

The Reverse Information Paradox: Can Open-Weight AI Protect Enterprise Knowledge?

Nael El Batji

nael.elbatji@gmail.com

Lycée Français de New York

Abstract

Artificial intelligence is transforming how corporations and government agencies operate, but its adoption carries a hidden cost. To extract meaningful value from AI, organizations must share their internal context: proprietary data, expert reasoning, workflows, and corrections. The more they share, the more useful AI becomes, while at the same time, the more of their competitive knowledge leaves their direct control. In 2026, Microsoft CEO Satya Nadella described this tension as the “Reverse Information Paradox,” building on Kenneth Arrow’s 1962 observation that revealing information to demonstrate the value of something risks giving that value away.

This paper examines how that paradox manifests itself in enterprise AI adoption, why it represents a genuine strategic risk, and what companies and agencies can do about it. It evaluates open-weight AI models, which can be run on private infrastructure controlled by users, as a partial mitigation strategy. Drawing on Stanford’s 2025 AI Index, it notes that the performance gap between leading open-weight and proprietary models has narrowed from 8% to approximately 1.7% between January 2024 and February 2025, making open-weight models a credible option for many professional workloads. It also discusses both the advantages and limitations of open-weight models vs closed model.

The paper concludes that the future of enterprise AI will be hybrid. Competitive advantage will be achieved not by companies that adopt the most powerful model universally, but by those that build an orchestration layer capable of routing workloads to the right model based on sensitivity, complexity, cost, and regulatory context, all while retaining control over the knowledge that matters most.

Table of Contents

I. Introduction	4
II. The Impact of the Reverse Information Paradox on Companies	5
III. How AI Learns From the Knowledge We Give It.....	5
IV. The Enterprise Value-Leakage Risk.....	7
V. Open-Weight AI as a Mitigating Alternative	7
VI. The Benefits of Open-Weight Models for Sensitive Workloads	9
VII — The Disadvantages of On-Premise Open Models.....	10
VIII: Conclusion: The Future will be Hybrid	11

I. Introduction

Artificial intelligence is rapidly becoming part of the infrastructure of modern companies. Engineers use AI to write and debug code, pharmaceutical researchers use it to analyze scientific results, lawyers to review documents and financial analysts to draft presentations, reports and even suggest stock to buy. However, this creates a paradox. To unleash the full power of AI, companies need to provide context: internal documents, proprietary research, code, customer information, technical specifications, hypotheses, corrections and sometimes the reasoning of their own experts. In other words, the more value companies want to extract from AI, the more of their business knowledge they have to share with AI.

In 2026 Microsoft CEO Satya Nadella recently described this problem as the “**Reverse Information Paradox**.” His argument builds on an older economic problem identified by Kenneth Arrow more than sixty years ago. Indeed, in 1962, economist Kenneth Arrow described the “Information Paradox”: a buyer needs to know enough about information about a product to assess its value, but if too much is revealed, the buyer may already have received part of that value. For example, a scientist trying to sell a new formula must explain enough to attract a buyer without giving away the key idea.

When Satya Nadella coined his “Reverse Information Paradox” he meant that companies are effectively paying twice for AI, first through subscriptions, API fees, or computing costs, and then by sharing some of their proprietary knowledge, thus exposing themselves to unintentional knowledge leakage. And this is a much higher price!

II. The Impact of the Reverse Information Paradox on Companies

Companies pay to access sophisticated AI systems, but to get useful results they also need to provide valuable contextual internal information, such as data, business strategy, research, software code, or expert knowledge.

This does not mean that every data submitted to a commercial AI system automatically becomes training data. OpenAI states that data submitted through its business products and API is not used to train its models by default, and Anthropic makes similar commitments for its commercial products. The concern is not that an AI provider might “steal” proprietary data, it is that AI may learn and get better through the contextual learning derived from the back and forth prompting and fine tuning during its interactions with users. The issue is one of control: companies need to understand where their proprietary information is processed, how long it is retained, and how it is used by models to improve and fine tune their answers and get better. The concern is **how much of a company’s knowledge leaves its direct control when it is fed into AI if AI becomes imbedded in a its daily business.**

III. How AI Learns From the Knowledge We Give It

Understanding the problem requires understanding, even if at a simplified level, how AI systems develop their capabilities. Large language models do not learn mathematics or scientific concepts in the same way humans do. Instead, they are trained on very large amounts of human-created material and learn patterns in language, reasoning and problem-solving.

Let’s use mathematics as an example. An AI model develops mathematical capabilities through several stages. It all begins with pre-training, during which the AI learns the “grammar” of mathematics. To do this, it is exposed to millions of mathematics books, research papers, and worked-out exercises. For example, after seeing many quadratic equations together with their solutions, it can learn patterns linking the form of an equation to methods such as factoring or the quadratic formula.

