Data for AI 2025

Web Data Infrastructure for AI:

The foundation of AI and powering its future.

bright data

Methodology

Bright Data commissioned Vanson Bourne to conduct an independent survey of 500 US and UK senior level decision makers working in analytics, IT and technology departments, business direction and strategy, or R&D.

For respondents to qualify they must develop or operate real-time AI agents, foundational models, or prediction models and use web data.

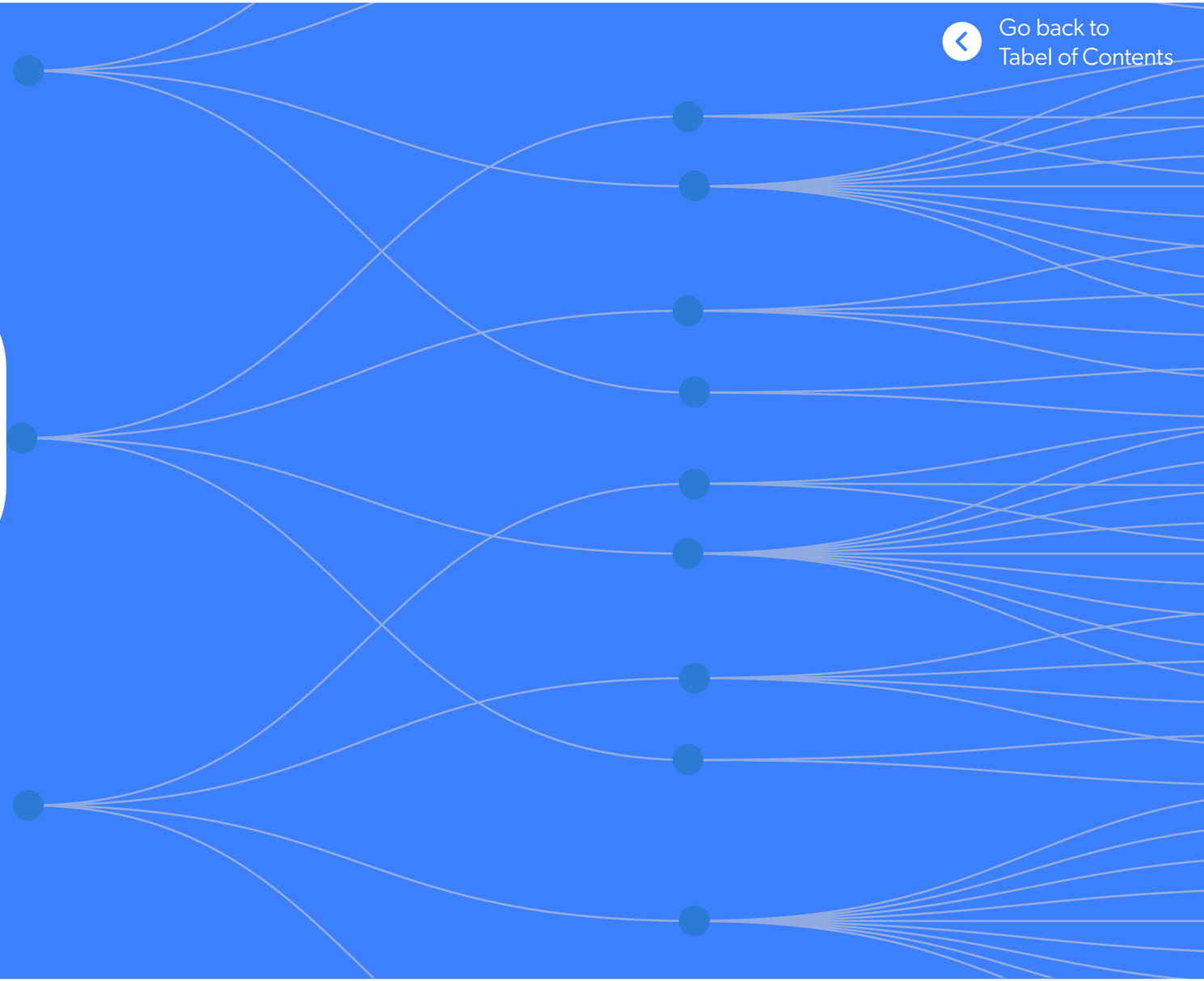
The purpose of the research is to identify and understand organizations data needs and challenges in the context of data for AI. This report highlights common barriers and pain points in data requirements.

