montana *Research Foundation*

INSIGHT REPORT · FINTECH

Agents in Financial Services

Adoption, customer-facing risk, model error rates, and the rules already in force

REPORT

2026-03

Dr. Ricardo Arcifa, Francieli Carra
Montana Research Foundation

AUGUST 2026

EXECUTIVE SUMMARY

Three quarters of UK financial firms already use AI, a third say they fully understand the systems they run, and more than a third of the US population has dealt with a bank's chatbot. Against that, a tribunal has held an airline liable for its chatbot's advice, a finance question-answering benchmark found a frontier model wrong or silent on four questions in five, and the EU has classified credit scoring as high-risk. This report puts the numbers together and sets out controls a risk function can sign off.

75%	34%	81%	$650.88
Of UK financial services firms already using AI, Bank of England and FCA, 2024	Of those firms report complete understanding of the AI they use	Of FinanceBench questions a retrieval-backed GPT-4-Turbo answered wrongly or refused	Damages Air Canada was ordered to pay for its chatbot's misstatement

CONTENTS

METHOD & SCOPE

Sources are published surveys, benchmarks, rulings, and regulation, listed at the end: the Bank of England and FCA 2024 survey of 118 firms, the CFPB's 2023 chatbot spotlight, Moffatt v. Air Canada (2024 BCCRT 149), FinanceBench (2023), Dahl et al. in the Journal of Legal Analysis (2024), the EU AI Act Annex III, and SR 11-7. Figures are quoted as published and were not re-run.

Every chart plots published numbers; none is illustrative. No financial institution's data was used. Regulatory references describe the general shape of obligations; firms should map them to their own rulebook. Findings are stated for the surveys, models, and cases named, and not as general properties of language-model systems.

KEY FINDINGS

01 Adoption is ahead of understanding. In the 2024 Bank of England and FCA survey, 75% of firms used AI and a further 10% planned to, while 34% reported complete understanding of the AI they use and 46% partial.

02 The customer-facing surface is large. By 2022, about 37% of the US population, over 98 million people, had interacted with a bank's chatbot, and all ten of the largest US commercial banks had deployed one.

03 A chatbot's statement binds the firm. A British Columbia tribunal found Air Canada liable for negligent misrepresentation by its website chatbot and awarded the customer $650.88; the airline's argument that the chatbot was a separate entity was rejected.

04 Frontier models get finance questions wrong at scale. On FinanceBench, GPT-4-Turbo with a retrieval system incorrectly answered or refused 81% of questions; in a separate study, models hallucinated on 58% to 88% of verifiable legal questions and could not reliably predict their own errors.

05 The rules are already written. The EU AI Act lists creditworthiness assessment and life and health insurance pricing as high-risk uses, and SR 11-7 has required validated, documented, and monitored models in US banking since 2011.

01

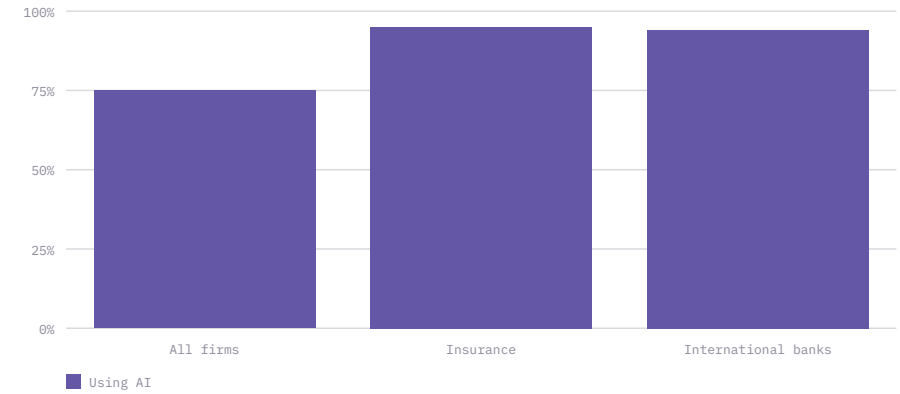
Adoption has outrun understanding

Financial services adopted machine learning earlier than most sectors and is now adopting language-model systems faster than it can explain them. The most recent supervisory survey puts numbers on both halves of that sentence.

