PLI Chronicle

April 2026

Responsible AI in Practice. Analyzing the Potential Risks of Procuring Off-the-Shelf Models in Financial Services Use Cases

Hannah Liu

Responsible AI Institute

Responsible AI in Practice is a series featuring practical, actionable guidance for teams navigating artificial intelligence governance and responsibility, authored by the experts at Responsible AI Institute (RAI).

Off-the-shelf (OTS) models are pre-built AI systems developed and trained by third-party providers, made available for adoption without requiring the end user to design or train the model from scratch. They have emerged as a leading convenience in this new frontier of AI, saving people time and money by providing solutions that integrate seamlessly into any use case. However, in the face of climbing technical innovation in this area, the risks associated with it are rarely discussed. Industries, especially financial services, that often face socially consequential scenarios require AI tools that yield high accuracy, low bias, and full transparency. Any process within an organization's AI agent, tool, or application that violates any of these three criteria immediately introduces institutional and regulatory risk.

When using OTS models, individuals are usually aware of the most visible risks: hallucinations in outputs, data leaks, and consumer-facing errors. However, the majority of the risks associated with these models are invisible. Is it evident why the model may be producing biased output? Is the model drifting (i.e. gradually producing less reliable outputs as real-world conditions change)? Exactly where in the model's reasoning are errors taking place? These are all questions that leaders across

industries—and especially in financial services—should be able to spot and address effectively before integrating an OTS model into their workflows. But how can they do so if many of the currently available governance frameworks fail to peer into these black-box applications?

In this article, we will discuss the risks that financial services organizations may miss when using off-the-shelf models to perform specific, potentially sensitive processes. We then introduce an agent-risk classification tool developed by the Responsible AI Institute to address this gap.

Off-the-Shelf Models: What Happens Before You Receive Them?

When developing off-the-shelf models, model providers use two training phases: pre-training to establish general learning patterns, and fine-tuning to adapt them to specific tasks. However, the fine-tuning stage can inadvertently weaken or distort safety and bias guardrails in foundational models, creating consequential effects during deployment. Salient examples can be found when we analyze fine-tuning in off-the-shelf large language models (LLMs). One study has shown that fine-tuning with either a few adversarial examples or even benign datasets can compromise the safety alignment of LLMs. It has also been observed that narrow fine-tuning can lead LLMs to exhibit misaligned behavior across a range of prompts, further underscoring the scale of unintentional harm that fine-tuning can cause. In agents, we see that narrow fine-tuning yields a similar result.

Across the literature reviewed, the root of this behavior is not easily identifiable. There are explainability and interpretability methods that provide transparency in model design processes. Attention visualization helps reveal what parts of an input the model focuses on. Feature importance scoring identifies which variables most influenced a decision. Counterfactual analysis tests how changing an input would change the output. However, when fine-tuning is applied without sufficient use of these methods, we may not fully understand how these processes affect the AI application's internal decision-making mechanisms. Especially since individuals require narrow fine-tuning to make their models highly domain-specific, these procedures are what open the door to the unintentional, harmful behavior we see in the literature. This inability of the models to generalize not only limits performance but also leads them to become significantly misaligned when applied to tasks outside of their original domain.

A Case Study into Financial Services

The research overview shows how the fine-tuning step in off-the-shelf models can erode alignment guardrails across LLMs and agents. The implications this has for their users and adopters are immense, especially in financial services use cases.

Since off-the-shelf models are pre-trained, **the lack of visibility into their internal structure creates uncertainty.** Financial service institutions frequently handle sensitive data, make consequential decisions, and can affect the livelihoods of many with one action. They thrive when they receive sufficient information, and an off-the-shelf model that fails to do so introduces risk that is difficult to quantify and harder to manage. Even worse, when it is integrated, any errors can spread downstream, making its origin hard to pinpoint and preventing efficient debugging and problem resolution.