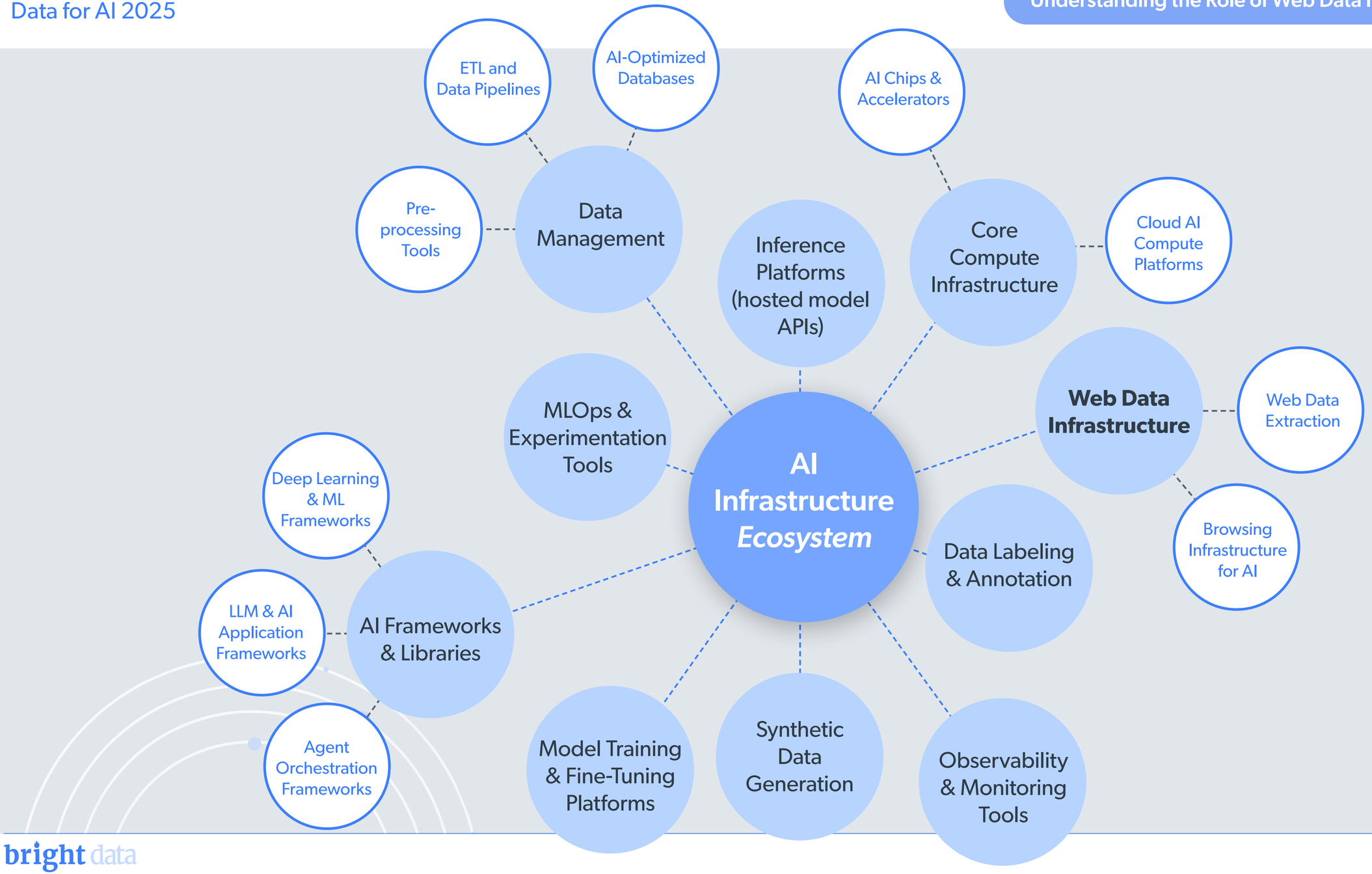
bright data

Table of Contents

Go back to
Tabel of Contents

Understanding the Role of Web Data Infrastructure for AI




AI Infrastructure Ecosystem
Data Management
ETL and Data Pipelines
AI-Optimized Databases
Pre-processing Tools
Inference Platforms (hosted model APIs)
Core Compute Infrastructure
AI Chips & Accelerators
Cloud AI Compute Platforms
Web Data Infrastructure
Web Data Extraction
Browsing Infrastructure for AI
Data Labeling & Annotation
Observability & Monitoring Tools
Synthetic Data Generation
Model Training & Fine-Tuning Platforms
MLOps & Experimentation Tools
AI Frameworks & Libraries
Deep Learning & ML Frameworks
LLM & AI Application Frameworks
Agent Orchestration Frameworks

Supporting the Entire AI Lifecycle

Training | **Inference**

Training Data

Pre-training, Tuning, Knowledge base

- Clean, structured data
- Discovery via web archive
- Annotated datasets

Data Feeds

Real-time insight, Grounding & Continuous pre-training

- Live data extraction
- Industry-specific pipelines
- Stealth data retrieval

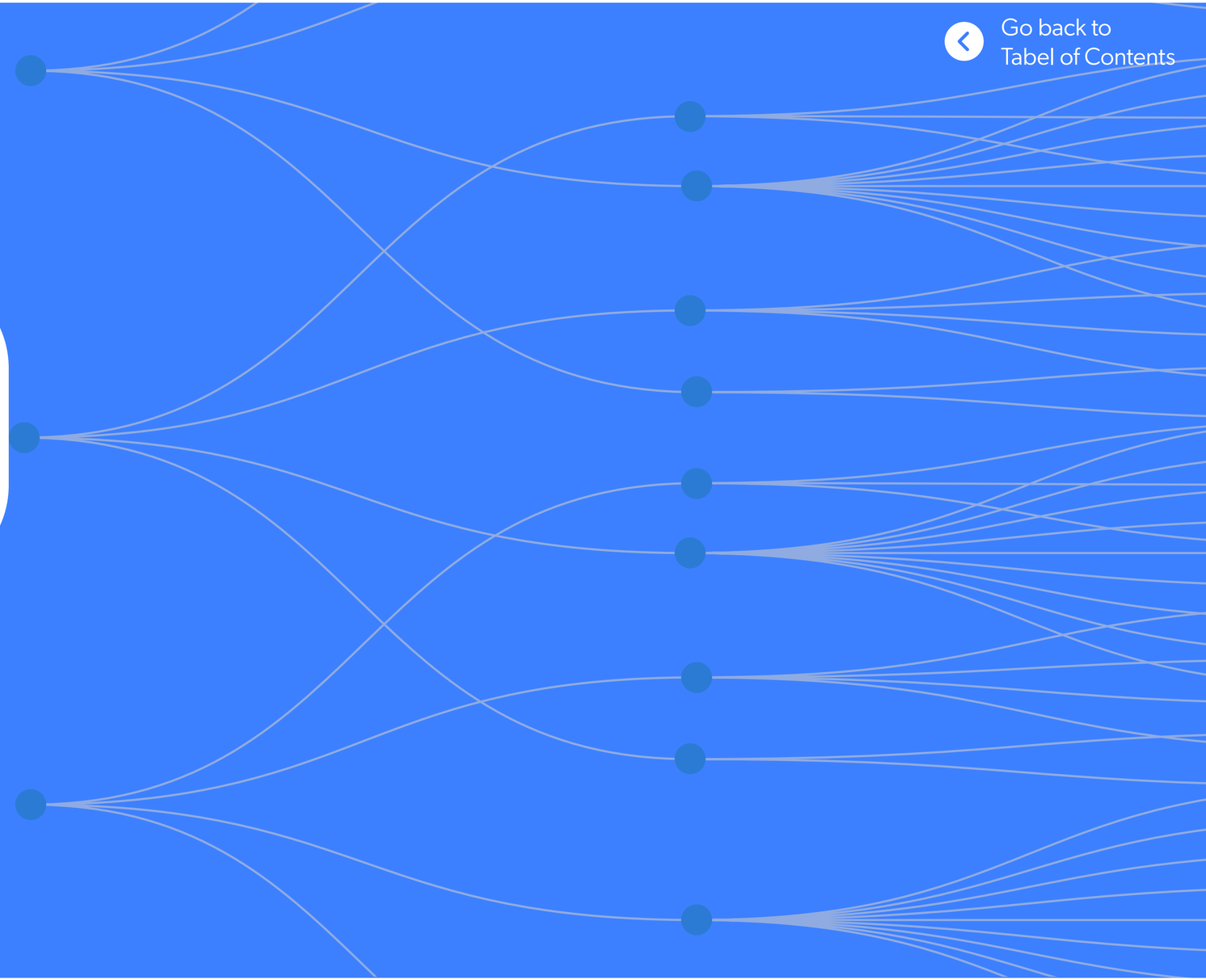
Web Access

Search, Navigation & Information extraction

- Browser infrastructure
- Crawling and extraction
- Search & answer engines

Go back to
Tabel of Contents

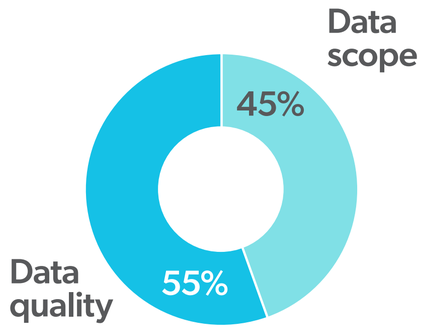
Summary of Business Segments



Startups

83 respondents

Primary differentiation for AI performance



52% are seeing positive financial impact/ ROI from web scraping

69% use public web data as a main source for collecting real-time, connected data

- Public web data's main role in AI strategy:
  Scaling AI capabilities with automated web data ingestion
- #1 challenge to collecting real-time, connected data:
  Data security and compliance
- #1 reason to work with data partner:
  Speed of data collection

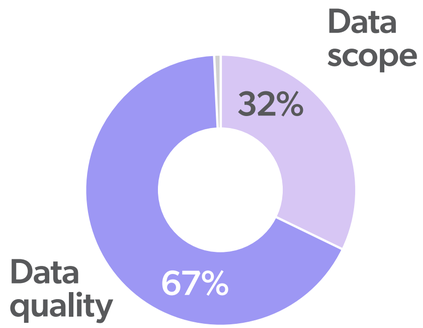
Types of public web data most critical for GenAI training

- Video **76%**
- Audio **70%**
- Text **72%**
- Language **67%**
- Image **71%**

SMBs

133 respondents

Primary differentiation for AI performance



43% are seeing positive financial impact/ ROI from web scraping

64% use public web data as a main source for collecting real-time, connected data

