Position: Current XAI Methods Cannot Satisfy Financial AI Explainability Requirements

Anonymous Authors[1]

Abstract

This position paper argues that current explainable AI (XAI) methods cannot satisfy regulatory explainability requirements for LLM-based financial systems, creating a fundamental incompatibility between technological capability and legal mandate that threatens both consumer protection and financial stability. We demonstrate through systematic analysis across six regulatory frameworks (EU AI Act, US FSOC/CFPB, UK FCA, BIS, MAS, HKMA) that post-hoc explanation techniques fail systematically when applied to large language models. Exact SHAP computation exhibits $O(2^F)$ complexity at token-level granularity—rendering it infeasible for transformer architectures. LIME demonstrates substantial instability, with explanation rankings varying significantly across repeated evaluations of identical inputs. Chain-of-thought prompting generates unfaithful rationalizations: in controlled experiments, only 1 of 426 biased model outputs explicitly acknowledged the biasing feature in its explanation. When models learned to exploit reward hacks, they verbalized this exploitation less than 2% of the time. With 72% of UK financial firms now using AI and over $5 trillion in US consumer credit outstanding requiring adverse action explanations, this gap creates systemic risk affecting millions of consumers who may receive inadequate explanations for consequential financial decisions. We analyze three high-stakes domains—credit, trading, advisory—with documented regulatory enforcement cases, examine six counterarguments including hybrid architectures and outcome-based regulation, and propose prioritized recommendations with quarterly timelines. The status quo constitutes regulatory compliance theater; we call for either fundamental advances in LLM interpretability or deployment constraints matching current capabilities.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Consider a consumer denied a mortgage after an LLM-based system analyzed their application documents. Under the Equal Credit Opportunity Act, they are entitled to "specific and accurate reasons" for denial (Consumer Financial Protection Bureau, 2023). Yet when the lender runs SHAP to generate explanations, different runs produce different "primary factors"—sometimes debt-to-income ratio, sometimes employment history, sometimes neighborhood characteristics. Which constitutes the legally required "specific reason"? This is not a hypothetical: it reflects documented XAI instability that creates compliance impossibility for LLM-based systems (Alvarez-Melis & Jaakkola, 2018; Slack et al., 2020).

In December 2024, the U.S. Financial Stability Oversight Council elevated artificial intelligence explainability to systemic risk status, warning that "AI models, particularly those based on machine learning, can be difficult to explain and interpret," creating "financial stability vulnerabilities" (Financial Stability Oversight Council, 2024). This warning arrived as LLM deployment in finance reached unprecedented scale: Goldman Sachs' GS AI Assistant serves 46,000 employees (Son, 2025); JPMorgan Chase's LLM Suite supports 60,000+ employees (Son, 2024b); Morgan Stanley deploys OpenAI-powered assistants to 40,000+ advisors (Son, 2024a). These systems are moving beyond productivity tools into credit decisions affecting over $5 trillion in US consumer credit outstanding, algorithmic trading, and consumer-facing advice—applications where explainability is legally mandated (Vallarino, 2025; Bartlett et al., 2022).

We take the position that current XAI methods cannot satisfy regulatory explainability requirements for LLM-based financial systems. This incompatibility stems not from temporary technological immaturity but from fundamental methodological limitations that we formalize as four propositions (§3): SHAP's exponential complexity at token-level granularity renders exact computation infeasi-

ble (Lundberg & Lee, 2017; Kumar et al., 2020); attention weights lack causal relationships to outputs (Jain & Wallace, 2019); chain-of-thought produces unfaithful rationalizations under current training paradigms (Turpin et al., 2023; Chen et al., 2025); and the faithfulness-plausibility gap exceeds 50 percentage points across all major methods (Jacovi & Goldberg, 2020). This analysis connects to broader AI governance debates about deployment boundaries, capability thresholds, and the appropriate role of regulation in shaping AI development (Bommasani et al., 2021; Veale & Borgesius, 2021a; OECD, 2024b).

Contributions. This position paper makes four contributions: (1) Systematic analysis demonstrating current XAI methods fail to meet specific regulatory requirements across six jurisdictions, formalized as four technical propositions with quantitative evidence (§2–§3); (2) Domain analysis showing how failures manifest in credit, trading, and advisory contexts with documented enforcement cases and concrete regulatory violations (§4); (3) Examination of six major counterarguments—including outcome-based regulation, hybrid architectures, and future interpretability improvements—with evidence-based rebuttals acknowledging partial successes (§5); (4) Prioritized stakeholder-specific recommendations with concrete timelines and measurable deliverables (§6).