The second stage consists of learning through rewards. Simply predicting the next word is not enough. Researchers therefore use reward-based methods. The reasoning produced by the AI is evaluated, and the model receives a reward when its reasoning leads to a correct result. When it makes a mistake, it is encouraged to revise its approach until it reaches the correct answer. This process helps optimize the performance of the AI agent.

The final element is how the model is prompted. The quality of an answer can depend partly on how a task is presented. A detailed prompt can specify the objective, provide relevant information, show examples or impose a particular method. This additional context helps the model identify which of the patterns it has previously learned are most relevant to the problem. For difficult tasks, AI systems can also be designed to break problems into smaller steps, verify intermediate results or use multiple agents to compare different approaches. These techniques can significantly improve performance on complex mathematical tasks.

This distinction becomes particularly important in a corporate environment. The underlying model may already contain broad capabilities learned during model training, but employees can

make it more useful by providing proprietary context such as internal documents, databases, instructions or workflows. In most cases, this does not mean that the underlying model is retrained every time an employee submits information. Instead, company knowledge is used as context that allows the AI system to fine tune its existing capabilities to specific problems.

The competitive asset is therefore not only the model itself. It is the combination of the model with a company's data and context. Together, these elements can represent a significant form of enterprise intellectual capital and competitive advantage.

Case Study 1: Mathematics and the Navier–Stokes Controversy

A recent controversy in mathematics shows why confidentiality is becoming an important issue in corporate and academical use. In 2026, NYU mathematician and professor Tristan Buckmaster and Levent Alpöge, a mathematician from Antropic, were working on a difficult problem related to the Navier–Stokes equations, one of the famous Millennium Prize problems, and were using AI tools including OpenAI's Codex as part of their research process.

Before their work was published, OpenAI launched its own intensive effort on similar mathematical problems. According to *The Washington Post*, the company first used around 100 coordinated AI agents on a related Euler-equation problem and later deployed roughly 10,000 agents for about 88 hours on the Navier–Stokes problem, after which it announced a possible solution.

The timing raised concerns within the mathematical community. Buckmaster questioned whether information from his earlier Codex interactions could have provided OpenAI's with the crucial idea and the direction to solve this equation. OpenAI denied leveraging his private inputs, and there is no evidence that his conversations were used to obtain the result. However, although the episode does not prove that confidential information was accessed, it raises an important privacy question: could Buckmaster's work, findings, or reasoning shared through Codex have influenced or guided OpenAI's own agents on a different task for a different user?

Now imagine the same issue in pharmaceutical R&D. A company could spend years and billions of dollars developing breakthrough treatment, using AI to analyze unpublished data, failed experiments, and hypotheses. If that knowledge were exposed or became accessible outside the company's control, a competitor could potentially accelerate its own research, eroding years of investment and destroying a major competitive advantage before the drug even reached the market.

Case Study 2: The New York Times v. OpenAI and Microsoft

The *New York Times* lawsuit against OpenAI and Microsoft raises a different but equally important question about privacy and control over information.

In December 2023, *The New York Times* sued OpenAI and Microsoft, claiming that millions of its articles had been used without permission to train AI models, something that, if true, would

be data theft. The newspaper also argued that these models could sometimes produce text very similar to its original articles. OpenAI and Microsoft deny these claims.

In addition to the data theft claim, this lawsuit also revealed another risk: information shared with an AI platform may not always stay only between the user and the AI provider. Indeed, during the lawsuit, a federal court ordered OpenAI to provide 20 million anonymized ChatGPT conversations. The messages were not made public, but the case shows that information shared with an external AI provider may sometimes have to be stored or disclosed for legal reasons.

For entities handling sensitive information or sovereign information, privacy is not only about whether the AI provider uses the data for training. It also depends on where the data is stored, how long it is kept, and who can access it.

IV. The Enterprise Value-Leakage Risk

For companies, a broad way to describe this risk is **value leakage**. Some of the most valuable knowledge inside a business does not exist in a single confidential document. It exists in how people work: how researchers interpret an unexpected result, how senior engineers correct a design, how a lawyer identifies contractual risk or how an expert repeatedly refines a prompt until an AI system produces the right result.

Over time, these decisions and corrections become part of what makes a company unique. As AI becomes part of everyday work, companies need to decide how much of that knowledge they are willing to share outside systems they control.

The emergence of increasingly capable open-weight AI models may give companies a path through this challenge. Instead of sending every workload to an external frontier-model provider, companies can increasingly operate powerful open-weight AI models inside a private infrastructure that they control. The question is no longer which AI model is the most powerful, but which model is best suited to each type of information, which security requirements it must meet, and who should ultimately control the data and knowledge created through its use.