> Public web data's main role in AI strategy:
Improving AI model accuracy and relevance

> #1 challenge to collecting real-time, connected data:
Data quality

> #1 reason to work with data partner:
Cost efficiency of data collection

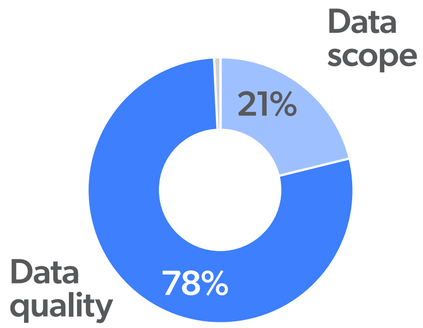
Types of public web data most critical for GenAI training

- Video **61%**
- Audio **57%**
- Text **82%**
- Language **77%**
- Image **74%**

Enterprises

284 respondents

Primary differentiation for AI performance



54% are seeing positive financial impact/ ROI from web scraping

69% use public web data as a main source for collecting real-time, connected data

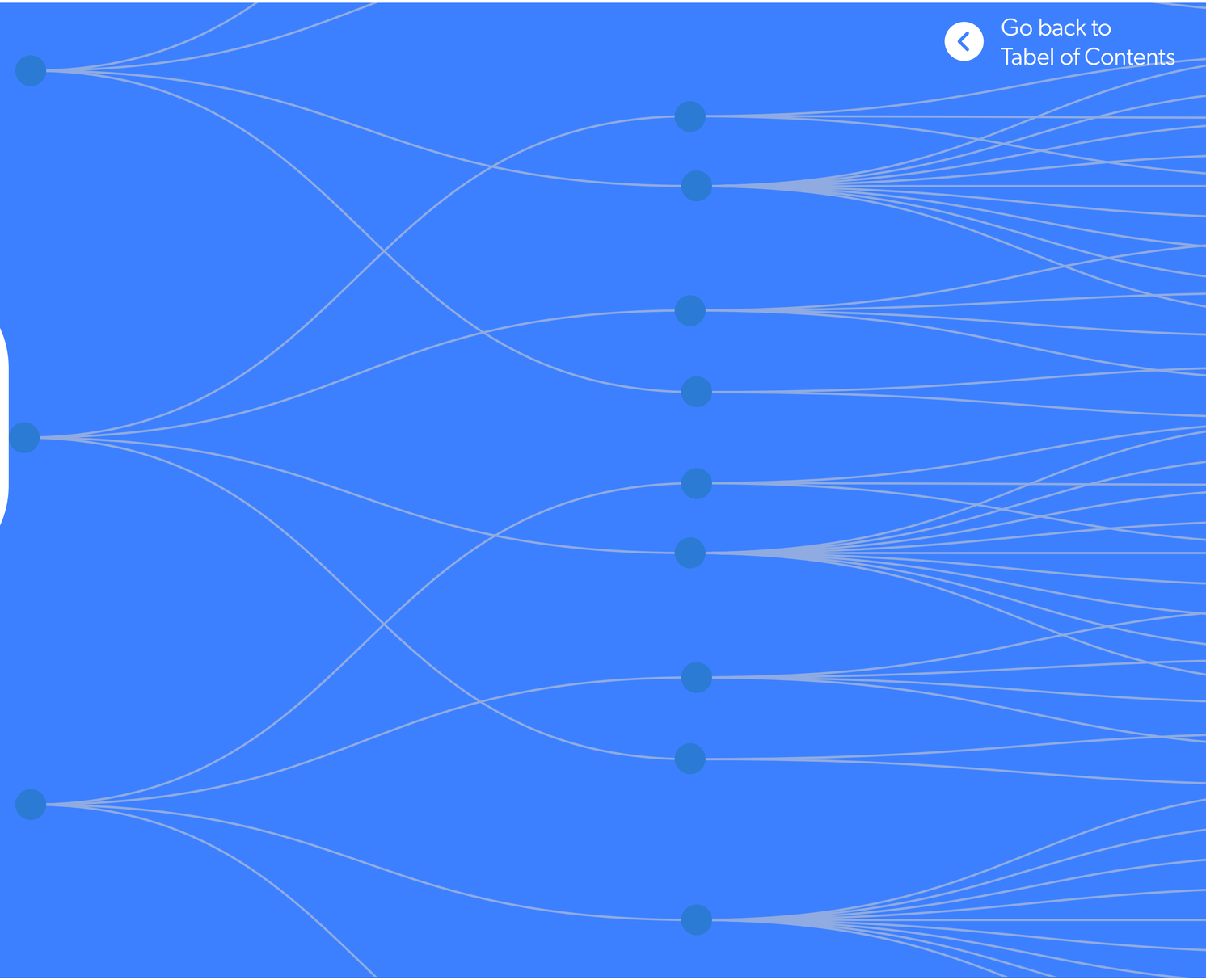
- Public web data's main role in AI strategy:
  Improving AI model accuracy and relevance
- #1 challenge to collecting real-time, connected data:
  Data security and compliance and data quality
- #1 reason to work with data partner:
  Completeness of data collection

Types of public web data most critical for GenAI training

- Video **58%**
- Audio **52%**
- Text **85%**
- Language **70%**
- Image **64%**

Go back to
Tabel of Contents

AI's Main Driver



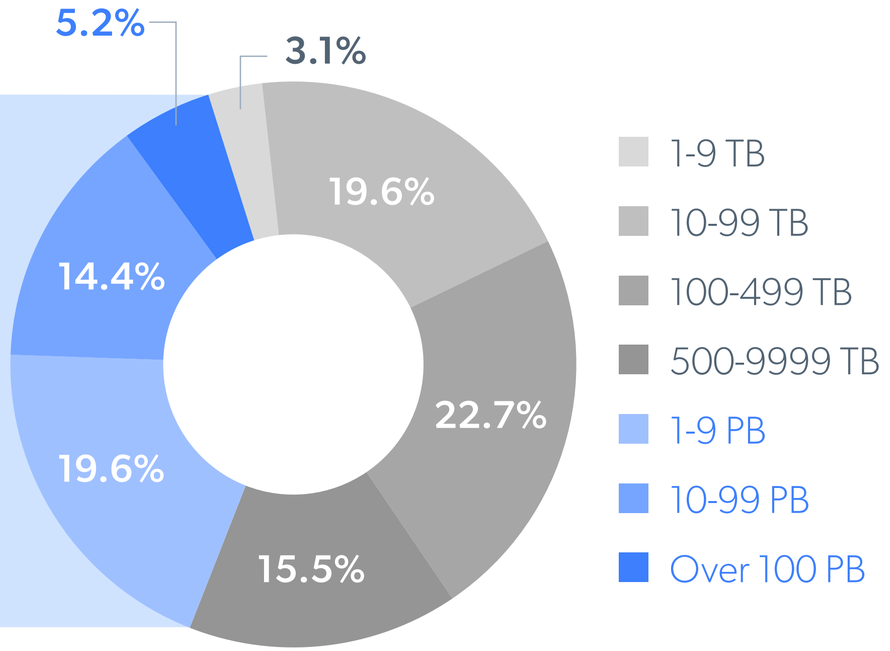
Data Volume Growth

38% of companies consume over 1PB of public web data annually.

Data needs are expected to grow by **33%** in the next year, as companies find more specific needs for images, audio, or text to improve outputs. Budgets for data acquisition will increase by 85%, reflecting the rising importance of web data in AI strategies.

Bright Data is scaling to meet this demand with over 7PB already stored.

