Enterprise Strategy Group™ by TechTarget

WHITE PAPER

# The Essential Role of Data and Data Quality in IT-related AI Model Training

By Gabe Knuth, Senior Analyst
Enterprise Strategy Group

February 2024

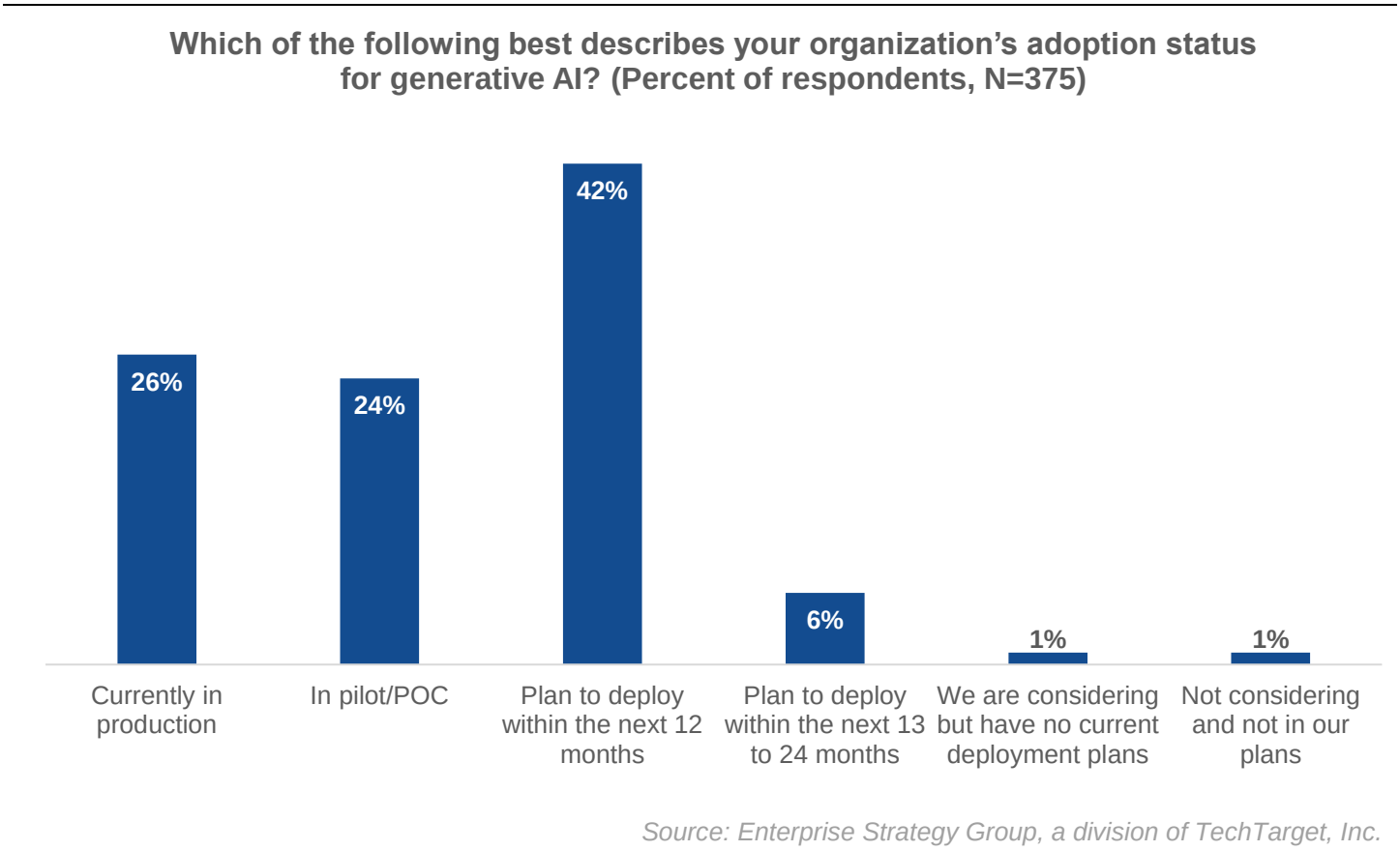
# Contents

# Overview – The Problem

The explosive growth in generative AI (GenAI) has it positioned among the most transformative trends in the history of computing, and the expected impact it will have on business operations, including IT, is hard to overstate. This is understandable, since the increased efficiency and insights gained are sure to have an effect on productivity, not to mention employee—and customer—experience.

