MLCommons

Financial Services Working Group

8 July 2026

Financial Stability Board
Centralbahnplatz 2
4002 Basel, Switzerland
Submitted electronically

Re: Response to the FSB Consultation Report, *"Sound Practices for Responsible Adoption of Artificial Intelligence (AI)" (10 June 2026)* — support for developing targeted AI transparency and assurance tools (Section 3.9 / Sound Practice 12)

Dear Members of the Financial Stability Board,

This response to the FSB Consultation Report "Sound Practices for Responsible Adoption of Artificial Intelligence (AI) is submitted on behalf of MLCommons. MLCommons is a non-profit consortium that aims to accelerate the benefits of machine learning and artificial intelligence. Our members and partners include over 125 organizations from around the world, many of which are leading technology companies, startups, academics, and nonprofits that are actively researching, developing, and deploying artificial intelligence products for customers. Critically, our founding membership includes academic researchers at the forefront of machine learning research, and the research community continues to be core to our membership, helping to lead many of our working groups. MLCommons acts as a neutral nexus for commercial and non-commercial actors to collaborate on tools that advance the field.

We create, operate and maintain community assets, especially benchmarks and datasets, that facilitate developing and evaluating artificial intelligence (AI) systems in pursuit of our mission to "make artificial intelligence better for everyone."[1] The original project that brought MLCommons into being is a benchmarking suite called MLPerf, which provides unbiased evaluations of training and inference speed for AI hardware and software.[2] These measurements enable a fair comparison of competing systems, accelerate ML progress through fair and useful measurement, enforce reproducibility to ensure reliable results, and do so in an open and collaborative way to keep benchmarking affordable for all participants. In addition to MLPerf, we have an AI Risk and Reliability working group which has launched benchmarks on safety and jailbreaks.

We recently established a Financial Services Working Group to promote safe and reliable AI in financial services. The working group is focused on addressing the lack of non-comparable evidence of AI agent reliability in financial services by developing standardized agent risk and reliability benchmark evaluations supported by a shared taxonomy. The working group currently

[1] Machine learning is one of the key techniques through which AI systems are built.
[2] Peter Mattson, et al, "MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance," IEEE Xplore, accessed February 1, 2024, https://ieeexplore.ieee.org/abstract/document/9001257.

includes several banks and is co-chaired by Michael Hsu, former Acting Comptroller of the Currency.

We welcome and support the FSB's consultation report, "*Sound Practices for Responsible Adoption of Artificial Intelligence (AI)*" (Report). This letter seeks to share our perspective on the Report's discussion of AI transparency and assurance tools. We regard this as one of the report's most consequential observations and encourage the FSB to develop it further.

1. The binding constraint on responsible AI adoption is assurance

Financial institutions and regulators face challenges in assessing, proving, and *verifying* the risk and reliability profiles of AI agents. Regulators, in particular, have limited tools to independently and confidently evaluate AI system deployments given the rapid pace of change, the lack of well-developed standards, and current reassessments of traditional risk management frameworks.[3]

The resulting uncertainty about how regulators might react to AI deployments and incidents at financial institutions is holding many banks back from adopting and deploying agentic AI, which can provide increased value and capabilities to consumers and operational and security benefits to organizations.

To overcome this uncertainty, improvements in transparency and assurance are needed. As the Report notes, "Financial institutions could collaborate with authorities and third-party AI providers to develop targeted AI transparency and assurance tools (e.g., financial sector-specific model or system cards)."

We encourage the FSB to form an AI Transparency and Assurance Workstream focused on developing tools, fostering consensus, and promoting standards development for AI agents in financial services.

2. Assurance requires shared, trusted measurements

Trust in banks' resiliency is enabled by well-developed, standardized risk metrics, such as capital and liquidity ratios, and by transparency through regulatory reporting and public disclosures. Agentic deployments at financial institutions will need to meet similarly high bars of assurance to meet financial risk management standards. But as the Report observes, the transparency tools currently available — disclosures, model and system cards, and certifications and standards such as ISO/IEC 42001 — are valuable but "generally not tailored to the financial sector."

In practice, each financial institution relies on different, non-comparable evidence. AI providers cite their own evaluations and institutions build bespoke internal tests. Risk managers, auditors, and examiners have little they can compare across firms or vendors. The absence of a shared

[3] See, e.g., Federal Reserve, *SR 26-2: Revised Guidance on Model Risk Management* (April 17, 2026), which explicitly scopes out AI from model risk management supervisory expectations, stating: "Generative AI and agentic AI models are novel and rapidly evolving. As such, they are not within the scope of this guidance" (footnote 3, p3). The U.S. federal banking agencies have signaled their intent to issue a Request for Information (RFI) on AI risk management.

language and shared risk and reliability benchmarks raises diligence costs, slows deployment, and contributes to regulatory uncertainty. The development of a taxonomy and standardized measures would address that and accelerate the adoption of responsible AI across the financial sector.