The amount of public web data respondents say their company consumes annually

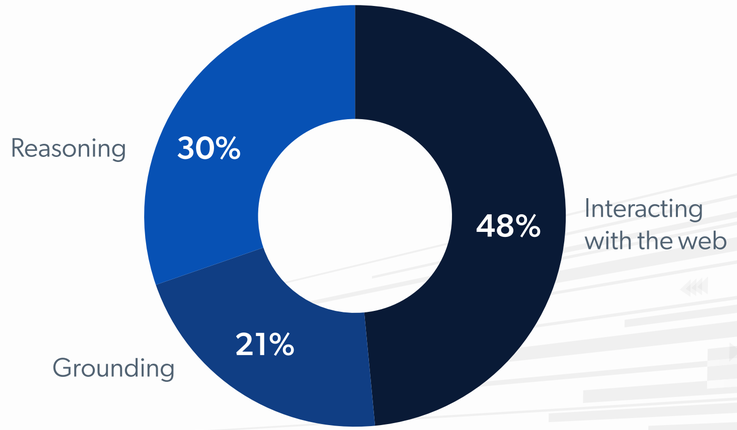


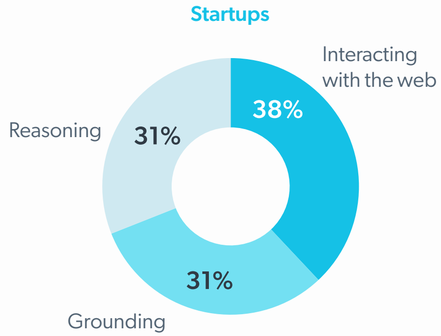
Real-time Data Use Cases

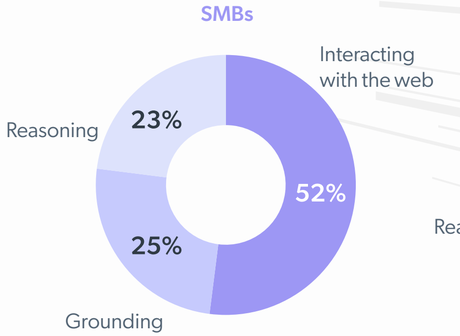
Use cases like interacting with the web, reasoning, and grounding require dynamic, up-to-date data that is collected without interruptions. Bright Data's unblockable infrastructure enables these capabilities at scale.

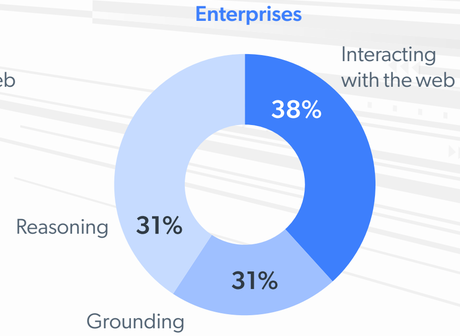
96% of organizations collect real-time web data for inference, but most face challenges doing so.

Organizations are mainly using real-time data during inference for:







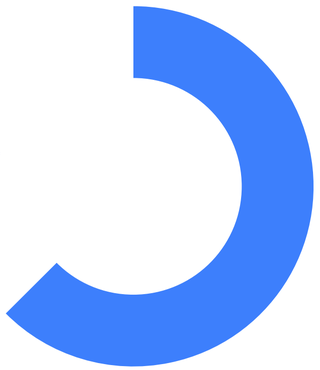


Data Types

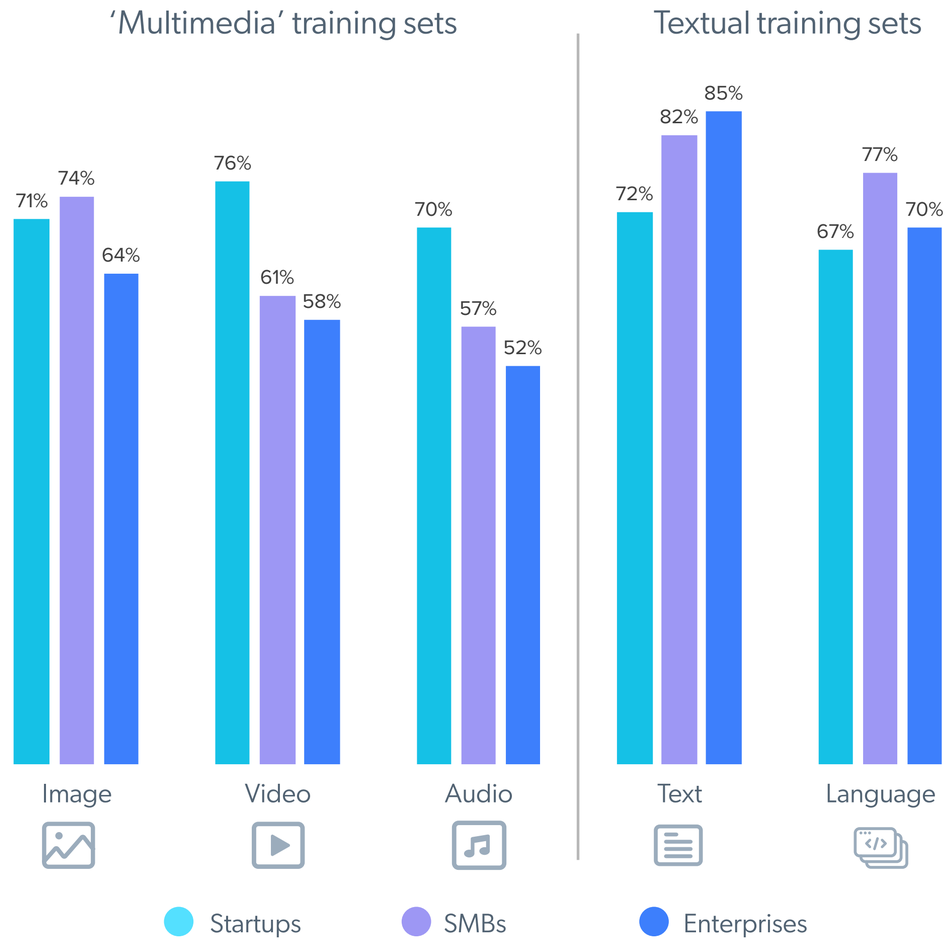
Organizations rely on both multimedia and textual data, with startups leaning toward video and image, and enterprises prioritizing text and language. Only 27% of businesses collect all five data types.

Bright Data supports diverse data collection, with **92%** of organizations stating vendor partnerships improve data variety.

65% of organizations use public web data as the primary source for AI training.