It's no wonder, then, that 92% of organizations surveyed in a recent study by TechTarget's Enterprise Strategy Group noted that they have already adopted or intend to adopt a GenAI-based solution in the next 12 months (see Figure 1). Still, given all the attention it has received, generative AI was a top priority for only 9% of organizations—well behind cybersecurity, digital transformation, and cost cutting and edging out automation as well as app modernization (see Figure 2).[1]

**Figure 1.** The Overwhelming Majority of Organizations Expect to Deploy an AI-based Solution in the Next 12 Months




*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

[1] Source: Enterprise Strategy Group Complete Survey Results, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

**Figure 2.** While Not a Top Priority by Itself, Generative AI Is Poised to Affect All the Initiatives That Rank Above It


*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

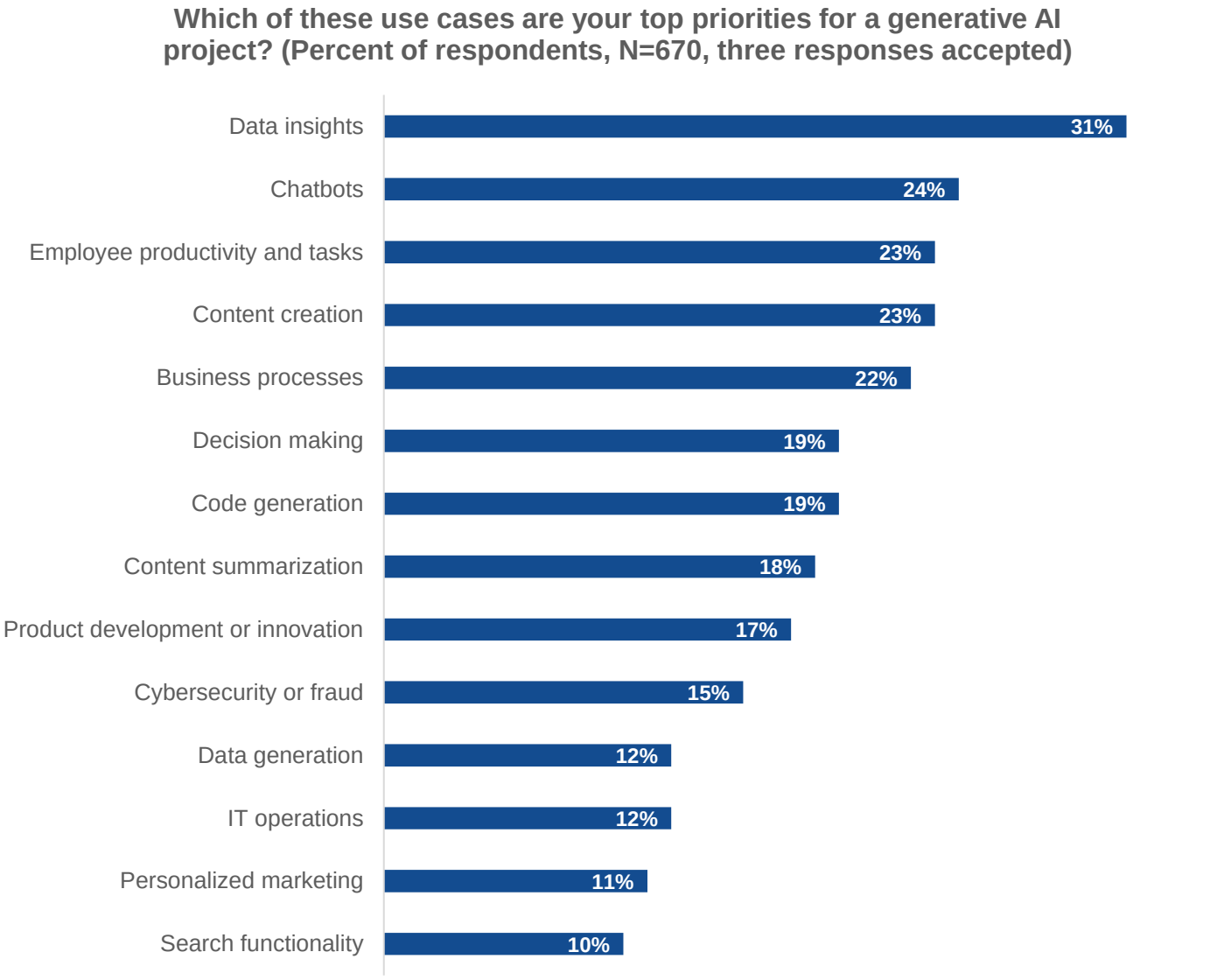
What's lost in this data point, however, is that GenAI is more than just a single organizational priority. While just 9% of organizations are prioritizing GenAI (which is still higher than cloud and sustainability initiatives), GenAI is poised to affect every single priority on that list, if it hasn't already. Below are some examples of this impact:

- **Cybersecurity.** GenAI is being used for automated threat detection and response as well as attack simulation.
- **Digital transformation.** GenAI is being used to automate and optimize business processes, as well as enhance customer experience.
- **Cost cutting and automation.** GenAI helps streamline operations, reducing manual labor and optimizing resource allocation using predictive analytics.
- **App modernization.** GenAI improves app functionality and user experience using analytics and personalization features, as well as automating updates and maintenance.

GenAI's potential impact becomes even more apparent when looking at the current top priorities organizations have for AI projects (see Figure 3).[2]

[2] Ibid.

**Figure 3.** High Expectations for Generative AI Begin With Data Insights but Traverse All Areas of the Business




*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

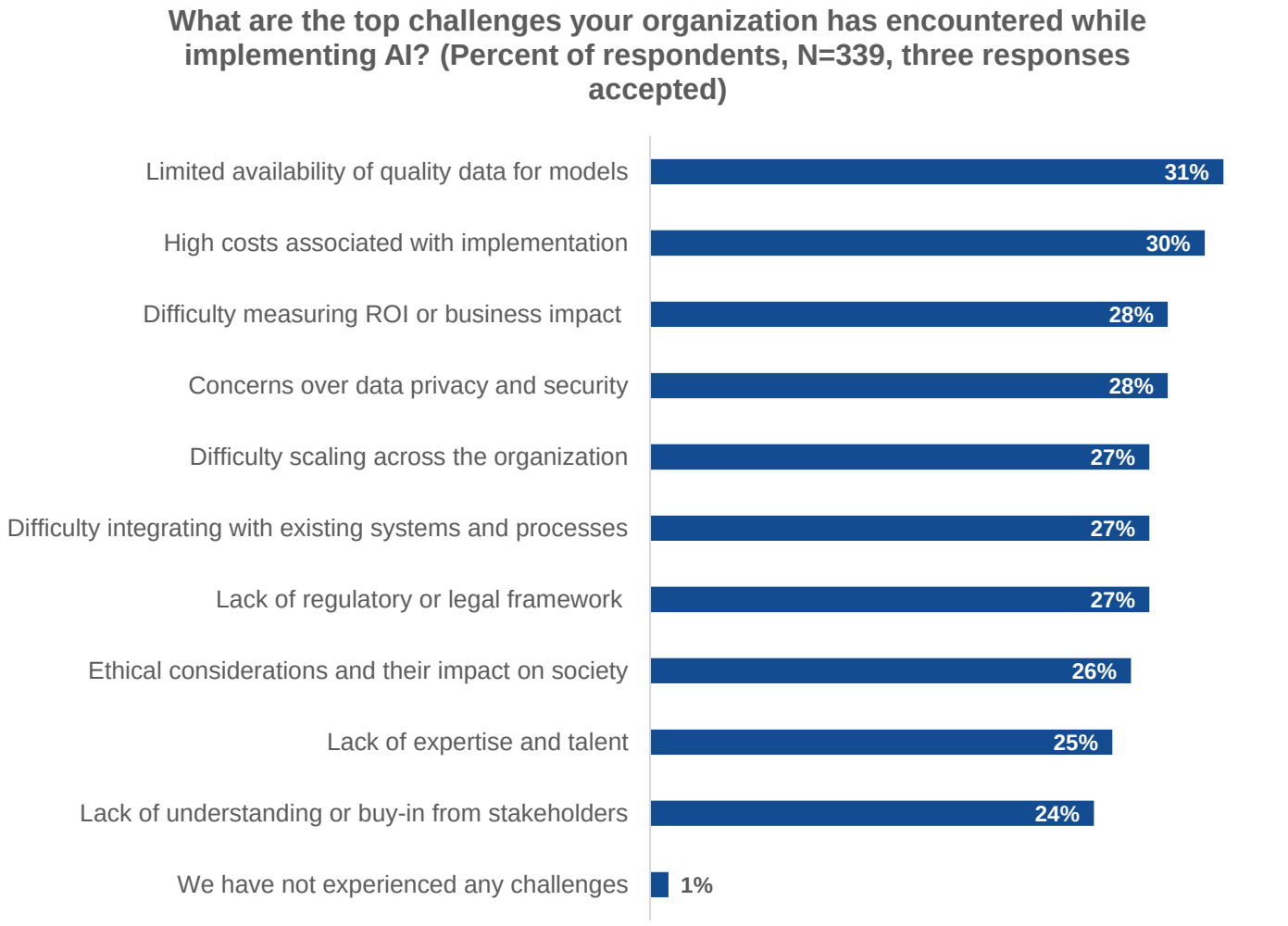
These projects are targeting multiple areas within organizations, from customer-facing things like chatbots and marketing to internal use cases like product development and code generation, and it's easy to see how closely they relate to IT. Plus, responses like "employee productivity" and "business processes" align with key initiatives of digital transformation, cost cutting, and automation.