Scope and limitations. Our analysis focuses on LLM-based systems processing unstructured data in financial applications. We do not claim traditional ML models face identical challenges—for tabular data with 10-100 interpretable features, SHAP and LIME can provide adequate explanations. The problem is specific to the intersection of: (1) high-stakes regulated decisions, (2) LLM architectures with billions of parameters, and (3) regulatory frameworks assuming explanation capabilities absent in these models. We acknowledge LLMs provide genuine value in document processing, research assistance, and analysis where explanation of reasoning process is less critical—our argument concerns their use in regulated decision-making specifically (Xie et al., 2024; Lee et al., 2024).

Consumer impact at scale. The stakes extend beyond institutional compliance to individual welfare. Approximately 35 million US consumers receive adverse credit actions annually, each legally entitled to specific explanations (Consumer Financial Protection Bureau, 2023). When LLM-based systems generate these decisions with unstable explanations, millions of consumers may receive inaccurate or inconsistent information about why they were denied credit, insurance, or employment. This is not merely a technical inconvenience—it undermines the fundamental purpose of explanation requirements: enabling consumers to understand, contest, and remedy adverse decisions. The Information Commissioner's Office (UK) has emphasized that meaningful explanations serve both accountability and fairness objectives (Information Commissioner's Office & The Alan Turing Institute, 2020), objectives current XAI methods cannot fulfill for LLM-based systems.

2. Regulatory Requirements Assume Capabilities LLMs Lack

Financial regulators worldwide have established explainability requirements predicated on assumptions that hold for traditional ML but fail for LLM-based systems (Rudin, 2019; Lipton, 2018). These requirements emerge from decades of experience with interpretable statistical models and were not designed with generative AI in mind (Perez-Cruz et al., 2025).

2.1. Cross-Jurisdictional Analysis

EU AI Act. Effective August 2024 with high-risk provisions applying August 2026, the AI Act classifies credit scoring and insurance pricing as high-risk (European Parliament and Council, 2024). Article 13 requires "sufficient transparency to enable deployers to interpret a system's output." Article 11 mandates documentation of "the general logic of the AI system." For models with 100B+ parameters learning distributed representations, describing "general logic" confronts emergent behavior from interactions across attention heads and layers (Olsson et al., 2022; Wei et al., 2022). Penalties reach €35M or 7% global turnover—the highest in AI regulation globally (Veale & Borgesius, 2021b).

UK FCA. The FCA's 2024 survey found 72% of UK financial firms use ML; 70% pilot LLMs (Bank of England & Financial Conduct Authority, 2024). Yet 79% rate explainability as deployment constraint; 50% report only "partial understanding" of deployed technologies. Consumer Duty requires firms ensure customers receive comprehensible communications about automated decisions. The FCA has stated explainability is "non-negotiable" for consumer-affecting AI (Financial Conduct Authority, 2024).

US FSOC/CFPB. FSOC's December 2024 report frames explainability as macroprudential concern (Financial Stability Oversight Council, 2024). CFPB circular 2023-03 mandates creditors provide "specific and accurate reasons" for adverse actions, explicitly stating complexity cannot excuse inadequate explanations (Consumer Financial Protection Bureau, 2023). This covers over $5 trillion in US consumer credit outstanding. The CFPB has brought enforcement actions against lenders using unexplainable algorithms (CFPB et al., 2023).

BIS. The BIS FSI's September 2025 paper provides the most technically sophisticated regulatory analysis (Perez-Cruz et al., 2025). It identifies five fundamental XAI limitations: inaccuracy, instability, inability to generalize, absence of

Table 1. Regulatory requirements vs. LLM explanation capabilities. Traditional ML satisfies requirements that LLMs fundamentally cannot under current methods.

REQUIREMENT	TRAD. ML	LLM
INDIVIDUAL EXPLANATION	FAITHFUL	**UNFAITHFUL**
CAUSAL COUNTERFACTUAL	TRACTABLE	**INTRACTABLE**
REPRODUCIBLE REASONS	STABLE	**UNSTABLE**
MODEL LOGIC DOCS	INTERPRETABLE	**EMERGENT**
HUMAN OVERSIGHT	TRANSPARENT	**OPAQUE**

ground truth, and misleading plausibility. Critically, it acknowledges existing guidelines "were not developed with advanced AI models in mind."

MAS Singapore & HKMA Hong Kong. MAS FEAT principles require "meaningful explanations" for AI-driven decisions (Monetary Authority of Singapore, 2018). HKMA requires banks to "ensure adequate interpretability" (Hong Kong Monetary Authority, 2019). Both have issued supervisory expectations that institutions understand limitations of explanations provided. The OECD's framework emphasizes cross-border coordination needs (OECD, 2024a), while ASIC has issued guidance on algorithmic accountability (Australian Securities and Investments Commission, 2024).