The potential harms of narrow fine-tuning also pose problems that financial institutions may not expect. We saw that a model fine-tuned on a narrow dataset for a specified domain may perform well in controlled evaluations but fail in edge cases that push the model to generalize. In financial services, these edge cases are common, often arising from unusual transaction patterns, atypical customer profiles, or market conditions underrepresented in training data. These challenges also interact with the first point. Since the fine-tuning process is also opaque to the adopter, institutions may not realize the model's limitations until a failure has already occurred. As a result, it forces these financial services companies to become reactive rather than prepared and proactive, an approach that should be circumvented with an OTS model that is ready-to-use.

Finally, **the absence of clear accountability frameworks around OTS models leaves financial institutions exposed when things go wrong.** Given the black-box nature of an OTS model and its development by a third party, determining who bears responsibility for errors is difficult.

Financial services is a highly regulated sector, and regulators will become increasingly unwilling to leave this tension unresolved. As a result, financial services companies will be pushed to create guidelines and policies to assume institutional responsibility. However, they may lack the technical and policy knowledge needed to build this sufficiently, creating a gap between technical innovation and responsibility.

Where the Responsible AI Institute Steps In

To help financial service companies proactively understand the risks that come with off-the-shelf models, we present the **Agent Risk Classification tool, or ARC.** Covering 7 agentic AI systems, ARC enables users to self-audit their organization's

AI applications. It is built on interlinking frameworks derived from the literature and our research findings, such as the GPA-IAT classification model for determining agency level, a twelve-dimensional scoring rubric for risk quantification that we developed in-house, and a five-level Levels-of-Autonomy framework that helps characterize human-agent interaction. Alongside this, it is grounded in several AI governance frameworks, such as the NIST AI RMF, the EU AI Act, ISO/IEC 42001, OWASP, MITRE ATLAS, and SR 11-7, as superseded and replaced by SR 26-2.

ARC helps pinpoint both the risks we can easily see and those buried beneath layers of the agent's reasoning architecture. The interactions between the interlinked framework structure and foundational AI governance policies significantly expand the scope of risk classification, and the underlying weighting algorithm helps users identify risks arising from the combination of several factors within a given AI agent or tool. For example, a financial services knowledge assistant may score low across the risk quantification rubric but have a high autonomy level. If we only looked at the risk-quantification rubric, we would conclude that, since the agent's responsibilities and allotted actions are categorized as low risk, the agent itself would be low risk. However, this would ignore the agent's high autonomy level, which automatically would deem it high risk.

ARC captures these interactions in its analysis, therefore helping users preempt any potential issues with their off-the-shelf model. ARC also includes general policy suggestions based on the risk classification outputs, and future work will directly feed ARC findings into a policy generator tool. This addition to our AI governance workflow provides financial services institutions with the technical and policy-domain knowledge foundation they need to begin building a robust regulatory safety moat.

Overall, we see ARC as a foundational first step toward bridging the gap between OTS model risks and integration into socially consequential industries such as financial services. It will allow individuals to fully leverage the advantages of OTS models in their use cases without falling into the associated risks.

Conclusion

When using cutting-edge AI technologies, individuals must do their due diligence before allowing AI agents, tools, and applications into their organizations. However, sometimes this due diligence is simply not enough because of the black-box nature of AI applications. To combat this, awareness and proactive measures toward understanding these underlying risks are a strong step forward. The Responsible AI Institute's research mission, as expressed through ARC, is to improve transparency and

clarity around these tools, enabling seamless integration in the future and increasing trust between humans and agents.

Hannah Liu is a Research Engineer in AI Policy and Governance at the Responsible AI Institute, where she helps develop new solutions and frameworks to bolster AI governance in organizations. She has been researching AI governance and socio-technical applications for the past 5 years and is interested in developing AI governance frameworks that are globally applicable, ethically adequate, and decentralized. Hannah holds a Bachelor's from the University of Pennsylvania in Cognitive Science and is currently a Master's student in AI Applications and Innovation at Imperial College London. You can reach out to her at hannah@responsible.ai.

Disclaimer: The viewpoints expressed by the authors are their own and do not necessarily reflect the opinions, viewpoints and official policies of Practising Law Institute.