The top priority—with good reason, as it's so important to the overall success of any AI endeavor—is data insights. Data is food for AI, and what's true for humans is also true for AI: You are what you eat. Or, in this case: The better the data, the better the AI.

# Data Is the Key to Success…But It Must Be Good Data

With so much riding on AI, it's no surprise that the list of challenges organizations face with regard to their AI implementations is long. While certain challenges, like security, privacy, and cost, are to be expected, the No. 1 response—and rightfully so—is the limited availability of quality data to feed into AI models (see Figure 4).[3]

**Figure 4.** The Need for Quality Data Tops a Long List of Challenges




*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

On the surface, IT might believe it has more than enough data to feed into AI models. After all, recent Enterprise Strategy Group research indicates that two-thirds (68%) of organizations have deployed more than 10 tools for security and management alone, not to mention tools for other things like performance analysis.[4] Data quantity should not be a problem—or so the thinking goes. But what good is a mountain of data if it's of poor quality?

While there are high hopes that AI itself can work out the kinks and determine good data from bad, this line of thought suffers from one critical flaw: AI doesn't know good data from bad. It just knows data, and whatever data is

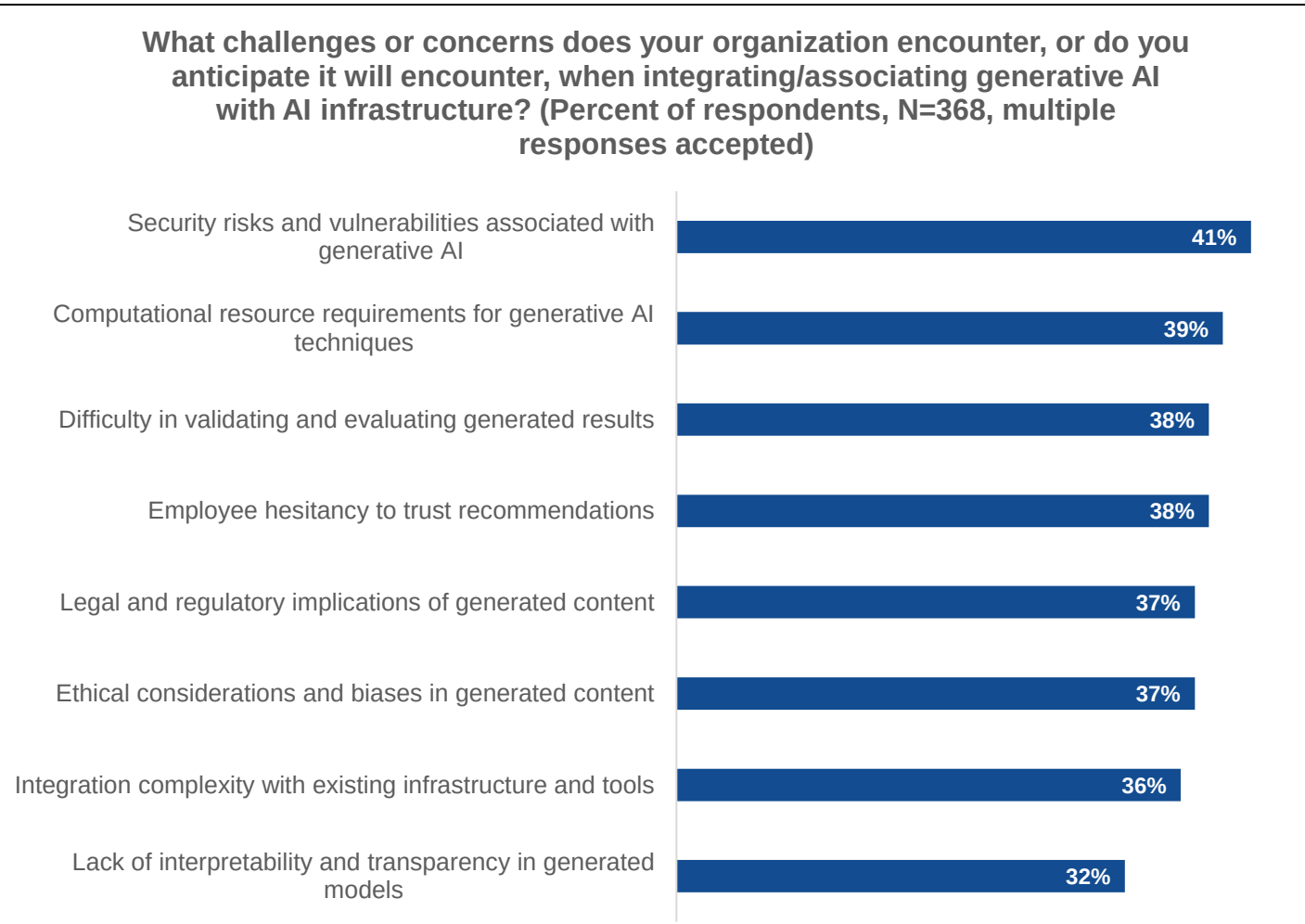
[3] Source: Enterprise Strategy Group Complete Survey Results, *Navigating the Evolving AI Infrastructure Landscape*, December 2023.
[4] Source: Enterprise Strategy Group Research Report, *Managing the Endpoint Vulnerability Gap*, May 2023.