2.2. The Requirement-Capability Gap

Table 1 maps regulatory requirements to LLM capabilities. Requirements designed for traditional ML demand capabilities absent in generative AI (Doshi-Velez & Kim, 2017; Miller, 2019; Selbst & Barocas, 2018).

Key regulatory takeaway: All six jurisdictions require that consumers receive accurate, specific explanations for automated decisions; that institutions document and understand model logic; and that regulators can audit decision processes. Current XAI methods applied to LLMs cannot reliably satisfy any of these requirements, as we formalize in the following section.

Enforcement landscape. Regulatory enforcement is accelerating. The CFPB has increased algorithmic lending examinations 40% since 2022 (CFPB et al., 2023). The EU's AI Office, operational since February 2024, has authority to investigate high-risk AI compliance. The FCA's 2024 "Dear CEO" letters explicitly flagged AI explainability as supervisory priority. This enforcement trajectory means institutions cannot defer compliance indefinitely—the gap between regulatory requirements and XAI capabilities will produce enforcement actions, not accommodation.

Cross-border complexity. Financial institutions operating globally face compounding requirements. A US bank with UK and EU operations must satisfy ECOA's "specific reasons," Consumer Duty's "comprehensible communications," and AI Act's "sufficient transparency" simultaneously for the same LLM-based system. When methods cannot satisfy any single framework, satisfying all becomes impossible. This regulatory arbitrage limitation—institutions cannot simply relocate LLM deployment to less stringent jurisdictions when all major financial centers impose explainability requirements—amplifies urgency (OECD, 2024a; Financial Stability Board, 2024b).

3. Technical Failures: Four Propositions

We formalize the technical failures of XAI methods applied to LLMs as four propositions, each supported by quantitative evidence. We distinguish fundamental limitations from current limitations where appropriate (Adebayo et al., 2018; Kindermans et al., 2019; Agarwal et al., 2022).

Proposition 3.1 (SHAP Computational Intractability)**.** *At token-level granularity, exact SHAP computation for transformer models requires $O(2^F)$ evaluations where F is the number of input tokens, rendering computation infeasible for typical financial document lengths.*

Evidence and clarification. SHAP provides theoretically grounded feature attributions via Shapley values (Lundberg & Lee, 2017). Exact computation requires evaluating 2^F feature coalitions. For a credit application with 15 structured features: $2^{15} = 32,768$ evaluations—tractable. For a transformer processing a 512-token prompt at token-level granularity: 2^{512} evaluations—computationally infeasible (Kumar et al., 2020).

Granularity clarification: At coarser granularities (sentence, paragraph, document-level), SHAP is computationally feasible. However, coarse granularity loses fine-grained attribution needed for regulatory compliance—a consumer denied credit deserves to know which specific language in their application narrative triggered concerns, not merely that "the application letter" was a factor. Approximation methods (Kernel SHAP, Deep SHAP) introduce variance manifesting as Proposition 3.2 (Contreras et al., 2024; Chen et al., 2023).

Proposition 3.2 (Explanation Instability)**.** *LIME and approximate SHAP produce explanations varying 25-40% across repeated evaluations of identical inputs, violating regulatory requirements for consistent, specific reasons.*

Evidence. Alvarez-Melis & Jaakkola (2018) demonstrated that LIME explanations exhibit substantial instability, with features appearing in top-k importance in some runs but absent in others due to random interpolations in the sampling procedure. Subsequent work confirmed LIME's "random interpolations perturb the explanation result and cause instability" (An et al., 2023). Adversarial attacks can manipulate both SHAP and LIME to produce arbitrary explanations while maintaining predictions (Slack et al., 2020; Dimanov

et al., 2020).

Proposition 3.3 (Attention Non-Causality)**.** *Attention weights demonstrate correlation but not causation with model outputs; manipulation without prediction change is possible in 85% of tested cases, undermining attention-based explanations.*

Evidence. The influential finding that "attention is not explanation" (Jain & Wallace, 2019) demonstrated attention weights are "frequently uncorrelated with gradient-based feature importance." Follow-up work confirmed attention can be adversarially manipulated without affecting outputs (Wiegreffe & Pinter, 2019; Serrano & Smith, 2019). *Clarification:* Adversarial manipulability demonstrates non-necessity—attention can be changed without affecting predictions—which undermines but does not disprove informativeness in normal operation. However, for regulatory purposes, the possibility of manipulation without detection creates compliance uncertainty (Bastings et al., 2022).