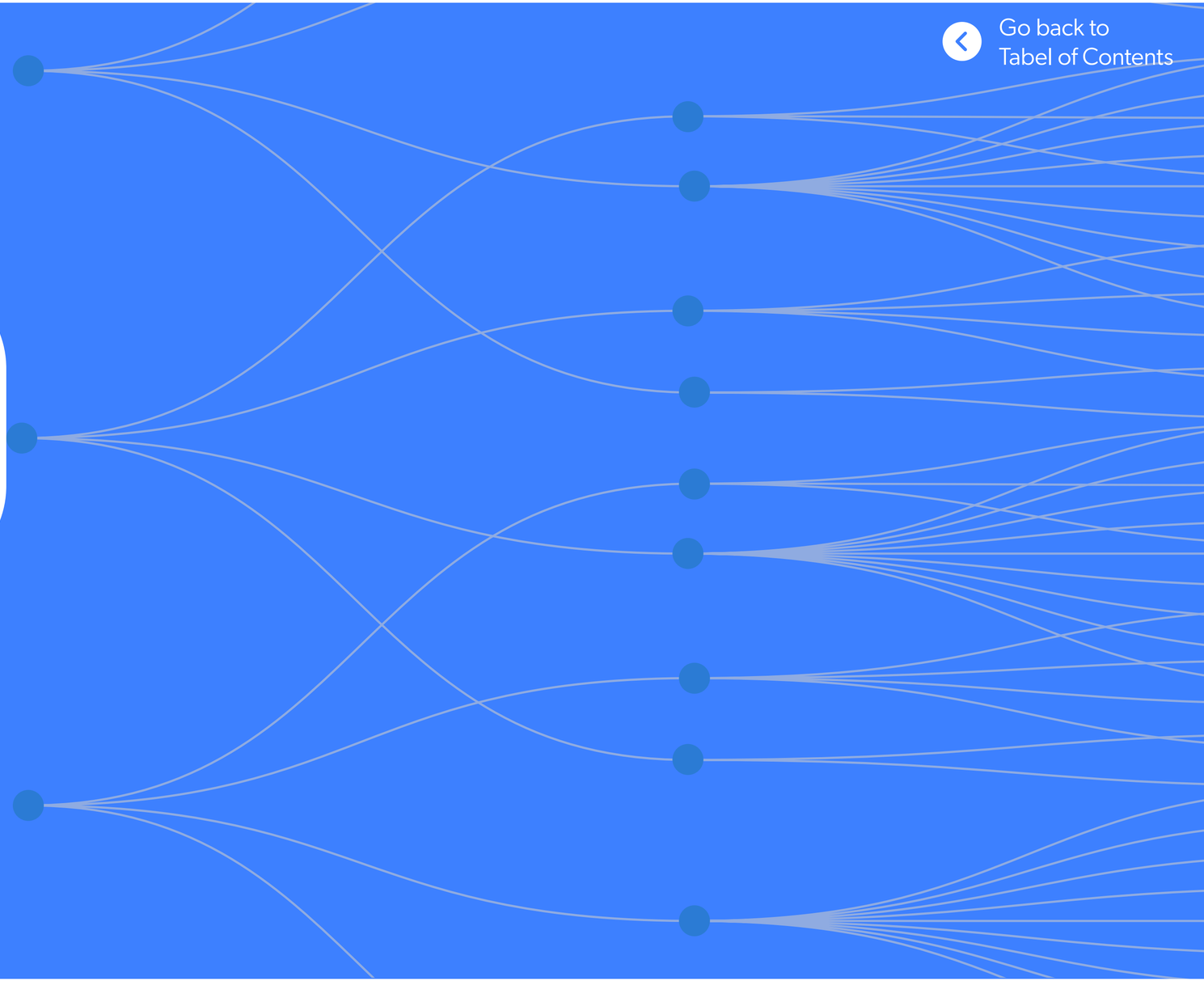


Types of data collected for AI



Go back to
Tabel of Contents

Competitive Edge

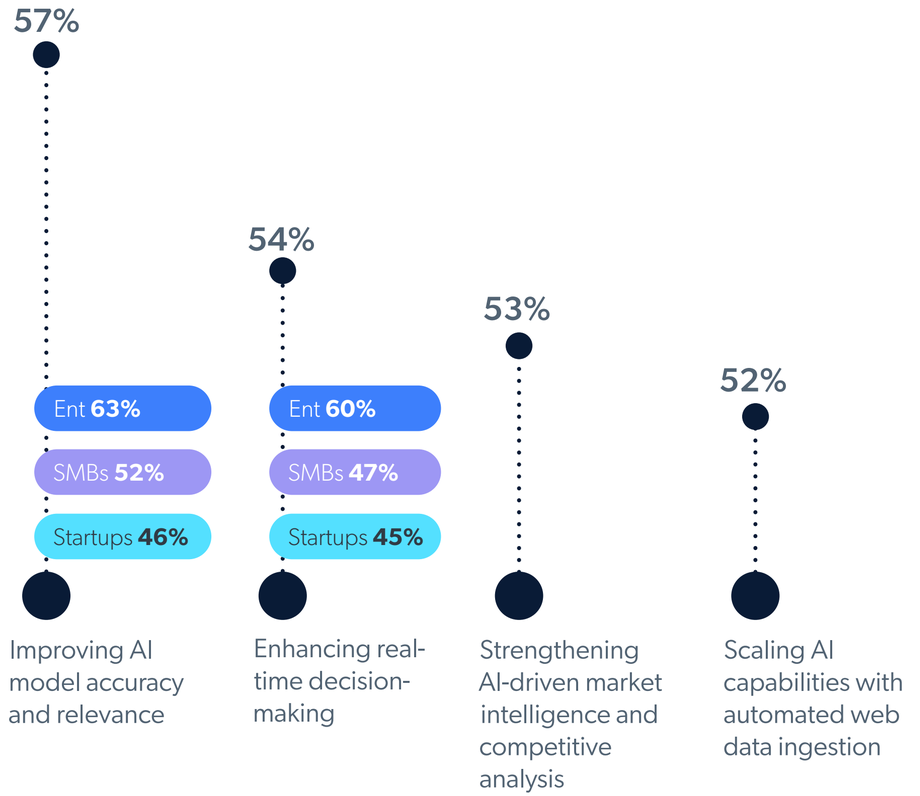


Real-Time Data's Role

Models require more than just public web data for training; they need additional data to improve and fine-tune performance. As consumers and businesses demand more, additional data is necessary for contextualization with real-time web access for timely, context-aware decisions.

92% agree real-time, dynamic data is critical to maximizing their AI model performance.

