Flying Cloud®

Ready to *really* protect your data

AI’s Return to Reality

It’s Still All About the Data.

TABLE OF CONTENTS

Good AI Never Comes from Bad Data

For five years now, we've been talking about the importance of knowing what's in your data. In our last white paper, *AI: Magic E-Wizard Machine or Maniacal Hell-bot?* we discussed what AI is and isn't. We covered critical data questions, like where AI tools like ChatGPT get their data (see the section *Where'd This Sh** Come From?*). In the section *Non compos mentis*, we described how bad data causes AI to lose its mind and why. And in *What's in the Water?* readers learned about the challenges organizations face when they have oceans of data but no idea what's actually in the water. You can get that paper here.

We're Gonna Say it Again

Since then, we watched the AI hype cycle take off. We've seen glowing predictions and stats, but we haven't really believed them. As the inevitable failures emerge and AI sentiment starts to slide into the slough of despond, we're going to say it again—AI results are always determined by the data fed to the model.

In this paper, you'll gain the context for what's failing and why. We also offer a data-first strategy for succeeding with AI projects. You'll see how a data integrity layer is the foundation of your model and ultimately, business value. It's going to be OK. **You can do this.**

AI—Not a Cheat Code to Business Success

Most companies' 10K reports in the past two years claimed that they have an AI strategy and implied that they are well down the road to success. We don't wish failure on anyone, but it was clear to us that 99% of all AI projects would fail, and of the 1% that succeeded, only a tiny fraction will actually deliver acceptable results.

Today, a growing number of reports are urging caution, scaling back, or even scrapping projects:

- This year, 65% of chief financial officers (CFOs) reported insufficient return on investment as a significant barrier to AI deployment in their companies.[1]
- At least 30% of GenAI projects will be abandoned after proof of concept by the end of 2025, due to poor data quality, inadequate risk controls, escalating costs, or unclear business value.[2]
- The number of Fortune 500 companies citing AI as a risk in their annual financial reports surged 473.5%, an increase from 2022.[3]
- The share of companies abandoning most of their AI initiatives jumped to 42%, up from 17% last year. The average organization scrapped 46% of AI proof-of-concepts before they reached production.[4]
- Over 80% of AI projects will fail—twice the failure rate for non-AI technology-related startup companies.[5]
- Two-thirds of businesses say they are stuck in generative AI pilot phases and unable to transition into production and 97% grapple with showing generative AI's business value. Nearly three in five respondents admitted to facing pressure to move projects along faster.[6]

The fatal flaw—poor data quality and governance

AI projects fail for multiple reasons, but the fatal flaw underlying most of them is poor data quality and governance. For example, the lack of a well-defined business objective is cited as a major reason for failure. A poorly defined problem results in feeding unknown, inappropriate, or unclear data to the model, which simply gets you to failure faster.

[1] How Many Companies use AI? https://gptzero.me/news/how-many-companies-use-ai/

[2] Gartner, https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025

[3] The Rise of Generative AI in SEC Filings, https://arize.com/wp-content/uploads/2024/07/The-Rise-of-Generative-AI-In-SEC-Filings-Arize-AI-Report-2024.pdf

[4] AI Project failure rates are on the rise: report, https://www.constructiondive.com/news/AI-project-fail-data-SPGlobal/742934/

[5] The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed, https://www.rand.org/pubs/research_reports/RRA2680-1.html

[6] Stuck in the pilot phase, https://www.te.com/en/about-te/news-center/reports/industrial-technology-index/2025-report-state-of-innovation-industrial-tech.html

What Were They Thinking?

How did this happen? At the end of 2022 the world was coming out of a pandemic and AI hadn't yet made main-street headlines. So we found it incredible that in less than two years, over 80% of companies are reported to be using AI. When you read the fine print however, only about 40% are using it and another 40% are "exploring" it. You'll read that 94% of business leaders believe AI is critical to success, 99% of the Fortune 500 companies use it, and nine out of 10 organizations view AI as "crucial," "critical," or "essential" to gaining a competitive advantage.

We have so many questions.

What does "AI use" even mean?

Turns out that the definition of "AI use" is a moving target. In 2017, the definition for AI use was using AI in a core part of the organization's business or at scale. In 2018–2019, the definition shrank to embedding at least one AI capability in business processes or products. Since 2020, the definition became even less ambitious, meaning that an organization has adopted AI in at least 1 function.[7] It *might* be possible that "everyone" is using it.

What is AI being used for?