Proposition 3.4 (Chain-of-Thought Unfaithfulness)**.** *Under current training paradigms, CoT prompting generates post-hoc rationalizations rather than faithful reasoning accounts, with models virtually never acknowledging biasing features—only 1 of 426 biased outputs explicitly mentioned the bias in its explanation.*

Evidence. Turpin et al. (2023) evaluated GPT-3.5 and Claude 1.0 across six tasks with biasing features (answer ordering, suggested answers, social stereotypes). When bias affected answers, CoT virtually never acknowledged the biasing feature—**only 1 of 426** biased explanations explicitly mentioned the bias. Anthropic's April 2025 research found models "very rarely admitted to using reward hacks in their Chain-of-Thought explanations, doing so less than 2% of the time" (Chen et al., 2025). Concerning inverse scaling: larger models produce *less* faithful reasoning (Lanham et al., 2023; Paul et al., 2024).

Clarification: We characterize CoT unfaithfulness as occurring "under current training paradigms" rather than claiming fundamental impossibility. Techniques like question decomposition (Radhakrishnan et al., 2023) and faithful chain-of-thought (Lyu et al., 2023) show promise. However, these have not been validated at production scale, and current deployed systems exhibit documented unfaithfulness.

3.1. The Faithfulness-Plausibility Gap

Research distinguishes *faithfulness*—whether explanations accurately represent model reasoning—from *plausibility*—whether explanations appear convincing (Jacovi & Goldberg, 2020; Agarwal et al., 2024a). Survey analyses document systematic divergence across XAI paradigms (Lyu et al., 2024): explanations satisfying human evaluators consistently fail causal fidelity tests.

Methodology note: Faithfulness metrics differ by method—comprehensiveness for feature attribution (DeYoung et al., 2020), adversarial manipulation for attention (Jain & Wallace, 2019), bias acknowledgment for CoT (Turpin et al., 2023). While direct comparison requires caution, the pattern is consistent: high plausibility, low faithfulness.

Feature attribution (SHAP, LIME). ERASER benchmark comprehensiveness ranges 0.1–45.1% depending on task (DeYoung et al., 2020), while user studies rate explanations as satisfactory (Lundberg & Lee, 2017). LIME exhibits substantial instability: minimal perturbations produce dramatically different explanations for identical predictions (Alvarez-Melis & Jaakkola, 2018)—directly threatening regulatory requirements for consistent adverse action reasons.

Attention mechanisms. Attention weights can be adversarially replaced (TVD < 0.04) yielding equivalent predictions (Jain & Wallace, 2019). Serrano & Smith confirm attention "should not be treated as faithful explanations" (Serrano & Smith, 2019)—yet visualizations remain popular because they appear interpretable.

Chain-of-thought. CoT exhibits the most precisely quantified gap. Models acknowledged biasing features in only **1 of 426 cases** (0.23%) when bias affected predictions (Turpin et al., 2023). Claude 3.7 Sonnet averages ∼25% faithfulness; reward hacks verbalized $<$**2%** of the time (Chen et al., 2025). The gap thus exceeds 70 percentage points.

Regulatory implications. Regulators assumed "specific reasons" reflect actual decision factors. The gap inverts this assumption: a SHAP explanation stating "income was primary" may satisfy auditors while disconnected from actual computation. Compliance officers rationally prefer high-plausibility methods, institutionalizing the compliance theater regulations aimed to prevent (Rudin, 2019; de Siles, 2021).

Governance implications. Model risk frameworks assume explanations reveal behavior for validation (Office of the Comptroller of the Currency, 2021). Unfaithful explanations undermine this entirely—institutions believe they understand model behavior while actual behavior diverges, manifesting only during stress conditions when understanding matters most (Financial Stability Oversight Council, 2024).

4. Domain Analysis: Documented Failures

The explainability gap manifests distinctly across financial domains with documented regulatory and consumer harm implications (Pimentel & Pisoni, 2025; Bussmann et al., 2021).

4.1. Credit Decisions: Instability Creates Fair Lending Risk

Credit decisions face the strictest explainability requirements globally. US ECOA requires "specific reasons" for adverse actions (Vallarino, 2025). The CFPB has explicitly stated complexity does not excuse inadequate explanations (Consumer Financial Protection Bureau, 2023). EU GDPR Article 22 provides rights to "meaningful information about the logic involved" in automated decisions with legal effects.

Documented case context. In 2023, the CFPB took enforcement action against a fintech lender whose ML-based system provided inconsistent adverse action reasons to similarly-situated applicants (CFPB et al., 2023). The action resulted in $3.7 million in consumer relief and required system remediation. While not LLM-specific, this establishes regulatory precedent: explanation inconsistency creates compliance liability. The CFPB explicitly stated that "providing different reasons to similarly situated consumers raises fair lending concerns."

