AI Transparency Is Slipping (And Thats a Problem)

AI models are quickly turning into critical infrastructure: they shape how people search, write, create, and make decisions at scale. Yet a Stanford-led research team argues that the industry is moving in the opposite direction on disclosure, with companies increasingly withholding information that would enable meaningful scrutiny.

The findings come from the *2025 Foundation Model Transparency Index*, which evaluates major AI developers on a 100-point scale across areas such as training data disclosure, risk mitigation, and broader impacts. According to the index, the overall level of transparency is low, and it has declined compared to the previous year.

What the index measures

The Foundation Model Transparency Index is designed to score how much companies disclose about their flagship foundation models and related practices. It spans multiple dimensions—such as where data comes from, what safety processes exist, and what is known about downstream use and impacts—because transparency isn't a single checkbox: it's a collection of concrete, verifiable disclosures.

In 2025, the index assessed 13 companies and found large variation in disclosure practices. The average score was about 40/100, and companies tended to fall into three rough clusters: top performers around 75, a middle group around 35, and low scorers around 15.

Big gaps between companies

One striking result is how uneven disclosure has become. IBM sits at the top with a reported 95/100—described as the highest score in the index's history—and is highlighted for providing unusually detailed information, including enough detail for external researchers to replicate training data practices and allowing access for external entities such as auditors.

At the other end, xAI and Midjourney are reported at 14/100, with the write-up stating they share essentially no information about training data, risks, or mitigation steps. The overall picture is not just "some companies are better than others," but that a meaningful portion of the market is operating with minimal public accountability.

The environmental blind spot

The Stanford HAI write-up emphasizes a major omission across the board: environmental impact. It claims that 10 of the assessed companies disclose none of the key information related to environmental impact (including energy usage, carbon emissions, or water use), which matters because datacenter expansion and training workloads have real-world resource costs.

This is a practical transparency issue, not a philosophical one. Without standardized disclosures, it becomes difficult for policymakers, researchers, and even customers to compare footprint claims or validate sustainability commitments.

Openness isn't the same as transparency

A useful distinction in the piece is that "open" (publishing model weights) doesn't automatically mean "transparent" (explaining practices and impacts). The write-up warns that

even influential open-weight developers can still be opaque about core items like training compute, risk assessment, and downstream use.

This matters because public debate often treats open-weight releases as a proxy for accountability. The index argues that assumption is risky: disclosure must be specific, structured, and repeatable to support real oversight.

Why this matters now

The researchers frame transparency as an "essential public good" for governance, harm mitigation, and oversight, and they point to growing policy interest in mandating disclosures for frontier AI risks. They also note that the index aims to help identify which transparency areas resist improvement without policy pressure.

Even if one disagrees with any single score, the direction of travel is the key signal: as foundation models become more central to economies and institutions, baseline disclosure is not keeping up.

Contributor: Alessandro Linzi