If everyone is using AI, what are they using it for? Search again and sources say that companies are using it to enhance the customer experience, through tools such as chatbots. Apparently, AI is also used in IT, cybersecurity, and fraud management. TE Connectivity[8] identifies specific use cases in industrial sectors:

Top AI use cases by industry

73% Wireless 5G
Network monitoring and fault detection

74% Energy
Forecasting demand

68% Auto/Commercial Transportation
Vehicle development and design

71% Industrial Manufacturing
Predictive maintenance systems

85% Data, Cloud Computing, and AI
Automated data cleaning

[7] McKinsey & Company, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[8] 2025 Industrial Technology Index, https://www.te.com/en/about-te/news-center/reports/industrial-technology-index/2025-report-state-of-innovation-industrial-tech.html

Wait, what kind of AI are we talking about?

That looks fairly convincing. If these use cases and statistics are true, we have to ask how this happened so fast. The breathless media has focused almost exclusively on generative AI (GenAI), playing out the maniacal hell-bot personality. Is *that* what all of these organizations are using? Are we really monitoring mission-critical networks with hallucination-prone GenAI apps?

We hope not. It's more likely these applications are based on machine learning (ML)—a subset of AI that has been around for quite awhile. A number of ML algorithms are used to recognize data patterns, predict numerical values based on linear relationships, and group similar classifications of data. It's ideally suited—and has been long used—to tackle questions that require filtering and organizing large data volumes. Machine learning applications might account for the vast majority of "AI" applications being used by "everyone."

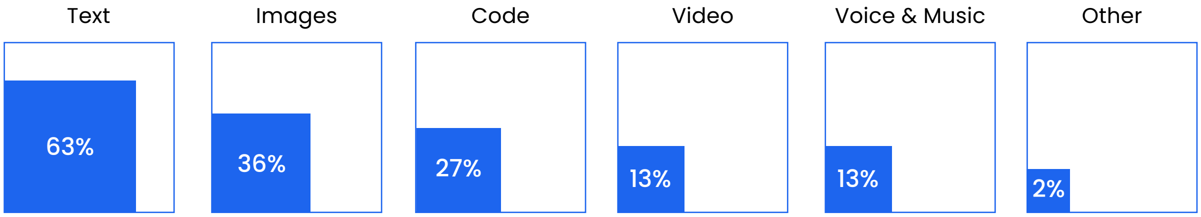
That's It?

Media hype also believed that AI's magic e-wizard machine personality would permanently change the world for good: "AI holds the potential to shift the way people access and use knowledge. The result will be more efficient and effective problem solving, enabling innovation that benefits everyone."

Well, packet networks did that. So did email. "Oh, but this is different." How? Apparently even ChatGPT can't come up with a convincing explanation, and if you've read *AI: Magic E-Wizard Machine or Maniacal Hell-bot?* you'll know why. Speaking of ChatGPT, it turns out that this shadow IT application that induces data security risk nightmares for CISOs is what constitutes the vast majority of GenAI use.

While text is the type of content that organizations are most commonly creating with gen AI, they are also experimenting with other modalities.

Types of content generated by gen AI at respondents' organizations,[1] % of respondents



[1] Only asked of respondents whose organizations regularly use gen AI in at least onee function. Figures were calculated after removing the respondents who said "don't know."

Source: McKinsey Global Survey on the state of AI, 1,491 participants at all levels of the organization, July 16-31, 2024

Congratulations. If you've used ChatGPT or a similar tool even once, your organization now has joined the ranks of AI-using companies. Wasn't that easy?

Then there are the data Karens

The point is there have been few concrete definitions or standards for what AI use is and isn't. Statistics lump small companies with global Fortune 100 companies, use cases with industries, and ML with GenAI. Once you start asking more pointed questions, you have a clearer picture of today's AI reality in business organizations. Don't fall for the guilt trips shaming you for not leaping head first, hair on fire into the GenAI fray:

- "Almost all companies invest in AI, but just 1 percent believe they are at maturity. Our research finds the biggest barrier to scaling is not employees—who are ready—but leaders, who are not steering fast enough."[9]
- "Businesses will either learn to take big leaps in AI adoption, setting the course of their businesses to welcome structural changes, or, struggle to catch up."[10]
- And this one, from a college professor with no P&L responsibility and certainly nothing to lose if you fail: "This is a time when you should be getting benefits [from AI] and hope that your competitors are just playing around and experimenting."

Frankly, we're appalled by these sentiments. The costs and risks of blindly adopting any new technology—especially AI—are high. According to Forbes, "While [AI] reshapes products and services, it also introduces unprecedented challenges: fairness, bias, transparency, and unintended societal impacts. These risks aren't hypothetical—they are real, urgent, and increasingly acknowledged in corporate decision making."