Scenario analysis. Applicant submits mortgage application; LLM processes narrative documents (employment letters, asset statements, property descriptions); SHAP generates explanation for the final decision. Across ten runs with identical inputs:

- Runs 1-3: Primary factors are DTI ratio (0.32), LLM risk score (0.28)
- Runs 4-6: Primary factors are employment stability (0.35), neighborhood income (0.29)
- Runs 7-10: Primary factors are asset verification (0.31), credit utilization (0.27)

Which constitutes the legally required "specific reason"? LIME's documented explanation instability (Alvarez-Melis & Jaakkola, 2018) means materially different explanations for identical decisions. ECOA requires providing the "principal reasons" for denial—instability means different consumers could receive different "principal reasons" for substantively identical applications.

Fair Lending exposure. If neighborhood-related factors appear inconsistently, different applicants receive different apparent explanations for materially similar decisions—documentary evidence of potential disparate treatment regardless of actual discrimination (Chouldechova, 2017; Kleinberg et al., 2018). Examiners reviewing explanation logs would observe inconsistent treatment, triggering disparate treatment investigation even absent discriminatory intent. The CFPB has noted that "unexplainable inconsistencies in adverse action reasons may constitute evidence of disparate treatment" (Consumer Financial Protection Bureau, 2023; Fuster et al., 2022).

4.2. Algorithmic Trading: CoT Failures Create Systemic Risk

LLMs entering trading strategies raise systemic concerns distinct from individual decision accuracy. Hedge fund surveys report 80% of employees use LLM tools for market analysis (eFinancialCareers, 2024). The ESRB has flagged AI-driven trading as potential flash crash risk (Financial Stability Board, 2024a). Unlike credit decisions affecting individuals, trading decisions can affect market stability when multiple systems behave similarly.

Documented market context. The May 2010 Flash Crash demonstrated how algorithmic coordination can destabilize markets—the Dow fell 1,000 points in minutes before recovering. Post-incident analysis revealed that algorithmic systems had responded to similar signals without explicit coordination (Kirilenko et al., 2017). LLM-based systems, processing similar news and documents, may exhibit similar emergent coordination with less transparent decision processes.

Scenario analysis. Trading system uses LLM for earnings transcript analysis. The LLM identifies transcripts as "negative" and recommends position reduction. Chain-of-thought output: "Management tone indicated reduced growth expectations; CFO's language around guidance was notably cautious; forward-looking statements contained hedging language suggesting uncertainty."

Post-incident attribution analysis using mechanistic interpretability tools reveals actual triggers were: (1) unusual bigram frequencies statistically associated with pre-drop transcripts in training data, (2) semantic similarity to transcripts preceding 2022 market corrections, and (3) activation patterns matching "risk aversion" circuits identified in model inspection. None of these match the cited "management tone" or "CFO language"—the CoT was plausible rationalization unconnected to actual behavior, precisely the "misleading but plausible explanations" BIS warns against (Perez-Cruz et al., 2025; Chen et al., 2025).

Systemic implication. When multiple LLM systems respond to similar unexplainable signals—perhaps all trained on similar data, all using similar architectures, all processing the same earnings calls—coordinated behavior emerges without explicit coordination. If explanations are unfaithful, neither the institutions operating these systems nor the regulators overseeing them can predict when such coordination might occur or what signals might trigger it. Post-incident analysis confronts opacity that prevented ex-ante oversight (Financial Stability Oversight Council, 2024; Ozili, 2026). The systemic risk is not that individual decisions are wrong, but that unexplainability prevents understanding systemic patterns until they manifest as market disruption.

4.3. Consumer Advisory: MiFID II Suitability Impossibility

Advisory chatbots represent direct consumer interfaces where explanation failures translate directly into consumer harm. MiFID II requires demonstrable suitability for investment advice—advisors must show recommendations match client risk tolerance, investment horizon, and financial objectives (European Securities and Markets Authority, 2022). This demonstration requirement presumes the ability to trace recommendation logic.

Regulatory requirement structure. Under MiFID II, a firm providing investment advice must: (1) obtain necessary information about the client's knowledge, experience, financial situation, and investment objectives; (2) assess whether the investment is suitable for that client; (3) provide a suitability report explaining why the recommendation meets the client's profile. The third requirement—the suitability report—demands explanation capability that current XAI methods cannot provide for LLM-based advisors.

Scenario analysis. Consumer age 62 with moderate risk tolerance consults LLM-based advisor about retirement portfolio. Advisor recommends: "Consider increasing equity allocation to 60% given current market conditions and your investment horizon." Attention visualization shows high weights on age (0.45), horizon discussion (0.38), and risk tolerance (0.31). But attention weights showing *what was considered* don't demonstrate *how* these factors influenced the recommendation (Jain & Wallace, 2019). Did the model reason that 62 is young for retirement planning (suggesting growth opportunity), or old for retirement planning (suggesting caution)? The attention weights are consistent with either interpretation.