Public web data acquisition's role in competitive AI strategies

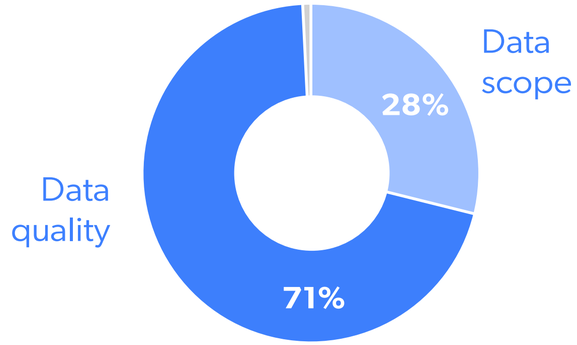


Future Differentiators

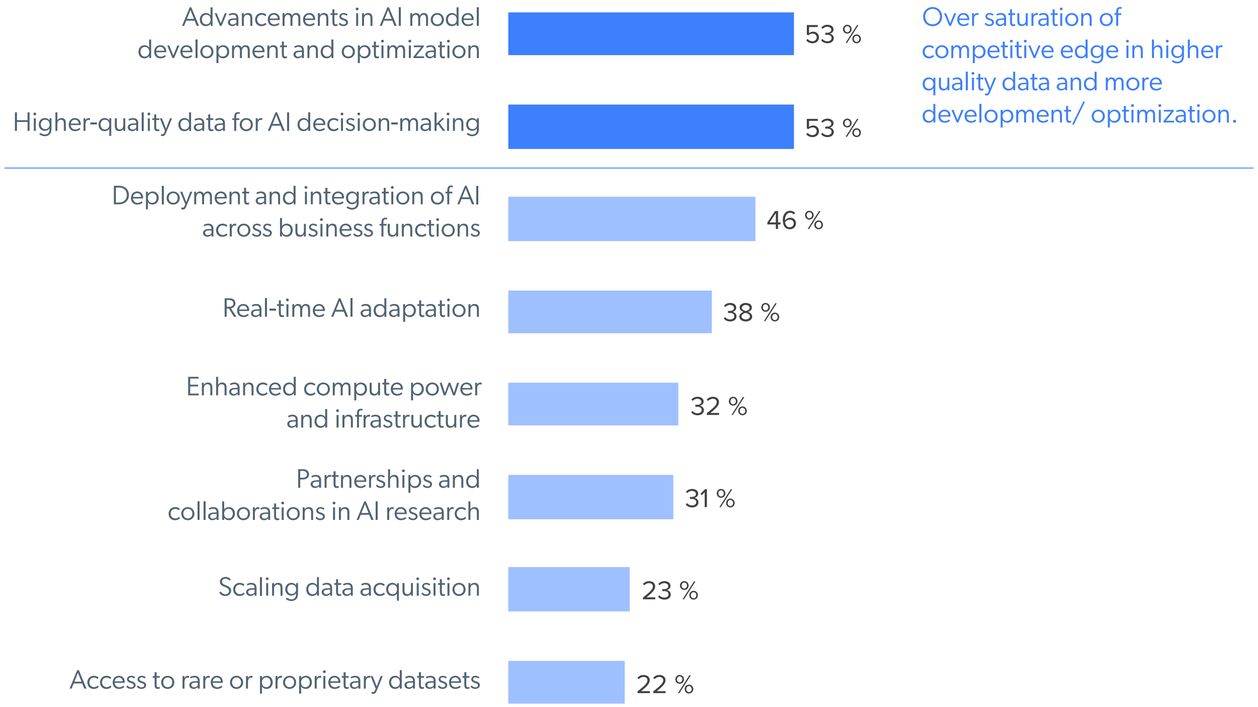
AI is no longer just a short-term research strategy; it's about gaining real-time insights into competitors. This requires extracting data in the right format and seamlessly integrating it into AI applications.

Data quality will be the competitive differentiator

The primary differentiator for organization's AI performance over the next 2 years



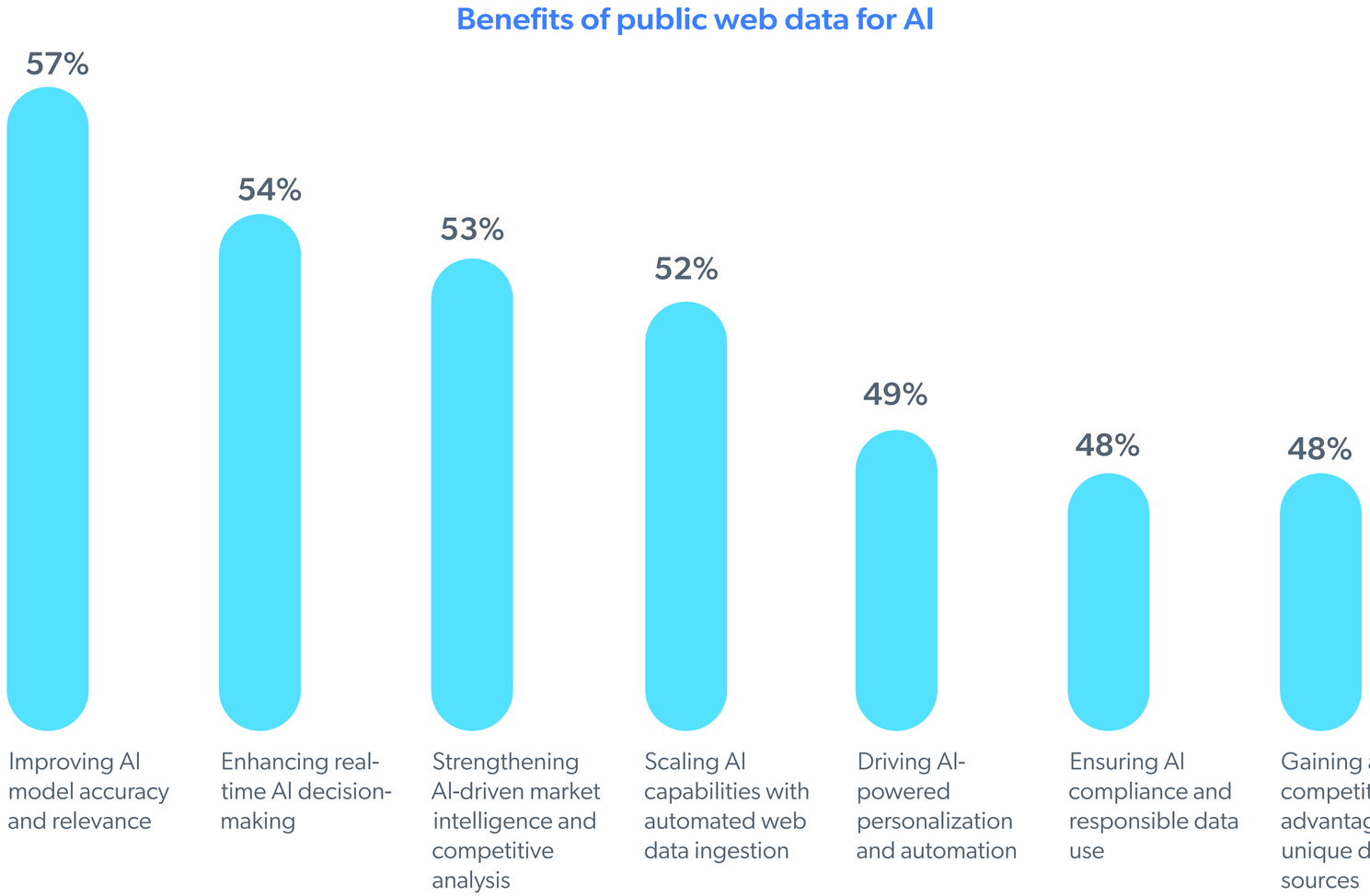
In the next two years, where do organizations anticipate their competitive edge in AI?



Fueling Competitive AI Strategy

Organizations are collecting real-time web data and powering their AI agents- a move towards building responsive context aware AI systems. The data is balancing historical accuracy and real-world responsiveness for optimal AI performance.

89% of the respondents say there are at least three definitive benefits to acquiring public web data.



Cost vs. Performance

On average companies report spending **41%** of their AI budget to data acquisition. The amount spent on public web data collection and acquisition is expected to more than double in the next 12 months to **89%.**

43% prefer a balance between cost optimization and model performance.

32% prefer cost optimization.

Mainly cost optimization, but some considiration of model performance.

25% prefer model peformance.

Mainly model performance, but some considiration for cost optimization.