being used to train a model will ultimately *become* the model. Paradoxically, if a model exists someday that actually can sort out the good from bad data, it will only be able to do this because it was trained to know what *good* data is, and the only way to do this is *with* good data.

## Why Is Good Data So Important?

Up to this point, the conversation has revolved around how organizations are getting their AI projects off the ground, what they hope to accomplish, and what the expected challenges are. The next step is the actual implementation of AI and its integration into the business. When asked what challenges will need to be overcome when integrating AI, the responses, though varied, indicate one core theme that is critical to success with AI: trust (see Figure 5).[5]

**Figure 5.** Trust Is a Common Thread Among Reported AI Challenges or Concerns




*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

Without good data to feed into the AI, trust can never be achieved. Without trust, full adoption can't happen. And without full adoption, the overall goals can't be achieved.

[5] Source: Enterprise Strategy Group Complete Survey Results, *Navigating the Evolving AI Infrastructure Landscape*, December 2023.

Plus, the risk of using poor data can negatively affect each of the areas listed above, and the ramifications can often go beyond that of simple trust. For example:

- With bad data, security risks and vulnerabilities could go undetected (false negative) or over-reported (false positive). Plus, bad actors are also using AI, which could amount to a data quality arms race in the context of security.
- Computational resources for training and fine-tuning models are expensive. Feeding in extraneous bad data not only affects the quality of the model but also costs more money.
- Legal and regulatory concerns, balanced by ethical considerations and biases, present legitimate liabilities for organizations and demand quality, accurate data to be useful.