We encourage the FSB to use its convening authority to sponsor a conference with key stakeholders to promote consensus in the evaluation/measurement of the risk and reliability profiles of AI agents in financial services.

3. Agents need to be scoped to be measured

The measurement problem for AI agents is challenging. As the Report notes, the object that must be assured is not a foundation model alone but the model together with its tools, controls, and workflow, operating in a defined context. Meaningful assurance must be scoped — to a defined operating context, a defined set of permitted and prohibited actions, and the institution's control environment — rather than expressed as a single, unbounded claim that a system is "safe." Tools that make bounded, well-scoped, and comparable claims are both more rigorous and more useful to risk functions and regulators.

We encourage the FSB to develop shared scoping conventions — defined autonomy levels, operating context, action classes, and control environment — so that assurance evidence is legible and comparable across institutions, settings, and jurisdictions.

4. Standardized evaluations can help build the solutions regulators need

The Report acknowledges the role that transparency, standards and certifications can play in developing assurance for third party AI deployments. Such efforts are relatively small in scale and influence currently, however, due to high start-up costs and uncertain recognition by regulatory authorities.

Standards bodies can contribute to assurance in ways that individual firms or vendors cannot:

- **Neutrality and independence.** Evidence is more credible when it is not produced by the party selling the system or grading its own work, such as an AI lab. Standardized evaluations can ensure that methodology and results are decoupled from self-interested commercial outcomes or vendor interests.
- **Comparability.** Shared methodologies and common assurance artifacts give institutions, risk managers, auditors, boards, and examiners a consistent basis for comparison rather than a patchwork of bespoke, incompatible tests.
- **Pooled expertise and shared cost.** Building realistic, sector-specific evaluation environments and ground truth is expensive and requires domain knowledge spread across institutions, providers, and researchers. Neutral parties like MLCommons can assemble that expertise and spread the costs across the sector — a burden no single agency, AI provider, or deployer can efficiently bear alone.
- **Complementing supervision.** Well-designed standards can efficiently operationalize principles-based supervision and existing frameworks — such as the NIST AI Risk Management Framework and risk-tiering approaches — into testable, reproducible

measures, giving firms and supervisors a practical instrument for applying the expectations authorities set.

Standardized evaluation efforts are also enable initiatives that are: developed collaboratively with authorities, institutions, and providers, consistent with the Report's call for collaboration; built on transparent, reproducible methodologies open to independent scrutiny; designed to produce standardized bounded, scoped assurance artifacts that carry an explicit validity window and an as-of system version; mapped to existing regulatory and risk-management frameworks; and governed to preserve integrity, with results not linked to commercial adoption.

We encourage the FSB to prioritize and facilitate the development of standardized evaluations of agent risk and reliability profiles. In particular, the FSB should issue a call for papers and/or host a convening of thought leaders on standardized agent risk and reliability evaluation frameworks in the context of financial services.

5. Regulators and standard setters need more "hands on keys" experience

Most regulatory agencies have approached the adoption of AI within their own institutions cautiously. While prudent, this has often resulted in limited regulator access to AI tools, especially frontier models and capabilities.

To engage productively and credibly on the challenges highlighted in the Report and this letter, regulators and standard setters must *use* frontier AI and *build agents*, not just study them. One cannot learn to swim from the shore. Hands-on-keys experience is the most efficient and effective way for regulators to gain the necessary AI fluency.

Barriers to AI adoption at regulatory agencies range from concerns about security to uncertainty about safeguarding confidential supervisory information. Agencies would benefit from sharing best practices and technical solutions to these obstacles.

We encourage the FSB to support and prioritize: (A) regulator training focused on AI evaluations, (B) secondments to AI providers, deployers, and third-party evaluators, and (C) sharing of best practices for rapid adoption of frontier AI within agencies.

The adoption of stronger standards and policies will be necessary, but not sufficient, to foster the development of useful assurance tools in financial services. Significant technical investment will also be required. We commend the FSB for highlighting these issues in its Report and would welcome the opportunity to contribute to the FSB's continued work in this area, including through technical input on methodology and scoping.

Respectfully submitted,

Rebecca Weiss
Executive Director
On behalf of the Financial Services Working Group
MLCommons