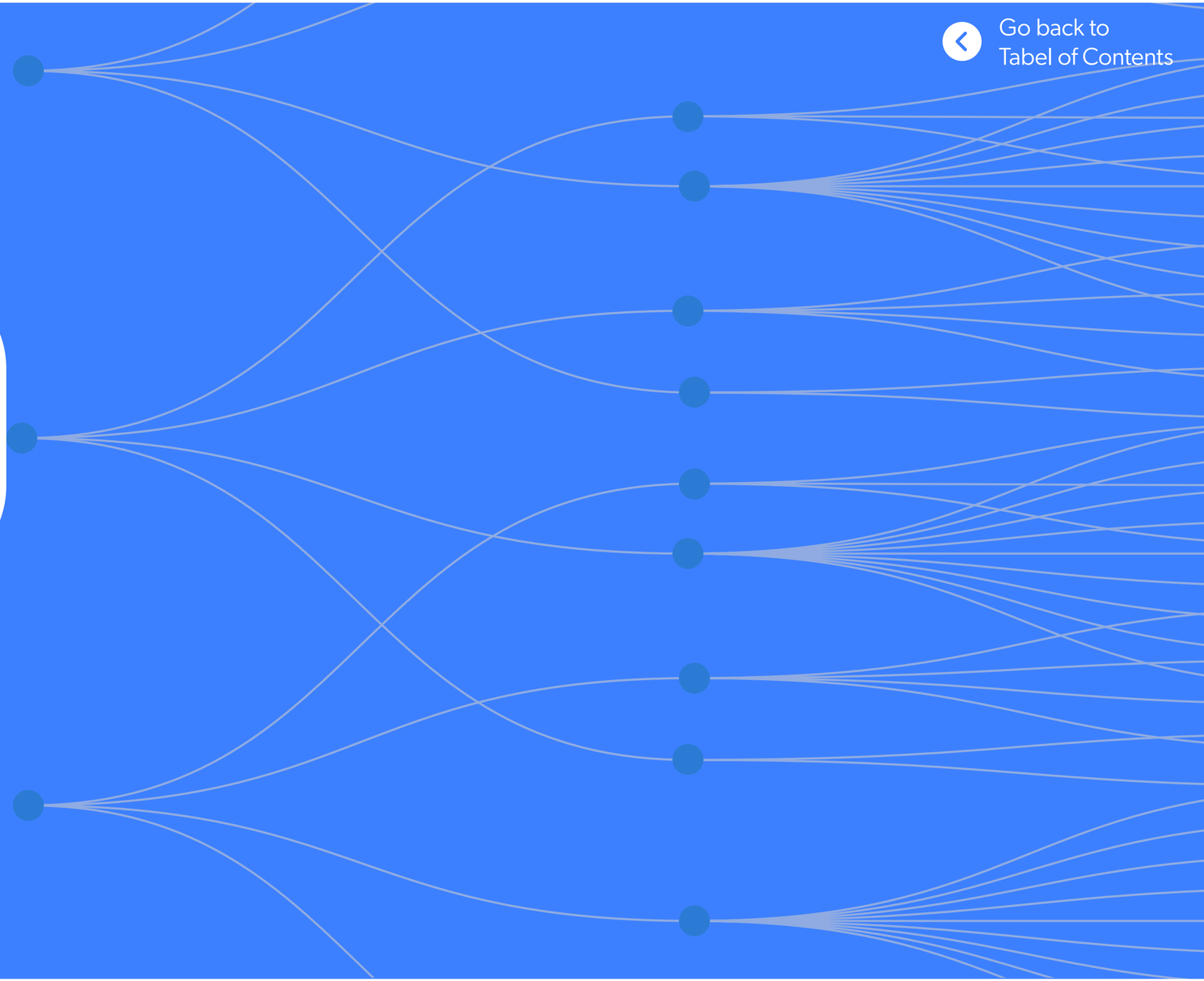
42% Startups
prioritizes a balance.

55% SMBs
prioritizes a cost optimization.

49% Enterprises
prioritizes a balance.

Go back to
Tabel of Contents

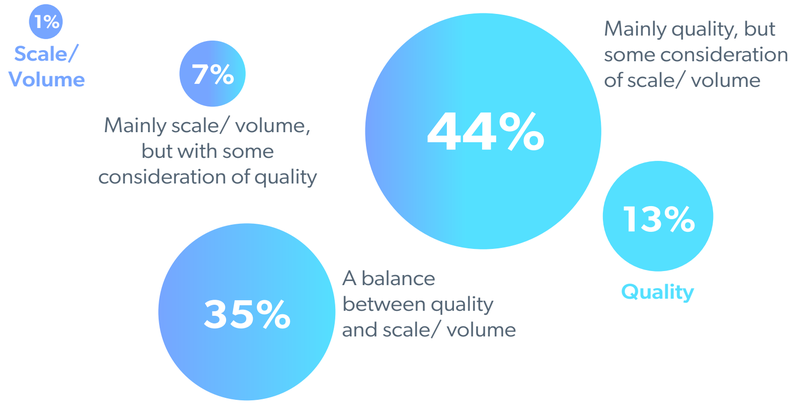
Overcoming Challenges



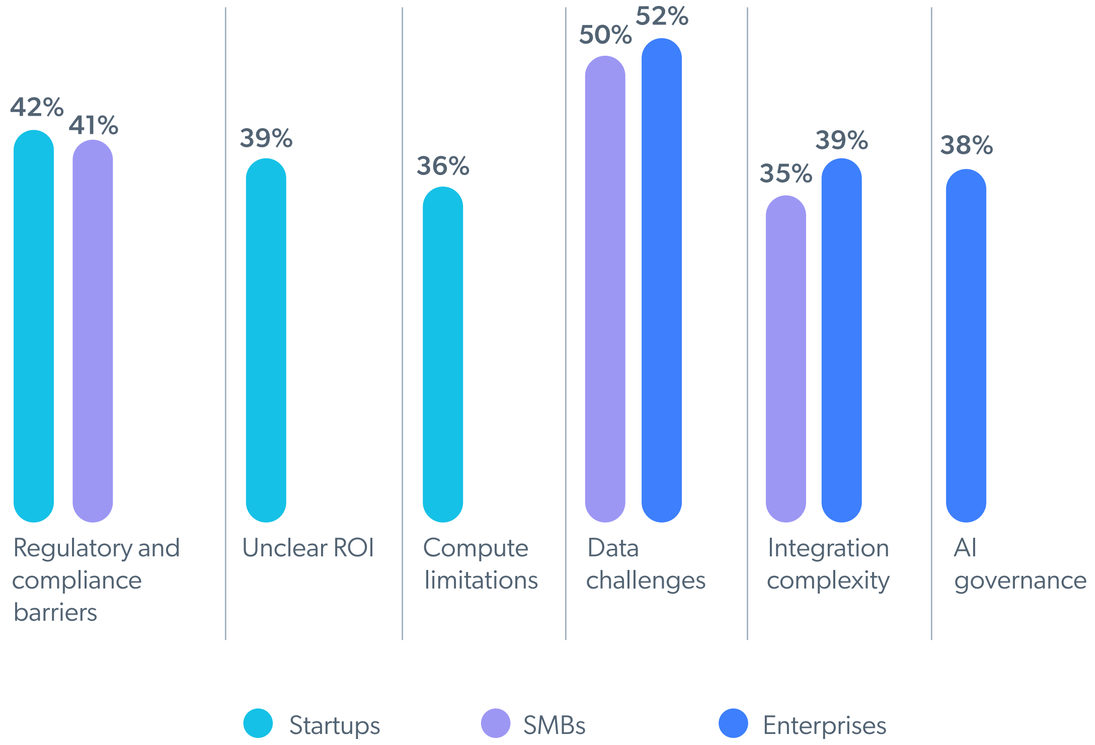
Future Differentiators

98% of organizations face challenges scaling data acquisition for AI. Without reliable, accurate, and timely data, even the most advanced AI models risk becoming irrelevant.

Is data quality or scale/volume more important to organization's models or agents?



