WHITE PAPER

The Data Deficit: What's Behind AI's Worst Flaws

Solving the LLM quality and safety problem

TaskUs™

A fintech platform issues higher interest rates for people of color. A recruitment tool screens out qualified women for technical roles. An autonomous vehicle (AV) fails to detect a pedestrian, resulting in a fatal crash. These are documented AI failures that have cost companies millions of dollars.

AI makes mistakes — sometimes at scale, often with devastating consequences. Although the headlines focus on hallucinations and rogue outputs, the real problem sits deeper in the stack.

Why AI goes off the rails

AI is only as good as the data it's trained on. Much of that information is often flawed.

The age-old principle of "garbage in, garbage out" still holds true. And unlike traditional software bugs that can be patched, data-driven errors require a complete overhaul of training datasets and model architectures.

The concern is undeniable. Recent research reveals that 55% of Americans are worried about bias in AI (when algorithms favor or discriminate against certain groups or individuals).

The business impact is equally problematic. One survey found that 62% have lost revenue to AI errors, 35% have paid legal penalties and 61% have watched customers walk away.

Yet, fewer than 25% of enterprises have AI governance frameworks in place.

Regulators are stepping in. The EU's AI Act could penalize businesses up to €35 million for violations. The United States, China and more are mandating bias testing and safety audits. For Level 4/5 vehicles, the U.S. National Highway Traffic Safety Administration is requesting more data and transparency from fleet operators and manufacturers. The message is clear: fix your algorithms, or pay the price.

AI safety isn't a code problem

Safety and accuracy are built (or broken) at the data layer. Here are 5 areas to focus on:

01 Robustness

AI needs to function in the real world and not just in a sandbox. When training data is brittle or blind to edge cases, models crack under pressure. Diverse, high-quality datasets with adversarial examples and noisy inputs ensure accuracy and reliability.

02 Alignment

AI behavior must match human values and intentions. Otherwise, systems will act in ways we neither want nor understand. Preference comparisons and human feedback help point models in the right direction.

03 Fairness

AI must prevent discriminatory outcomes. Bias stems from skewed, imbalanced datasets. The fix requires auditing data for representation gaps and using diverse human annotators.

04 Misuse prevention

AI trained on the open web can learn a lot of things — including how to misbehave. The right guardrails, like refusal patterns, prompt filters and red teaming teach the model what it should never do (e.g., help bad actors).

05 Long-term risk

As AI becomes more capable and autonomous, control and human oversight becomes more critical. Techniques like constitutional AI, reward modeling and simulated deliberation help codify human values in ways machines can understand and follow.

3 common data traps

Three things often cause data quality and safety challenges:

The Western web problem

Most AI models are trained on internet data that reflects Western, primarily English, language and nuances. This information on its own reflects narrow viewpoints or skewed patterns which can distort how AI interprets the world, producing biased outputs. (See Figure 1)

Some examples include: voice assistants fumbling non-native accents, image recognition models underperforming on darker skin or chatbots giving culturally off-base answers.

The quantity trap

The idea of "more data, better AI" often backfires. When companies prioritize massive datasets over quality, annotators tend to rush through tasks. Mislabeled and noisy data leads to poor performance.

The one-size-fits-all approach

AI systems developed in one region are often deployed everywhere, but without local adaptation. A driverless car system trained on U.S. roads, for instance, might misinterpret road signs, lane markings or driving norms in countries with different traffic laws and infrastructure. Global models need local context and labeling.

Figure 1

Common types of machine learning bias	
Reporting Bias	Overrepresentation of sensational or rare data skews model understanding of frequency and importance
Selection Bias	Skewed sampling from specific demographics or sources can cause misrepresentation
Prejudice Bias	Human annotator assumptions based on culture, gender, ethnicity, etc., lead to discriminatory model behavior
Measurement Bias	Mismatches between training data conditions and real-world application environments (e.g., sensor calibration issues)
Exclusion Bias	Unintentionally omitting edge cases or rare events during preprocessing or annotation

Building safer AI data pipelines

Data isn't a static asset. It must be intentionally (and constantly) designed, tested and maintained like any other core technology. Consider the following:

Design for quality

Set the standards before data collection. Build in checks and governance. Make quality everyone's responsibility and not just the annotators'.

Customize workflows

AI use cases aren't created equal. For example, a medical model needs different oversight than a content recommender. Build for specificity and let your pipelines evolve.

Track everything

Log every data point and make every annotation auditable. Real-time monitoring for bias signals and annotation quality helps prevent issues.

Keep humans in the loop

Machines don't know what's inappropriate, unsafe or unfair. Humans do. Combine tech with expert review to get both scale and nuance.

Here's an example of what the data process might look like.

Type of annotation workflows

Model-assisted review

AI proposes initial labels. Humans review, correct and approve. It combines speed with oversight to reduce errors and automation bias.

Taxonomy evolution

What qualifies as harmful or biased shifts over time? Workflows must track these changes, update guidelines and teams must re-label data to stay current.

Federated annotation

Distributed teams label data with regional legal, cultural and linguistic insight to ensure accuracy and inclusivity.

MLOps integration

Build annotation into the model lifecycle. That means faster fixes, smarter retraining and tight feedback loops when issues like bias or blind spots emerge.

Building in-house vs. external support

Building AI models in-house is possible provided there's budget, capacity and, most importantly, the right skill sets.

Many companies are opting to focus their team on innovating for the core business and turning to an external partner for scale, and specialized AI expertise.