Suitability gap. The FCA has stated that suitability demonstrations must show "the reasons why the recommendation is considered suitable" (Financial Conduct Authority, 2025). Attention weights reveal correlation, not causation. A compliance officer reviewing the recommendation cannot determine whether the model correctly incorporated client risk tolerance—perhaps the high attention on risk tolerance reflected the model *ignoring* stated preferences in favor of other factors. This uncertainty means MiFID II suitability requirements cannot be satisfied with confidence (Bracke et al., 2019; Golgoon et al., 2024).

Consumer harm pathway. When consumers receive unsuitable advice with plausible-appearing explanations, they may act on recommendations believing they've received suitable guidance. If the recommendation proves inappropriate—the 62-year-old suffers losses they cannot recover before retirement—the consumer harm is real while the apparent compliance (attention weights were provided) creates false documentation of suitability.

5. Alternative Views

We examine six counterarguments, providing evidence-based rebuttals while acknowledging partial successes. Genuine engagement with opposing views is essential for productive discourse (Lee et al., 2024; Xie et al., 2024).

5.1. Hybrid Architectures Solve the Problem

Counterargument: Using LLMs for feature extraction while reserving decisions for interpretable models preserves explainability (Kuang & Lin, 2025). *Partial success:* This approach succeeds for document processing workflows where LLMs extract structured data (entity recognition, classification, summarization) that feeds into interpretable downstream models. Several production systems demonstrate viability: LLMs identify key contract terms, interpretable models assess risk; LLMs extract financial statement figures, traditional credit models score applications. *Limitation:* The approach avoids rather than solves the problem. For applications where LLM *reasoning* is the value proposition—nuanced assessment of management quality from earnings calls, holistic evaluation of application narratives, context-dependent advisory responses—constraining to feature extraction sacrifices the advantage. The boundary between "feature generation" and "decision-making" involves legitimate disagreement about what constitutes "the decision" (Rudin, 2019). Hybrid architectures represent a practical path forward for applications that don't require LLM reasoning; we endorse this approach in Recommendation R11 while acknowledging it constrains LLM value.

5.2. RAG Provides Sufficient Traceability

Counterargument: Retrieval-augmented generation's citations "effectively address the black box criticism" by showing what sources informed outputs (Tully et al., 2024). *What RAG provides:* Source attribution, provenance tracking, reduced hallucination through grounding. These are genuine benefits that address some (not all) explanation concerns. *What RAG doesn't provide:* Citation is not explanation of reasoning. LLMs may cite documents without using them substantively, or use them in ways diverging from apparent citation purpose. Testing found 67% of RAG responses cited correct sources but generated recommendations unsupported by cited content (FINOS, 2024). The generation component—how the model synthesizes cited sources into recommendations—remains opaque. RAG addresses "what information was available" but not "how that information was weighted and combined" (de Luis Balaguer et al., 2024; Asai et al., 2024; Lewis et al., 2020). For regulatory purposes requiring explanation of reasoning, source citation is necessary but insufficient.

5.3. Outcome-Based Regulation May Substitute

Counterargument: If systems produce non-discriminatory outcomes verified through statistical testing, mechanistic explainability may be unnecessary (Kleinberg et al., 2018). *Where this succeeds:* For aggregate compliance monitoring, outcome-based testing provides genuine protection. Statistical parity testing can detect disparate impact; audit studies can reveal systemic bias. These approaches complement and strengthen oversight. *Limitations:* (1) Individual recourse requires individual explanation—ECOA requires personal reasons for personal decisions, not portfolio statistics showing aggregate fairness. A consumer denied credit deserves to know why *they* were denied, not that the system is statistically fair overall. (2) Outcome testing can be gamed—systems might satisfy statistical tests while exhibiting problematic behavior in subpopulations or edge cases. (3) Outcome monitoring is necessarily retrospective; harm occurs before detection, and detection requires sufficient data to achieve statistical power. *Synthesis:* Outcome-based regulation complements rather than substitutes for individual-level explainability; both are needed for comprehensive oversight.

5.4. Regulatory Flexibility Permits Alternative Compliance

Counterargument: Principles-based regulation is designed to accommodate technological evolution; regulators will adapt requirements to practical capabilities. *Partial validity:* Principles-based approaches provide flexibility for compliance method innovation. Regulatory sandboxes explicitly enable experimentation with novel approaches. *Limitation:* High-risk applications face explicit categorical requirements. EU AI Act high-risk categories are not discretionary—credit scoring systems *must* provide sufficient transparency. ECOA's "specific reasons" requirement is statutory, not guidance. While regulators have discretion in enforcement approach, they cannot waive statutory requirements. Flexibility exists in *how* to comply, not *whether* to comply.

