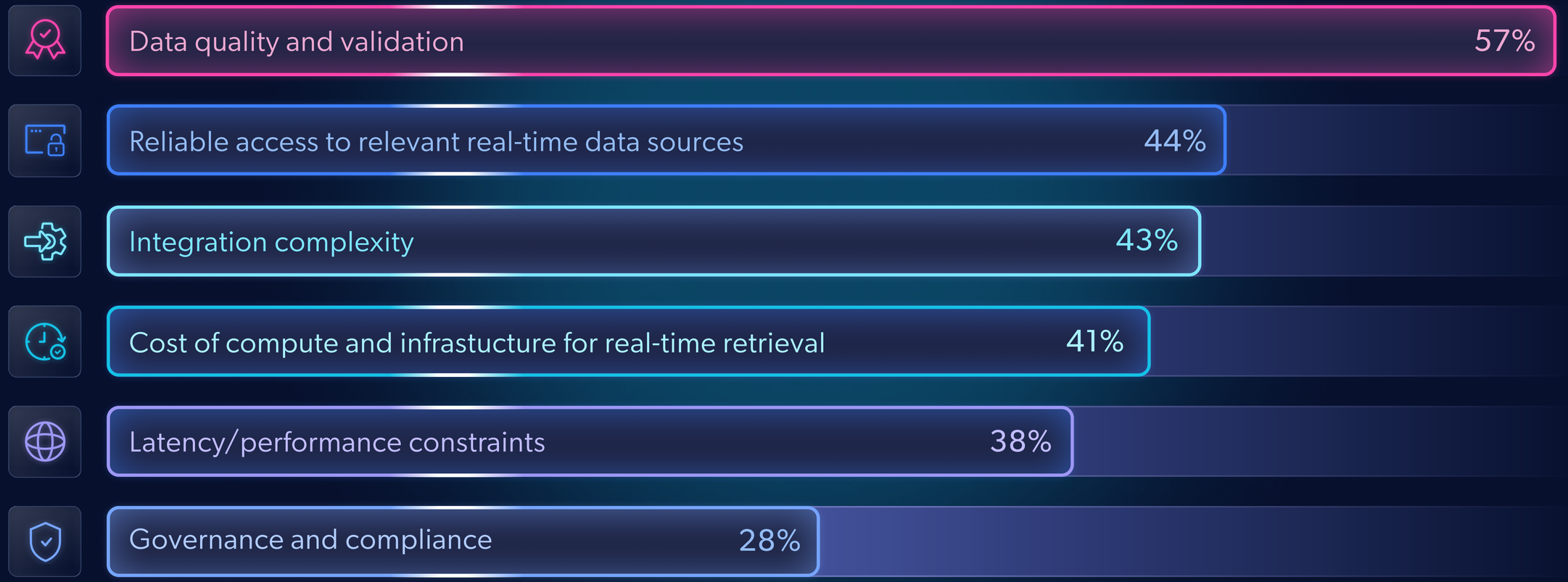
Respondents overwhelmingly (98-99%) said all the following are important or critical points about real-time and refreshed data for models. Additionally 82% of respondents said relying on stale datasets make it difficult for AI to remain accurate.



Access and Integration are the Bottleneck

Challenges when enabling real-time reasoning in AI

Top challenges AI leaders face



Robotics

Robotics + Foundation Model Overlap (Cross-Reference Insight)

Organizations using robotics training data also report:

Foundation model usage

85%

Predictive model usage

79%

Robotics training organizations are aligned with foundation model-centric stacks, where fresh external data and strong pipelines become a multiplier.

Notable qualitative signals

- Interest in “Functional AI”
- Multiple manufacturing references to perception + manipulation models in production environments