The 2024 Bank of England and FCA survey

In November 2024, the Bank of England and the Financial Conduct Authority published their third survey of AI in UK financial services, covering 118 firms across banking, insurance, investment, and market infrastructure. 75% of firms were already using AI and a further 10% planned to within three years, up from 58% using machine learning in the 2022 survey. Adoption was highest in insurance, at 95% of firms, and international banks, at 94%. Most adopters ran few systems: 56% reported ten or fewer use cases, and 10% more than fifty.

AI adoption in UK financial services, 2024



Share of surveyed firms using AI, by group. All firms: 75% using, 10% planning within three years. Insurance and international banks reported the highest use (Bank of England and FCA, November 2024; 118 firms).

> **Why this is a control problem.** No supervisor in the survey argued against AI. The question they asked, and the one this report answers, is whether the firm can evidence what the system does and what happens when it is wrong.

Understanding lags

The same survey asked firms how well they understood the AI they had deployed. 34% reported complete understanding and 46% partial understanding. The top perceived benefits were data and analytical insight, anti-money-laundering and fraud, and cybersecurity. The top perceived risk was cybersecurity, with third-party dependencies, model complexity, and embedded models named as the risks growing fastest. A firm that runs a system it partly understands, sourced from a third party, and embedded in a vendor product has three of the survey's top risks in one deployment.

34%

Of firms with complete understanding of the AI they use; 46% partial

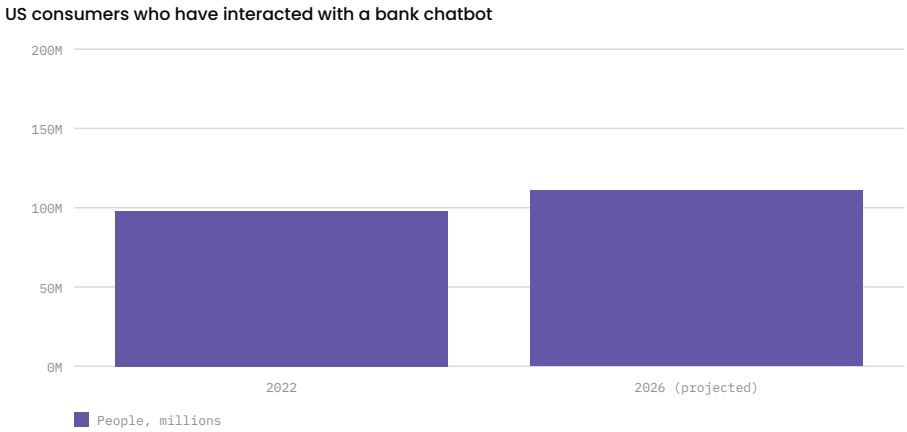
02

The customer-facing surface

The deployment that reaches the public first is the chatbot. It is also the one where a wrong answer has a named customer, a written record, and, as one tribunal has now confirmed, a liable firm.

How many people the chatbots reach

The US Consumer Financial Protection Bureau's June 2023 issue spotlight found that all ten of the largest US commercial banks had deployed chatbots, and that by 2022 about 37% of the US population, more than 98 million people, had interacted with a bank's chatbot. The figure was projected to reach 110.9 million by 2026. The CFPB's concerns were specific: chatbots that give inaccurate information, that fail to recognise a dispute, and that block access to a human, any of which can amount to a violation of consumer financial law.



Millions of people. 2022 figure as reported by the CFPB (about 37% of the US population); 2026 figure is the projection the CFPB cited (CFPB, Chatbots in consumer finance, June 2023).

Moffatt v. Air Canada

In November 2022, Jake Moffatt asked Air Canada's website chatbot about bereavement fares after a grandmother's death. The chatbot said a reduced fare could be applied for retroactively within 90 days. The airline's actual policy did not allow retroactive applications, and Air Canada refused the refund. In February 2024, the British Columbia Civil Resolution Tribunal found Air Canada liable for negligent misrepresentation. The airline had argued that the chatbot was a separate legal entity responsible for its own actions; the tribunal called that a remarkable submission and rejected it. Moffatt was awarded $650.88 in damages, the difference between the fare paid and the bereavement fare, plus interest and fees.

$650.88

Damages awarded against Air Canada for its chatbot's advice

> *Air Canada argues it cannot be held liable for information provided by one of its agents, servants, or representatives, including a chatbot. It does not explain why it believes that is the case.*
>
> MOFFATT V. AIR CANADA, 2024 BCCRT 149

The sum is small. The principle is large: a firm is responsible for what its chatbot says, the chatbot's statement is treated like any other representation on the firm's website, and the customer is entitled to rely on it. For a bank, the equivalent misstatements are a fee, a limit, a rate, or an eligibility rule, each of which is also a conduct matter.