So that's what organizations were thinking. Today, mid-2025, it appears that perceptions of AI are returning to reality—it's *still* all about the data.

[9] McKinsey & Company, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

[10] 2025 AI Business Predictions, https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html

You Can’t Model or Prompt Your Way Out of Bad Data

AI models simply interpret the data fed to them. They can’t compensate for poor data quality. If you don’t have enough data to adequately train your model, you won't get your desired results. You can add synthetic data to make up deficits, but synthetic data without fresh data infusions degrades results.

Just *any* data won’t work either. Data must be examined, qualified, managed, and continuously curated to ensure that it’s suitable. According to a survey by Gartner, 63% of organizations either don’t have, or are unsure if they have, the right data management practices for AI. In fact, Gartner predicts that through 2026, organizations will abandon 60% of AI projects unsupported by AI-ready data.

But isn’t data just…data? What makes data AI-ready?

Getting to AI-ready data

AI-ready data meets defined quality, quantity, and management standards before it can be used in an AI model. The problem is that most organizations have not defined these standards for their data. They’ve assumed that data in their systems is fine the way it is. Until it becomes clear that it’s not.

Getting data ready for AI is going to require significant changes to current data management practices:

- Traditional data management practices are structured, rigid, and too slow for AI teams.
- Data is collected in a host of different formats and stored across diverse repositories and platforms, making it challenging to evaluate and prepare.
- Data usage is rarely documented.
- Most organizations don’t really know the actual data they have, how it’s used, or where it originates.

You should have established practices and competence with data warehousing, business intelligence, and traditional analytics before you can expect successful AI deployments. The infrastructure must be flexible to accommodate changing AI use cases. Over time, real-world data drifts and models must be re-trained with fresh data—where will you get it? Ongoing data validation, curation, and governance will be essential for successful AI projects.

Building a Data Integrity Foundation

If your organization is already proficient with data management and analytics, you're better prepared to dive deep into your data and define standards for AI-ready data. Here are seven steps to take.

1. Know what *your* data looks like

The first step to building a data integrity foundation is to know your organization's data, or at least, the data that you plan to use for AI. You might know generally what kinds of data live in repositories like an inventory management database, call center system, email, or product lifecycle management (PLM) system. It's likely that you don't know who created any specific file, the content, whether it's an original file or a composite of data from multiple documents and creators, or where it has moved outside of the system.

Looking at log files for these systems won't give you the granular insight you need. They might log the instance of when data was created. They might—or might not be—logging actual data content. If you plan to use it for AI, you need to see the content and complete DNA of the actual document or binary data itself, not just the ledger and the log file.

Answers to these questions will give you a clearer picture of your data now:

- **What is the data provenance?** Where did the data come from—persons, devices, systems, external sources, the internet? If it came from outside of the organization, where was it gathered? How did it arrive?

- **Is it original?** If generated by a person, who? If by a device, which device? Understanding which person, group, or device created the data provides clarity into its original purpose.

- **What was its purpose?** Is its original purpose aligned with the AI model purpose?

- **Is the data qualifiable?** Is it complete? Is it verifiably accurate?

- **What is its structure?** Does the data represent text, an image, video, numerical values, or other format?

- **Does it have integrity?** Is the data original or a derivative of a dataset? Has it changed from its original state, and if so, how? Is it synthetic data?

- **What is its destiny?** Where is data allowed to move and be used? Who should be allowed to use it? What is allowed to happen to it—whether it's acted on by people, other systems, devices, applications, or models?

2. Verify data sources to establish provenance

Automated data discovery tools find and map simple information across an organization's systems. For data security compliance requirements, simply identifying sensitive data and knowing where it lives might be sufficient. For AI, that's not enough. To validate AI-ready data, you need to know where it originated and its journey—you need to know its provenance.

For instance, was the data generated internally according to a defined process, such as product ordering data? Was it collected with random input from external sources, such as a chatbot assisting a customer? Is it data that was simulated in another AI model? Knowing the provenance helps you know if it's suitable for your model, what preparation it might need, and how it should be weighted.

3. Classify, label, and document your data

Like data discovery, organizations use data classification to distinguish sensitive and other categories of data for security and compliance purposes. In AI, data classification must go further, and it's why metadata management and better classification tools are needed. You need the ability to distinguish high-quality data from low-quality data, as well as redundant, outdated, inaccurate, trivial and other data that will negatively affect your results.