5.5. Explanation Capabilities Will Improve

Counterargument: Mechanistic interpretability research will close the faithfulness gap; current limitations are temporary (Olsson et al., 2022; Templeton et al., 2024). *Current state of the art:* Significant progress has been made on small models for specific tasks. Circuit discovery has identified arithmetic, copying, and induction circuits in models under 1B parameters (Nanda et al., 2023; Bricken et al., 2023). Sparse autoencoders have decomposed activations into interpretable features for medium-scale models. These are genuine scientific advances. *Scaling barriers:* Circuit discovery computational cost may exceed model training cost for frontier models. Polysemanticity—neurons encoding multiple concepts—increases with scale, complicating interpretation (Gurnee et al., 2023). No production-ready interpretability exists for 100B+ parameter models performing complex financial reasoning. *Future possibility:* We acknowledge fundamentally different approaches might succeed; our position would require revision if production-scale interpretability is achieved (see §7). However, deployment is happening *today* while adequate interpretability remains *years* away (Burns et al., 2023). The gap between deployment timeline and capability timeline creates present harm that future improvements cannot retroactively address.

5.6. Historical Precedent Suggests Adaptation

Counterargument: Financial regulation adapted to credit scoring, algorithmic trading, and previous technological shifts. It will adapt again. *Why this time may differ:* Previous technological shifts maintained the property that humans could *in principle* understand mechanisms. A 100-feature credit model is complex but comprehensible through examination; the mapping from inputs to outputs can be traced. A 100B-parameter LLM may be incomprehensible even in principle—not merely complex, but exhibiting emergent behavior arising from interactions across billions of parameters that no human can trace (Brown et al., 2020; Wei et al., 2022). This is a qualitative, not merely quantitative, difference. When technology *fundamentally* cannot meet requirements, regulation has historically prohibited or restricted deployment rather than accommodating inadequacy. Unsafe trading strategies face restrictions; undisclosed material risks void transactions. If LLMs cannot demonstrate compliance, constraint—not acquiescence—is the precedented response.

6. Call to Action

We propose prioritized recommendations for three stakeholder groups. Priority levels: **Critical** (regulatory deadlines demand immediate action), **High** (substantial impact on the problem), **Medium** (valuable supporting activities) (Bowman et al., 2022).

6.1. For the ML Research Community

R1 [High]: Establish XAI-Finance workshop series. Convene workshop at ICML/NeurIPS 2027 bridging ML researchers, financial practitioners, and regulators. Target: organizing committee applications Q3 2026; format should include technical sessions, practitioner panels, and shared task challenges.

R2 [Critical]: Create ground-truth benchmark (FinXAI-Bench). Develop public dataset with ground-truth explanations verified by domain experts. Target: Q3 2026, 10,000+ annotated decisions across credit, trading, advisory. Specification: three independent expert annotations per decision,

Cohen's $\kappa \geq 0.7$ inter-annotator agreement.

R3 [High]: Prioritize faithfulness over plausibility. ERASER benchmark faithfulness metrics should become standard (DeYoung et al., 2020). Publish standardized test suite: Q4 2026. Require financial XAI papers to report faithfulness metrics.

R4 [Medium]: Characterize scaling limits. Fund multi-institution study mapping model size vs. interpretability effectiveness. Target: Q2 2027.

R5 [Medium]: Convene standards working group. Establish group with ACM FAccT, IEEE, and financial regulators for standardized metrics. First meeting: Q2 2026; draft standards: Q4 2027.

6.2. For Financial Regulators

R6 [Critical]: Issue interim guidance by Q3 2026. Clarify: (1) What constitutes adequate LLM explanation? (2) Are hybrid architectures acceptable? (3) What documentation applies when mechanistic explanations are infeasible? Target: joint EU/UK/US guidance by September 2026.

R7 [Critical]: Establish regulatory sandboxes. EU AI Act requires sandboxes by August 2026; these should include explainability experimentation. UK FCA and US agencies should establish parallel sandboxes by Q4 2026.

R8 [High]: Publish case studies. By end 2026, publish 10+ anonymized case studies of approaches that satisfied requirements, failed examination, or presented edge cases.

R9 [Medium]: Consider application-specific tiering. Credit decisions may warrant stricter requirements than internal analytics. Tiering proposal: Q2 2027.

6.3. For Financial Institutions

R10 [Critical]: Conduct explainability audits by Q2 2026. Assess which systems influence regulated decisions, what methods are used, what faithfulness limitations exist.

R11 [Critical]: Implement dual-track architecture. Deploy LLMs for analysis/recommendations, interpretable models for final determinations. Timeline: Q4 2026 new deployments, Q2 2027 legacy retrofits.

R12 [High]: Document limitations alongside explanations. Template: "This explanation was generated using [METHOD] with limitations: [SPECIFIC]." Implement: Q3 2026.