03

What the models get wrong

Chatbots fail in the way the underlying models fail. Two benchmark studies, one on financial filings and one on legal questions, measured those failure rates on questions with checkable answers.

FinanceBench

FinanceBench, from Islam and colleagues, is a set of 10,231 questions about publicly traded companies, each with an answer and the supporting passage from the company's filings. The questions were written to be clear-cut, as a minimum standard. On a manually reviewed sample of 150 cases across 16 model configurations, GPT-4-Turbo used with a retrieval system incorrectly answered or refused to answer 81% of questions. All models examined showed weaknesses including hallucination. These are the questions a support or research assistant in a bank would be asked about a client's own documents.

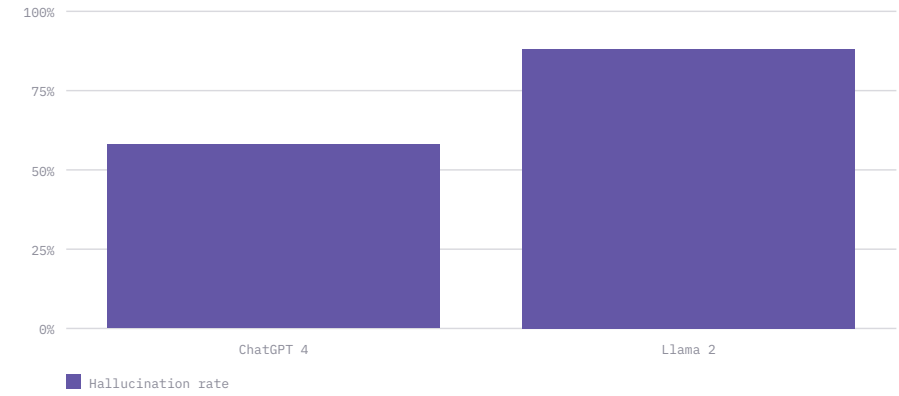
Legal hallucinations

Dahl, Magesh, Suzgun, and Ho, in the Journal of Legal Analysis, asked public models specific, verifiable questions about random US federal court cases. Hallucination rates ran from 58% with ChatGPT 4 to 88% with Llama 2. Two further findings weigh more for a regulated deployment: the models often failed to correct an incorrect legal premise built into the question, and they could not reliably predict when they were hallucinating, so their own confidence was not a usable signal of error.

81%

Of FinanceBench questions a retrieval-backed GPT-4-Turbo got wrong or declined

Hallucination rate on verifiable questions about US federal cases



Share of responses containing a hallucination when asked specific, verifiable questions about random federal court cases, by model, as reported in the abstract (Dahl et al., Journal of Legal Analysis 16, 2024).

What the two studies mean for escalation

Most chatbot designs route a conversation to a person when the model's confidence drops. Both studies undercut that design: the models answer confidently and wrongly, accept false premises, and cannot predict their own errors. Escalation should be triggered by the case, not by the model's self-report.

04

The rulebook is already written

Firms sometimes treat language-model systems as a new category awaiting regulation. In the two largest jurisdictions the relevant obligations already exist, and both reach the systems described in this report.

The EU AI Act

Annex III of the EU AI Act lists the uses classified as high-risk. Point 5(b) covers AI systems used to evaluate the creditworthiness of natural persons or establish their credit score, with an exception for fraud detection. Point 5(c) covers risk assessment and pricing in life and health insurance. High-risk status brings conformity assessment, technical documentation, logging, human oversight, and accuracy and robustness requirements, and deployers of these systems must carry out a fundamental rights impact assessment before use. A language-model component that informs a lending or pricing decision is inside that perimeter.

SR 11-7

In the United States, the Federal Reserve and the OCC's Supervisory Guidance on Model Risk Management, SR 11-7, has applied since 2011. It requires that models be validated by parties independent of their developers, that validation cover conceptual soundness, ongoing monitoring, and outcomes analysis, and that the firm keep an inventory and documentation sufficient for a third party to understand how each model works and where it is used. Nothing in the guidance limits it to statistical models. A language-model system that produces outputs a firm relies on for a business decision is a model under SR 11-7's definition.