Data should be labeled because supervised learning relies on trustworthy, ground-truth labels. Everything about your data also should be documented, including sources, label definitions, and collection methods.

4. Control data quality

Forbes contributor Bernard Marr[11] describes a company that invested heavily in an AI project, and its highly sophisticated algorithms failed because no one addressed the stores of fragmented, inconsistent data across 27 legacy systems. The AI solution "was like buying a Formula 1 car when you only have dirt roads to drive on." In another example, "a healthcare system wanted to use machine learning to predict patient readmissions. Six months into development, the team discovered that their historical patient records—the data they were using to train the AI—contained significant biases in how various conditions were coded across different facilities. The AI was learning these inconsistencies rather than genuine medical patterns."

These examples illustrate how critical data quality is to a model. Is your data accurate? Was it entered and processed correctly from its creation? Does it truly reflect the real-world thing or state that it's intended to describe? Is it complete? Are there inconsistencies, especially across multiple data sources? If you don't have processes in place to verify accuracy, you will need humans to do it.

[11] Forbes , https://www.forbes.com/sites/bernardmarr/2025/03/19/the-ai-graveyard-7-deadly-mistakes-that-kill-most-enterprise-ai-projects/

Second, the data must be relevant to the range of scenarios that your AI model is expected to handle. With low relevancy, your AI model will learn from inconsistencies instead of meaningful patterns. Too little data leads to inaccurate results.

Finally, data must be as free from bias as possible. People with training in math and statistics will be familiar with basic types of bias, but when it comes to AI-ready data, there are a number of new types of bias emerging. Models naturally pick up and amplify biases buried in data sets.

5. Ensure data privacy and security

Data gathered for AI use must be handled with awareness of global privacy and industry regulatory compliance requirements. Mixing regulated data with non-regulated data in a data pool for AI use can land you in regulatory trouble and increase the risk of a data breach.

But there's more to data privacy than its regulatory status. Stakeholders from different departments in your organization are closest to the data they use and might have a much different view of data risk than the AI team. Before data is harvested for use in AI, it's best to involve those data owners or teams responsible for it.

Finally, ensure that your data access methods, repository, and safeguards meet your security team's standards. Data should not be moving out of the organization, across models, or to third-party LLMs unless you specifically allow and monitor it.

6. Pick fresh data

Data for training should reflect your current operating conditions. If it doesn't, the model will deliver sub-par performance in the live business environment. Over time, models need fresh, real data to maintain predictability and the results you want. This means establishing a process with human oversight to ensure that data—and your data pipelines—stay accurate, relevant, compliant, and deliver high quality over time. Datasets need to be continually monitored, updated, and refined for quality AI-driven insights.

7. Curate and prepare your data FIRST

This is how *not* to curate your data: "We check our data when it's going into the model." This is the fastest way to add significant processing overhead, model bloat, and cost. Your results will be so expensive that they will outweigh the value of the findings. Set your standards, curate, and prepare your data before it goes into the pool for AI use.

Data Validation, Simplified

As you're probably feeling right now, validating data for use in AI is a little daunting. We'd wager that almost every organization launching an AI initiative—instead of working through the steps outlined—started from an assumptive perspective. They assumed:

- The needed data resides in one or two places
- The data is relevant, complete, and correct
- If any curation or cleanup is needed, it's minimal
- Simply writing an AI script will remedy any data inconsistencies
- If data is incomplete or missing, they can make it up or fill it in with other data
- Made-up or added data comprises such a small percentage of the dataset that statistically the probabilities are good enough

Let's see how this plays out.

AI, help me know why products were returned

Suppose you're a retailer and your company brought on two new manufacturers for this new season. At the end of the season, you want to know how their products performed before deciding to renew the agreements. One way to determine success is to find out how many of these products received complaints or were returned—and why. You want AI to help answer that question.

Let's say that Flying Cloud has been asked to help with this data validation piece of their AI project.

Flying Cloud (FC) What data do you want to see?

Retail company (RC) We want to know how many people called to return a product from either of these two manufacturers and their reason for the return.

FC Where is that data? What systems house it?

RC It would be in our call center system with recordings and transcripts.

Flying Cloud connects to the call center system via API. Now we can automatically see all of the call center data currently on the network and records created for past "X" number of days. We ingest that data. Our data accountability assessment sees the actual call transcript content for each record during that time period, and we can see which calls were about product returns.

FC For the X-day timeframe, some of the transcripts in your system are incomplete. They don't give a reason for the return or they wanted to exchange the product for something else. Of these calls, how many products were actually returned to inventory?

RC I don't know. We'd need to see our warehouse inventory database.