Challenges organization face in scaling data acquisition for AI applications



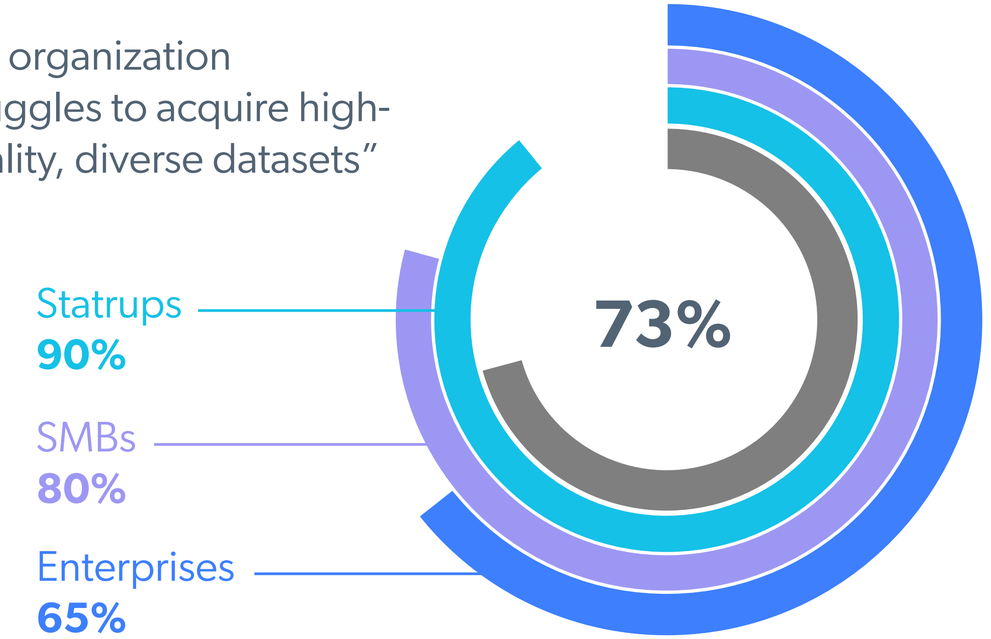
Better Data = Better Predictions

An example of an AI project involving public web data

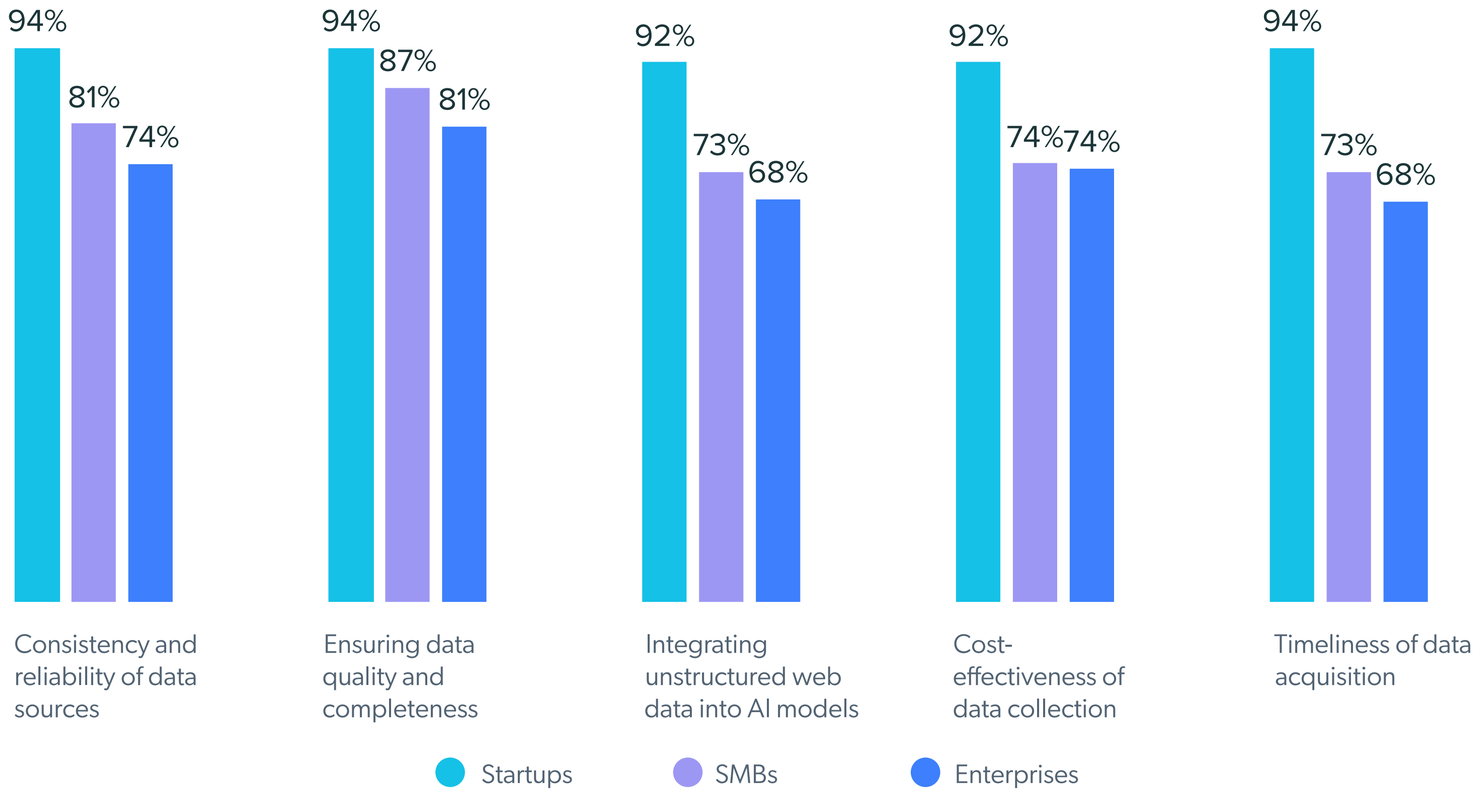
*"We used **publicly available real estate data** to build a predictive model for property valuations. This significantly improved our pricing accuracy, leading to faster sales and increased profit margins. **Basically, better data meant better predictions**".*

Enterprise organization in manufacturing

"My organization struggles to acquire high-quality, diverse datasets"



Percentage of organizations who find collecting, cleaning and processing public web data for AI models challenging for the following…



A trusted partner eliminates the need to build complex data pipelines. This allows AI teams to focus on innovation accelerating time to market and improving model performance. A data partner ensures compliance with evolving regulations, enhances data diversity, and improves cost-efficiency. For AI to succeed, especially in inference and agent-based systems, partnering with a specialized web data provider is no longer optional, it's strategic.

What is the main reason that would drive organizations to work with a data partner for their AI projects?

Startups

65%
Speed of data collection

61%
Completeness of data collection/ability to collect data in one batch

58%
Accessing web data for training sets

SMBs

70%
Cost efficiency of data collection

69%
Completeness of data collection

59%
Speed of data collection

Enterprises

70%
Completeness of data collection

68%
Cost efficiency of data collection

67%
Speed of data collection

Go back to
Tabel of Contents

Message from the CEO

AI has transformed the world, and in business, it's either fueling growth or leaving non-adopters behind. Companies that harness the open web are advancing rapidly and driving greater innovation. 99% of survey respondents say public web data is a key part of their company's competitive AI strategy. This report highlights the common challenges organizations face in deploying AI agents or building AI models, but also reveals the strategies that set leaders apart. Public web data is the foundation of AI innovation, essential for training and powering models at inference. For AI providers, a robust web data infrastructure is critical. Without rapid innovation to overcome the challenges machines face in accessing and processing this data, many AI breakthroughs will fall short, and we risk never realizing AI's full potential.

Or Lenchner,
*CEO **Bright Data***

bright data

www.brightdata.com

Contact us

Follow us