This is where Open-Weight AI becomes strategically interesting.

V. Open-Weight AI as a Mitigating Alternative

AI models can broadly be divided into three categories: closed proprietary models, open-weight models, and more fully open models.

Closed Models

Closed models are controlled by the companies that develop them. Users generally access them through a website, application, or API, while the provider retains control of the model weights and, in many cases, the infrastructure on which the model runs.

Examples include proprietary models from OpenAI, Anthropic, and Google.

The advantages are significant. Companies can access highly capable AI without purchasing GPUs, deploying inference infrastructure, maintaining model software, or managing the underlying system. Much of the technical and operational burden remains with the provider.

The trade-off is dependence. The customer relies on the provider for model availability, pricing, upgrades, product policies, and most importantly the infrastructure through which its information is processed.

Open-Weight Models

An open-weight model is a model whose trained parameters, or weights, can be downloaded and operated on infrastructure controlled by the user. During training, a model adjusts billions of numerical parameters that encode patterns learned from data. These trained weights can then be released back into the open community, for another organization to download the model and run it independently again.

However, companies may decide not to release trained loads into the open, instead keeping them in their private control. Alternatively, a developer may release the trained weights while keeping the complete training dataset, training code, data-processing pipeline, or other elements of the development process private.

Prominent open-weight model families include Llama, Mistral, Qwen, DeepSeek, and Gemma. For companies, these models can be operated on infrastructure that the company itself owns and controls.

Are Open-Weight Models Powerful Enough?

Historically, companies faced a relatively straightforward trade-off: proprietary frontier models offered the strongest capabilities, while open models offered more control but significantly weaker performance. However, that gap has now narrowed.

Stanford's 2025 AI Index reported that the difference between the leading closed and open-weight models on the Chatbot Arena leaderboard fell from 8.04% in January 2024 to approximately 1.70% by February 2025. Recent benchmark results suggest that the performance gap between open-weight and proprietary AI models has narrowed significantly. The results of these benchmarks show that open-weight models are becoming strong competitors to proprietary AI systems. Models such as Llama, Qwen and Mistral now perform close to leading closed models on many benchmarks, and can sometimes do better in areas such as coding or mathematics.

However, no model is the best at everything. Proprietary models may still have an advantage for some complex reasoning tasks, while open-weight models can be just as effective for more specific technical uses.

This means that although open-weight models have not yet completely caught up with proprietary AI, but the gap is now small enough for companies to seriously consider them for many key professional activities and domains.

VI. The Benefits of Open-Weight Models for Sensitive Workloads

The analysis below lists the benefits that can be derived from using Open-Weight Models over Closed Models for sensitive work loads

Greater Protection of Sensitive Data

One of the main advantages of open-weight models is that companies can run them on private infrastructure that they control. Prompts, documents, corrections, and other sensitive information remain within the company's own environment rather than being sent to an external AI provider.

This is particularly important for industries handling confidential data, such as healthcare, finance, legal services, or R&D or confidential government work. A healthcare organization, for example, could use an open-weight model to analyze private patient records while keeping that data within its own secured infrastructure. Companies can finetune the original models without sharing the benefits with their competitors.

Better Performance on Specialized Tasks

Open-weight models can be adapted to a company's specific needs using internal data, domain knowledge, fine-tuning, Retrieval-Augmented Generation, or specialized workflows. This can make the model become highly performant for a particular industry, language, or business process. The main advantage is not that the model itself is more powerful, but that it can be adapted to the company's precise needs using its own data, needs, and ways of working.

More Predictable Costs

Open-weight models are often free to access, but they are not free to operate. Companies still need computing power, storage, technical staff, security, and maintenance. This comes at a cost. However, this cost is predictable, and exposure to future changes in API or licensing prices from Closed Models is avoided. In addition, for large and repetitive workloads, owning or renting dedicated infrastructure can be economically advantageous.

Greater Transparency and Auditability

Open-weight models can also give companies greater visibility and control over how the system is deployed. They can inspect and manage the model version, configuration, access controls, logs, connected tools, and supporting software more closely.

This does not mean that companies can fully explain how the model "thinks." Instead, the real advantage is greater transparency, auditability, and control over the AI environment.

VII — The Disadvantages of On-Premise Open Models

Open-weight models do not only come with advantages: they also shift more overall responsibility to the companies that use them. Some of these disadvantages are listed below:

Less Vendor Support and Protection

Compared with enterprise AI services, companies will receive fewer guarantees around Service Level Agreements, technical support, warranties, and legal indemnification. The companies using them are pretty much on their own to fix any technical issues if and when they arise

Greater Cybersecurity and Compliance Responsibility

When a company runs AI on its own infrastructure, it becomes responsible for securing the system. This includes access controls, backups, software updates, cybersecurity, monitoring, and regulatory compliance.