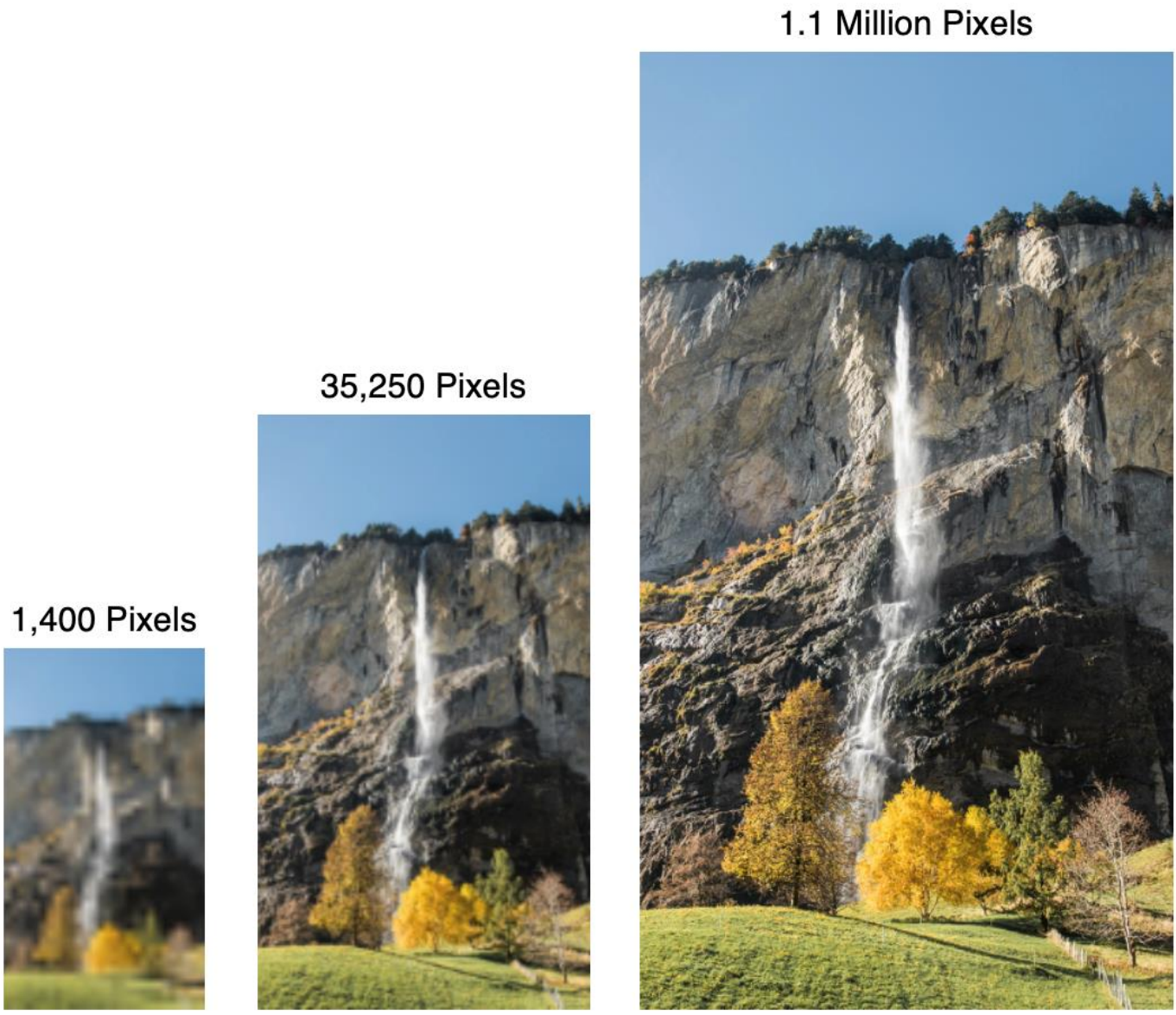
## Trustworthy Data Is More Than Just Good Data

To get trustworthy data—and therefore a trustworthy AI model that doesn't become a liability to achieving overall objectives (or to the company at large)—the depth, breadth, history, and quality of the data all matter.

- **Depth.** Accurate, domain-specific data is more useful than generic data for AI models. For example, a chatbot trained for end-user support on a generic foundation of data will be much less effective than one that's been trained on the apps, processes, and performance data of the end user.
- **Breadth.** Working hand in hand with depth, breadth is also critical to trustworthy data. A model that's been trained across a wide range of quality, contextual data is important for building trust and maximizing value. In the chatbot example, the more systems, people, processes, apps, procedures, and so on that it knows, the more helpful it can be without having to escalate to a person.
- **History.** Historical data provides additional context and a wealth of data on which to train a model. This helps identify performance trends, areas that can be optimized, and much more that is critical to contextual awareness and trust-building.
- **Quality.** Finally, none of this matters unless the data is high quality, well structured, and free of noise. As mentioned above, with so many tools deployed for management and security, there are often overlapping data, blind spots, and conflicting data from systems that aren't integrated with each other. This can lead to quality issues—not to mention duplicative and/or useless data that increases the cost of training without providing any real value—that affect the overall trustworthiness of the data.

There's a lot that goes into building a data set full of good information that can train a trustworthy model. In general, more data leads to more accuracy, as long as the source data is good. Think of an image at various resolutions. At low resolution, the image is blurry and hard to make out because there just aren't enough pixels to form a complete picture. But as you add pixels, the details start to emerge (see Figure 6).