> **The consequence.** A chatbot that quotes a fee is making a representation. A model that feeds a lending decision is high-risk in the EU and a model under SR 11-7 in the US. The obligations attach to the use, whatever the technology.

REGIME	REACHES	REQUIRES
EU AI Act, Annex III 5(b)	Creditworthiness assessment and credit scoring of natural persons	Conformity assessment, documentation, logging, human oversight, fundamental rights impact assessment
EU AI Act, Annex III 5(c)	Risk assessment and pricing in life and health insurance	As above
SR 11-7 (Fed and OCC)	Any model used for business decisions in a US banking organisation	Independent validation, ongoing monitoring, outcomes analysis, inventory and documentation
Consumer financial law, per the CFPB	Chatbots giving inaccurate information, mishandling disputes, or blocking human access	Accurate information, dispute recognition, access to a person
Negligent misrepresentation, per Moffatt	Statements a chatbot makes on the firm's site	The firm answers for them

05

Controls a risk function can sign off

None of the evidence argues against deploying these systems. It argues for treating them as what the rules already say they are: models, with validated inputs, documented behaviour, and monitored outputs, whose customer-facing statements bind the firm.

FOR	ASK FOR	BECAUSE
Model risk	The system in the model inventory, with independent validation and an owner	SR 11-7 already requires it; 34% of firms say they fully understand their AI
Conduct and compliance	Every quoted fee, rate, limit, or rule traced to a retrievable source, or escalated	Moffatt: the firm answers for the chatbot's statement
Product owners	Escalation triggered by case features, never by model confidence alone	Models cannot predict their own hallucinations
Credit and underwriting	High-risk classification and a fundamental rights impact assessment where the system informs a decision	EU AI Act Annex III 5(b) and 5(c)
Customer operations	A visible route to a person in every conversation	The CFPB names blocked human access as a potential violation
Internal audit	An error rate measured on the firm's own documents, with the benchmark method disclosed	FinanceBench: 81% wrong or refused on clear-cut questions

> **The one-line version.** Put it in the model inventory, ground every number it quotes, escalate on the case, and give the customer a person. Then the deployment is evidenced.

A checklist for the first customer-facing deployment

1. **Inventory and validation.** The system is listed as a model, with an owner, a validator independent of the build team, and a monitoring plan.
2. **Grounded statements.** Any fee, rate, limit, eligibility rule, or deadline is quoted from a retrievable policy source, or the system says it cannot answer.
3. **Case-driven escalation.** Monetary actions, unverifiable customer claims, complaint language, and ungrounded policy questions route to a person regardless of confidence.
4. **Disclosure and access.** The customer is told they are speaking to a system, and a route to a person is available at every turn.
5. **Own-document error rate.** Accuracy is measured on the firm's own product documents and filings, with method and sample size recorded.
6. **Regime mapping.** Each use is mapped to Annex III, SR 11-7, and consumer law before launch, and the mapping is kept with the model documentation.

7. **Incident record.** Every misstatement found in production is logged with the conversation, the source that should have been quoted, and the fix.

The surveys and studies above cover particular firms, models, and question sets, and the two benchmarks measure earlier models than a firm would deploy today. A firm's own error rate, on its own documents, is the number its supervisor will ask for.

Sources

1. Bank of England and Financial Conduct Authority (2024). *Artificial intelligence in UK financial services – 2024.* 21 November 2024.
2. Consumer Financial Protection Bureau (2023). *Chatbots in consumer finance.* Issue spotlight, June 2023.
3. *Moffatt v. Air Canada*, 2024 BCCRT 149 (British Columbia Civil Resolution Tribunal, 14 February 2024).
4. Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., Vidgen, B. (2023). *FinanceBench: A New Benchmark for Financial Question Answering.* arXiv:2311.11944.
5. Dahl, M., Magesh, V., Suzgun, M., Ho, D. E. (2024). *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models.* Journal of Legal Analysis 16(1), 64–93.
6. European Union (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act), Annex III.*
7. Board of Governors of the Federal Reserve System and OCC (2011). *Supervisory Guidance on Model Risk Management,* SR 11-7 / OCC 2011-12.

AUTHORS

Dr. Ricardo Arcifa
Montana Research Foundation

Francieli Carra
Montana Research Foundation

montana *Research Foundation*

An independent non-profit research institute doing open, rigorous research on the systems everyone is building. Every number in our work traces to a committed artifact, and every report cites the studies behind it.

montanaresearch.org

MRF-R-2026-03 · August 2026 ·