Robotics Training Data: Volume and Modality Shifts

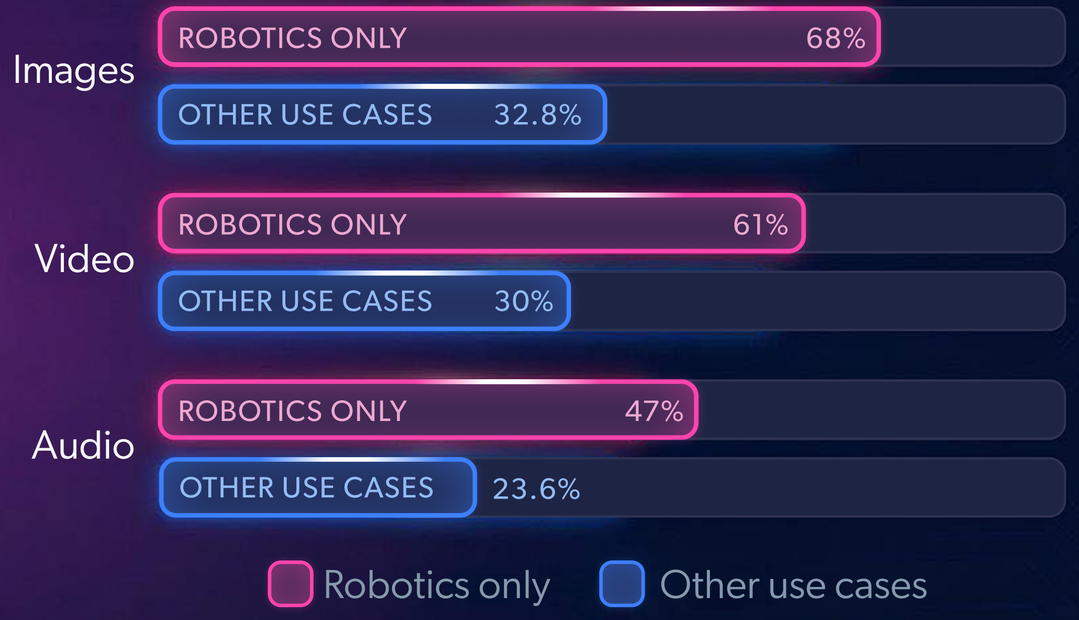
Average training data volume increase

+133%

Growth in gata volume

Significant surge in data acquisition across all sectors

Robotics training modality preferences



Takeaway: Multimodal data acquisition is more pronounced in robotics, consistent with perception + manipulation training needs.

AI Agents in Robotics: The Challenge Landscape

Robotics teams are fighting **controllability**, **validation**, and **integration**, not just data acquisition, for real-time agentic workflows.

TRAINING DATA

45%
Data quality & consistency

45%
Controlling what agents collect/avoiding unintended capture

38%
Cost + provenance + legal uncertainty

SCALING AI SYSTEMS

45%
Integration/ orchestration complexity

42%
Reliable up-to-date external data

41%
Infrastructure/ ops cost

LIVE REASONING

52%
Data quality/ validation

45%
Integration complexity

38%
Reliability of live sources and compute cost

Regulation & Blocking Friction

The Compliance Paradox

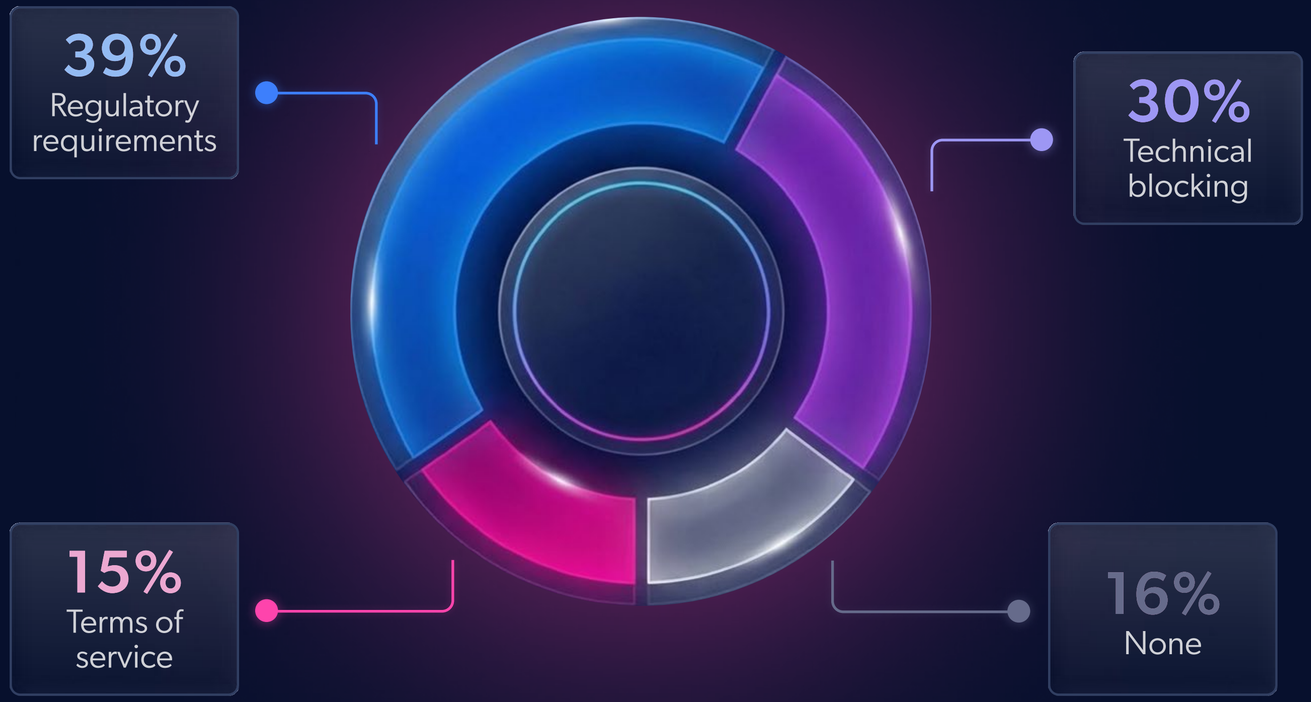
AI demands more web data, while regulation and blocking increase. This is causing great friction for innovation, as AI leaders balance their needs with overcoming challenges and making ethical decisions.