Flying Cloud now connects to the warehouse inventory database and ingests return records for the X-day time period.

FC The inventory returns database had records for only 30% of the timeframe. Would returns have gone somewhere else during the rest of this time period?

RC There is a smaller warehouse that received product back, but they aren't tied into this system.

FC Do you have other channels for customers to return products? Maybe email or chatbot assistance? Maybe they fill out a return form sent back in the package with the item?

RC There is a self-service portal where customers can return items. That's a different system.

Flying Cloud now connects to the customer portal system and after ingesting data for the X-day timeframe, finds that another 15% of customers used it to return the product—most of the returns could be verified through the inventory records, but not all. There were still some alleged returns that were not able to be traced to goods received back in inventory. They could be misplaced in a warehouse or still in process.

Now what?

Our retailer now has a much better picture of the data they *do* have, and it's clear that it's not ready for the AI model yet. The data found is incomplete for the desired time period. This may result in lacking enough data for training the AI model. Data quality is inconsistent. For example, the call records don't directly correlate with actual returns and not all call transcripts provide reasons for the return. Different warehouses coded returned items differently. Data is in different formats and systems—call center audio files, text transcripts, structured database inventory records, and chat streams.

At this point, the existing data must be curated and normalized. Missing data must be found, records must be made consistent, and there must be sufficient data for the desired time period to even make an AI effort worthwhile. This will require human assistance to bring the data up to an acceptable level of AI readiness.

This is also the time when customers ask, "can we just skip that part?" Well, you could. You would join the 99% of failed AI projects because the model can only work with what it receives and it will find the anomalies and double down on them. You won't get the answers you need.

Set up your data program to succeed

A better way to create your data integrity foundation is to use Flying Cloud for a data accountability assessment of your data sources before you tap them for an AI project. The more data sources Flying Cloud can connect to, the higher the fidelity of the results. In other words, the more data assessed, the clearer the picture you will have of your actual data. For every record and binary, Flying Cloud can tell you:

- When, how, and by whom (or what) it was created
- Where it's stored
- If, when, and how the data changed after it was created
- Where the data moves
- If it is complete
- If it's the right data for the use case
- If sensitive or regulated data is attached to source documents and files
- If there are gaps from the data source, indicating a potential system or recording problem
- ...and many more details, depending on the specific questions being asked

We start a data accountability assessment with a simple data repository, such as email. Flying Cloud instantly sees all of the data, which is automatically fingerprinted so it can be tracked. We also see into attachments and binaries—regardless of file type or data format—and can even read text embedded in images.

We gather findings and review them. For one organization, our initial email assessment quickly found a disturbing amount of regulated data—PII, credit card numbers, and patient IDs—written into the emails themselves or attached in other documents, such as a spreadsheet. Regulated data was leaving the organization unseen, putting them at risk of data breach and compliance fines.

Gain ledgerized data

A data accountability assessment results in a report that details findings and recommendations. These findings are your stake in the ground for establishing data verification, validation, and normalization standards.

Flying Cloud can initially verify and continue to monitor your data sources for data completeness, security, and notification of any changes. We validate data by assigning policies to the data and enforcing them to ensure that you are notified if they change or fail to meet your desired standard.

Next, Flying Cloud automatically processes and ledgerizes validated data, giving you a chain of custody from when data was created to when it is processed by the AI model. Validated data can be used to create a “clean” data lake for AI use. New data streams from other repositories can be validated before they are added to the LLM, ensuring that everyone working on an AI project pulls water from the clean pond.

Determine your AI data policies

Policies should define which data can be used for AI purposes. They should include decisions about:

- Where data sets can come from
- Data quality thresholds, based on assessment findings
- Which data users are allowed to feed into other organization’s AI applications, knowing that the queries become part of the other organization’s AI model
- Ensuring that sensitive data or IP is not used

Set human AI policies

These might include articulating rules or best practices for developing AI projects. Think about revisiting user privilege and password policies to update them if necessary for AI security. Ensure that humans are part of AI-based processes with the authority and ability to monitor, intervene if necessary, and question results.

Care and feeding of your AI

The bottom line with AI will always be the data fed into it. Data accountability is essential to a data integrity foundation for any AI project or tool. With 12 data surveillance patents, Flying Cloud is enabling companies to look at their data analytically and forensically with the ability to completely control where it moves. For the first time, you can easily see, track, characterize, and defend your data—to build a solid foundation for your AI future.

Next Steps

For more information about a Flying Cloud data accountability assessment, visit www.flyingcloudtech.com/contact/