Uncertain Upgrade Path

Any upgrade to a new version of an Open Weight Model when it is released by the open community will not be straight forward activity. And again, the responsibility for upgrade compatibility rests solely with the users.

In short, greater control comes with greater operational responsibility.

VIII: Conclusion: The Future will be Hybrid

The choice between proprietary and open-weight models involves striking a balance between performance and control. Frontier proprietary models may still offer stronger capabilities for some complex tasks, while open-weight models give companies greater control over sensitive data and deployment. The key question for a company is therefore whether the additional performance of a proprietary model justifies sending a particular workload outside its direct control.

Open-weight models are particularly attractive for sensitive, high-volume, specialized, or regulated workloads where privacy, customization, and control are more important than access to the absolute strongest model or when cost predictability over the medium term is essential. However, as the responsibility for designing, deploying, securing, and maintaining the system remains with the company, Open-Weight Models come at a greater technical complexity price.

The future of enterprise AI is unlikely to be purely closed or purely open. It will be hybrid.

Proprietary frontier models will remain important when companies need the highest performance, fast deployment, and strong reasoning capabilities. At the same time, open-weight models, which can be run on a company's own servers or private cloud, will be a safer choice when sensitive data, intellectual property, strict regulations, or large amounts of data are involved.

The main challenge will be choosing the right AI model for each task. Instead of using one model for everything, companies will need an AI orchestration layer that will help them route prompts to the right models depending on the type of data, the level of security required, the complexity of the task, the cost, and the regulatory requirements. In this environment, competitive advantage will not come from adopting the most powerful model everywhere. It will come from designing an AI architecture that uses the right model for the right task, while preserving control over the knowledge that matters most.

On a personal level, I have started experimenting with Qwen, an open-weight AI model that I run locally on my Macbook Pro M5 with 16GB of ram. I have been using Qwen3 8B with Ollama and n8n to build a personal AI workflow that can analyze emails, identify important tasks, and extract deadlines. Through this project, I have learned how to connect a local AI model to external tools and automate everyday tasks while maintaining control over the system.

This experience has shown me that AI is not just about developing powerful models, but also about finding practical ways to use them. As someone who enjoys coding, robotics, and experimenting with technology, building my own AI workflow has strengthened my interest in understanding how AI can be integrated into real-world applications responsibly and effectively.

References

1. Interviews with Julien Mazaudier, Head of Sales of Accenture Technology, Switzerland
2. Satya Nadella (2026). The Reverse Information Paradox. July 12, 2026. [The Reverse Information Paradox | Satya Nadella \(X\)](#)
3. Arrow, K. J. (1962). Economic Welfare and the Allocation of Resources for Invention. National Bureau of Economic Research. [Economic Welfare and the Allocation of Resources for Invention | NBER](#)
4. Shao, Z., Wang, P., Zhu, Q., et al. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv. [DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models | arXiv](#)
5. The Washington Post (2026). He Was Close to a Huge Math Breakthrough. Then He Got Scooped by AI. [He Was Close to a Huge Math Breakthrough. Then He Got Scooped by AI | The Washington Post](#)
6. OpenAI (2026). On the Navier–Stokes Millennium Prize Problem. [On the Navier–Stokes Millennium Prize Problem | OpenAI](#)
7. OpenAI (2025). Fighting the New York Times’ Invasion of User Privacy. [Fighting the New York Times’ Invasion of User Privacy | OpenAI](#)
8. Stanford Institute for Human-Centered Artificial Intelligence (2025). Artificial Intelligence Index Report 2025 — Technical Performance. [Technical Performance | The 2025 AI Index Report | Stanford HAI](#)
9. Hubert, T., Mehta, R., Sartran, L., et al. (2026). Olympiad-Level Formal Mathematical Reasoning with Reinforcement Learning. Nature. [Olympiad-Level Formal Mathematical Reasoning with Reinforcement Learning | Nature](#)
10. McGraw Hill (2026). Why AI’s Biggest Players Want Open-Weight AI. [Why AI’s Biggest Players Want Open-Weight AI | McGraw Hill](#)
11. OpenAI. Business Data Privacy, Security, and Compliance. [Business Data Privacy, Security, and Compliance | OpenAI](#)
12. Anthropic. Is My Data Used for Model Training? Anthropic Privacy Center. [Is My Data Used for Model Training? | Anthropic Privacy Center](#)
13. Open Source Initiative (2024). The Open Source AI Definition 1.0. [The Open Source AI Definition 1.0 | Open Source Initiative](#)