The right partner brings global talent with deep experience in bias detection, edge case identification and policy design, as examples. Plus, they have the tools for complex labeling operations — from nuanced sentiment analysis to adversarial input detection.

Flexible staffing and operations models also allow your data pipelines to grow, pivot and adapt effortlessly as AI systems, use cases and risks evolve.

The right partner focuses on speed and safety to deliver trustworthy models without slowing your edge.

TaskUs: A trusted partner to top AI firms

At TaskUs, we combine human intelligence, ethical guardrails and flexible infrastructure to power smarter, safer systems.

We've been recognized as a Leader in Everest Group's Trust & Safety PEAK Matrix three years in a row, and as a Leader in Data Annotation and Labeling (2024).

Our strength lies in our deep expertise in not just AI and data services but customer experience (CX), risk mitigation and compliance as well.

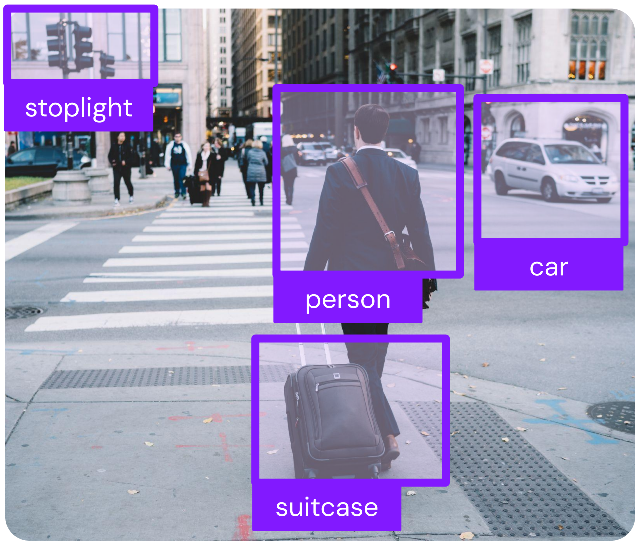
What we're known for

Global scale

60,000+ full-time employees and 30 delivery centers across 13 countries, plus 1M+ freelancers via our TaskVerse crowdsourcing platform (including PhDs and subject matter experts in mathematics, biology and more). We also support 100+ languages and dialects to reduce annotation bias and train culturally intelligent models.

Tech-flexible pipelines

We work with your existing technology and can also partner with leading data labeling platforms to boost productivity, reduce latency and provide end-to-end visibility across the annotation lifecycle.



Agile workflows

We support the entire data development stage from collection, classification, validation, consensus and enrichment. Our dynamic task routing adapts to labeler expertise, task complexity and urgency, ensuring the right work goes to the right people at the right time.

Rigorous QA & red-teaming

Our teams conduct multi-layer reviews, real-time error detection and sampling to catch issues before they scale. We can also implement red-teaming to test edge cases.

Full data security

Clients retain full control of the data which remains securely stored in their own cloud environments. We enforce strict role-based access, so that only authorized reviewers, supervisors and auditors see what they need, and only for as long as needed. We also comply with the highest industry standards, including SOC 2 Type II, ISO 27001, GDPR and HIPAA.

Project examples and results

For three leading AI firms building foundational models, we conducted rigorous adversarial testing, building expert teams and creating adversarial datasets to expose edge-case failures.

Our teams developed a new training curriculum aligned with the latest red-teaming insights, conducted peer-calibrated quality reviews and refined testing tactics. As a result, the firms saw a 71% drop in CSAM-related outputs.

In another project, we collaborated with an open engineering group to establish new benchmarks for AI safety testing and evaluation. Our red team designed 200,000+ prompts across 13 hazard categories covering English, French, Simplified Chinese and Hindi.

And when a leading European foundation model developer sought to improve its Reinforcement Learning from Human Feedback (RLHF) pipeline, we delivered structural and process upgrades.

We assembled a diverse team of reviewers across five locations, then re-engineered workflows to include consensus-first reviews, ranking disagreement audits, label inspection criteria and robust feedback loops.

For a top AV company, our annotators raised system performance with precise training data — achieving 98% accuracy. But our support doesn't stop at model training. For another major player, we delivered 24/7 remote assistance and real-time emergency response to keep operations running safely and smoothly.

Protecting annotator well-being

Recognizing the emotional toll the work can take on annotators, we implement evidence-based wellness programs to reduce cognitive load and enhance psychological resilience.

Our Wellness & Resiliency team provides proactive support, conducts research and recommends improvements to reduce bias and boost consistency when applying complex policy decisions.

The bottom line

Data is the origin point of every AI decision. Bias, inaccuracy and unsafe behavior often start — and can be stopped — at the dataset level.

Whether you're fine-tuning an LLM, building a safety testing protocol or designing workflows from scratch, we bring the right people, tools and ideas to help you get it right.

About TaskUs
TaskUs is a leading provider of outsourced digital services and next-generation customer experience to the world's most innovative companies, helping its clients represent, protect, and grow their brands. Leveraging a cloud-based infrastructure, TaskUs serves clients in fast-growing sectors, including social media, e-commerce, gaming, streaming media, food delivery and ride-sharing, technology, financial services, and healthcare. As of March 31, 2025, TaskUs had a worldwide headcount of approximately 61,400 people across 28 locations in 12 countries, including the United States, the Philippines and India.

For more information, please visit:
www.taskus.com/services/ai-data-services/

Copyright© 2025 TaskUs.
All rights reserved.