Extent of Limitations

90%

agree restrictions are limitations on innovation

Which Restriction Creates the Greatest Friction?



More Challenges Ahead

An overwhelming majority (88%) of respondents agree that access to public web data is becoming increasingly restricted through gatekeeping measures. Here are their predictions of what will happen in the near future.

Regulatory Legislation

Website Blocking

Ethics and Compliance are Non Negotiables

They also present additional challenges within their web data infrastructure and collection process.

Most essential for ethical and compliant access

Ethics And Compliance Are Non Negotiables

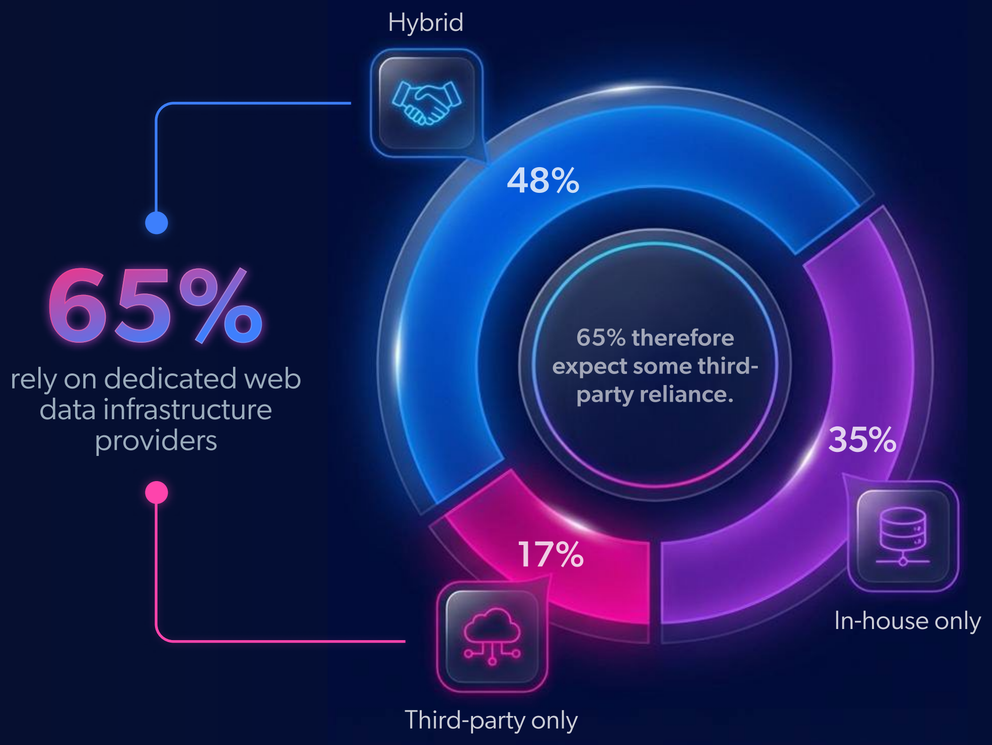
They also present additional challenges within their web data infrastructure and collection process.

Main ethical risks in acquiring data

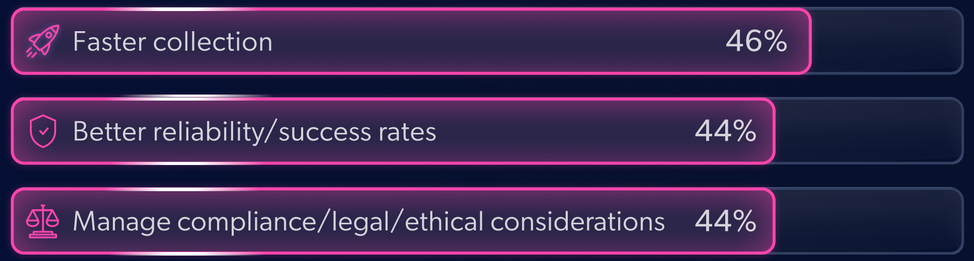
Web Data Infrastructure Providers Are Strategic Partners

In a landscape where rules vary by region, AI practitioners rely on dedicated web data infrastructure providers to collect data for them to stay compliant and consistent with website changes.

Approach Over Next 12 Months



Reasons to Use Third Parties



Ethical/Compliant Essentials

“There is a fundamental architectural shift in how AI systems are built and operated. 97% of organizations now connect their AI directly to live web data sources, which represents the exponential growth of the foundational data infrastructure layer. The era of static training datasets is over.

Whether you're building search, agents, predictive models, or physical automation, the common denominator is access to reliable, real-time public web data. Organizations are scaling despite friction because they have no choice and most rely on dedicated web data infrastructure providers to manage those complexities.

The winners in this space will be those who deliver speed, reliability, and compliance simultaneously. That trifecta is what defines the permanent infrastructure layer for AI."

Or Lenchner, CEO, Bright Data

bright data

www.brightdata.com

FOLLOW US

AI SUMMARY

CONTACT US