**Figure 6.** Image Clarity Depends on the Number of Pixels, Just as AI Model Quality Depends on the Number of Quality Data Points

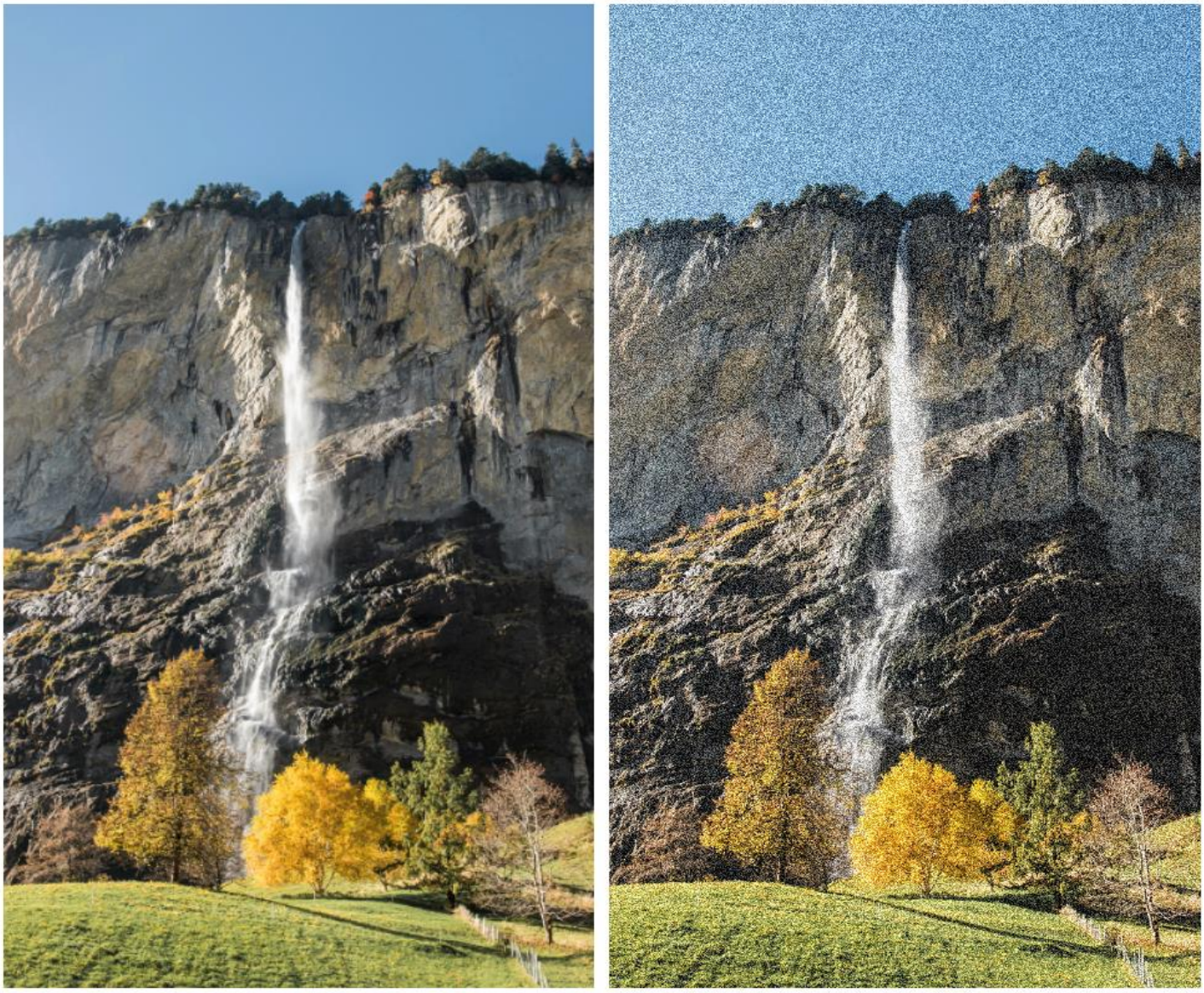



*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

Data used to train AI models is similar. The more data it's trained on, the more accurate it can become. If data points are the pixels of the AI model, collecting as much relevant data as possible helps to build a more accurate, trustworthy model.

That said, collecting more data is only good if the data quality is good. Just as adding noise to an image affects its overall quality, regardless of resolution, using noisy, irrelevant data to train an AI model results in a poor-quality model, regardless of how many data points are used (see Figure 7).

**Figure 7.** It's Not Enough to Collect Lots of Data: If the Data Is Noisy, the Model Quality Suffers



*Source: Enterprise Strategy Group, a division of TechTarget, Inc.*

While the depth, breadth, history, and quality of data all matter, the risk of simply adding more data to hopefully achieve these goals is that bad data seeps in—a scenario that's potentially even worse than not having enough data. Plus, pulling data from multiple sources can lead to confusion and conflicts because there's no single source of truth.

## Ensuring Data Quality With Human Expertise

By now, it should be clear that data sets used to train AI should be based on high-quality, contextual data, but there's one other element that can help ensure a trustworthy result: human expertise. When training a model, human expertise is useful to provide context about the data that's being ingested. Think of it as coaching someone

to learn a sport, as opposed to throwing a person onto a field with a bag of balls and hoping they figure it out. Plus, as usage of the model grows and trust improves, more contextual data can be collected that drives further improvement in a loop that grows exponentially.

Beware, though: That exponential growth is going to happen no matter what, and just as investments made in data quality today will have a positive effect on data quality and trust, the opposite is undoubtedly true if poor quality data is used.

# Lakeside Software's Role in a High-quality IT Data Foundation

To overcome the challenges that AI introduces, it's important to consider all the data sources that are used to build models. While organizational goals will rely on data from multiple sources, many of the objectives outlined above—such as enhanced analytics and insights, improved operational efficiency, and improved employee engagement and satisfaction—rely on quality, IT-driven data.

Ultimately, this means that if you want the AI model to speak the language of IT, you need to feed it massive amounts of quality IT data, and that's where Lakeside comes in.

Using the Lakeside SysTrack platform, organizations can collect massive amounts of trustworthy, relevant data from their endpoints, virtual desktops, and virtual applications, giving IT a complete visibility of endpoint performance and IT health across the digital workplace. This data can be used to fuel an AI that has been trained on IT data, ultimately resulting in a model that is not only IT-aware but also attuned to an individual customer's specific usage patterns and needs.

A model trained on this kind of data can be useful for end users, IT, and the business at large. Imagine a natural language query asking, "How many of the engineering team's laptop batteries are under warranty?" and getting a real-time, accurate result. A proactive request like this can lead to reduced downtime and fewer calls to the help desk, and that's just one small example of the time and cost savings that deep, contextual awareness can lead to.

The best part? None of this is new to Lakeside. For more than 25 years, they've been known not only for the amount and quality of the data they collect but also for how they put that data to use. Their capabilities include:

- **Structured data.** Lakeside SysTrack derives understanding and meaningful intelligence from its well-structured data so its embedded AI models deliver relevant, contextual data insights. Data is organized, labeled, and formatted so it is conducive to analysis and interpretation for IT teams to "shift left."
- **Performance analytics.** Extensive, historical data about hardware, OS, and application performance is correlated so that IT has multiple views into an issue, accelerating root cause analysis while getting users and IT personnel back to maximum productivity.
- **Capacity planning.** SysTrack uses performance analytics and historical data to aid in capacity planning for VDI and DaaS environments, helping IT choose the right VM sizes and configurations to optimize cost and performance for both on-premises and cloud use cases.
- **Natural language queries.** As IT has evolved, so has Lakeside, which is pioneering the use of natural language queries that make it easier to extract insights from the data collected for IT.
- **Machine learning.** One of Lakeside's core competencies, machine learning, has been used to build many of the features in the platform over the entire course of Lakeside's existence.
- **Digital employee experience.** Lakeside is a driving force behind digital employee experience (DEX) and end user experience management (EUEM), which aim to increase end-user productivity and satisfaction through performance analytics, workflow optimization, and automated remediations.

# Conclusion

AI occupies an increasing part of enterprise mindshare, and its impacts will be widespread throughout each company that adopts it. Successful implementations will be predicated on many things, including security and compliance, but it's hard to find a more important, fundamental component to AI success than data quality.

This applies to all areas of the business, but it's especially true for IT. Without good data—and lots of it—AI models themselves simply cannot be good. Bad models lead to bad outcomes, wasted money, loss of trust, loss of competitive positioning, and so much more—and that's not even considering security and compliance issues!

Every organization that is on an AI journey needs to be thinking about data today. Even if there are no concrete plans for how to use AI, having good data when the time comes will help accelerate adoption and build trust among stakeholders, end users, and even customers. Investments made in trustworthy data today will undoubtedly pay off in the future, and organizations that are struggling with quality IT data would be wise to look at Lakeside to help them establish a solid, trusted data foundation.

## About Lakeside

Lakeside Software is how organizations with large, complex IT environments can finally get visibility across their entire digital estate and see how to do more with less. For far too long, IT teams have struggled to see what's going on in their dark estate—where costly inefficiencies, poor employee experiences, and unresolved problems hide. Only Lakeside lets you give everyone a better view, so they can see the hidden issues, see the smartest fixes, and see the biggest savings. That's why so many of the world's leading global brands rely on Lakeside. And it's how Lakeside's customers see an average ROI of more than 250%. Learn how you can get a better view at www.lakesidesoftware.com.

Lakeside®
Give everyone a **BETTER VIEW.**™

**About Enterprise Strategy Group**
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com