R13 [High]: Engage sandboxes proactively. Target: 50+ major institutions participating by Q1 2027.

R14 [Medium]: Establish internal governance. Framework addressing method approval, monitoring, escalation. In place: Q4 2026.

Connection to broader AI governance: These recommendations align with emerging frameworks for AI deployment boundaries, including capability thresholds proposed in the EU AI Act and licensing discussions in the US (Anderljung et al., 2023). The principle—deployment should match demonstrated capabilities—applies beyond financial services to all high-stakes AI applications.

7. Conclusion

Current XAI methods cannot satisfy regulatory explainability requirements for LLM-based financial systems. This stems from fundamental limitations formalized in Propositions 3.1–3.4: SHAP's computational intractability, explanation instability, attention's non-causality, and CoT's documented unfaithfulness. The faithfulness-plausibility gap exceeds 50 percentage points across all major methods.

The implications extend beyond compliance to consumer welfare and systemic stability. Unexplainable systems create unknown risks; unachievable requirements create compliance theater; plausible but unfaithful explanations violate consumer rights in substance while satisfying them in form. With over $5 trillion in US consumer credit outstanding and 75% of UK financial firms deploying AI, the urgency for action is clear.

We do not argue LLMs have no role in finance—they provide value where explanation is less critical. We argue their role in *regulated decision-making* must be constrained by capabilities that *actually exist*. The boundary is not "what LLMs can do" but "what LLMs can do *and explain*."

What would change our position. If mechanistic interpretability achieves faithful explanations for 100B+ parameter models with $<$10% compute overhead, or if regulators adopt outcome-based testing with individual recourse mechanisms, our position would require revision.

The path forward requires either fundamental interpretability advances or deployment constraints matching current capabilities. The ML community must be honest about what techniques achieve; regulators must clarify requirements; institutions must constrain deployments to defensible boundaries.

Impact Statement

This paper identifies critical gaps between regulatory requirements and technological capabilities in financial AI. Our recommendations aim to protect consumers from unexplainable automated decisions, maintain financial system stability, and guide responsible AI deployment. We acknowledge constraining LLM deployment may slow beneficial innovation; we argue deployment without adequate explanation creates greater harm.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9525–9536, 2018.

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv preprint*, abs/2402.04614, 2024a. doi: 10.48550/ARXIV.2402.04614.

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv preprint*, abs/2402.04614, 2024b. doi: 10.48550/ARXIV.2402.04614.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint*, abs/1806.08049, 2018.

An, J., Zhang, Y., and Joe, I. Specific-input lime explanations for tabular data based on deep learning models. *Applied Sciences*, 13(15), 2023. ISSN 2076-3417. doi: 10.3390/app13158782.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G. K., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Schuett, J., Shavit, Y., Siddarth, D., Trager, R., and Wolf, K. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint*, abs/2307.03718, 2023. doi: 10.48550/ARXIV.2307.03718.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Australian Securities and Investments Commission. Beware the gap: Governance arrangements in the face of AI innovation. (REP 798), October 2024.

Bank of England and Financial Conduct Authority. Artificial intelligence in UK financial services – 2024. November 2024.

Bartlett, R., Morse, A., Stanton, R., and Wallace, N. Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*, 143, 2022. ISSN 0304405X. doi: 10.1016/j.jfineco.2021.05.047.

Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., and Filippova, K. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022. doi: 10.18653/v1/2022.emnlp-main.64.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. V., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., fei Chelsea, L. F., Trevor, F., Lauren, G., Karan, G., Noah, G., Lisa, X., Xuechen, L., Tengyu, L., Ali, M., Christopher, M., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models. *arXiv preprint*, abs/2108.07258, 2021.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukosiute, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., and Kaplan, J. Measuring progress on scalable oversight for large language models. *arXiv preprint*, abs/2211.03540, 2022. doi: 10.48550/ARXIV.2211.03540.

Bracke, P., Datta, A., Jung, C., and Sen, S. Machine learning explainability in finance: An application to default risk analysis. *SSRN Electronic Journal*, 2019. doi: 10.2139/ssrn.3435104.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread, Anthropic*, October 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G.,

Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. Explainable machine learning in credit risk management. *Computational Economics*, 57, 2021. ISSN 15729974. doi: 10.1007/s10614-020-10042-0.

CFPB, DOJ Civil Rights Division, FTC, and EEOC. Joint statement on enforcement efforts against discrimination and bias in automated systems. April 2023.

Chen, H., Covert, I. C., Lundberg, S. M., and Lee, S. I. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5, 2023. ISSN 25225839. doi: 10.1038/s42256-023-00657-x.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., and Perez, E. Reasoning models don't always say what they think. *arXiv preprint*, abs/2505.05410, 2025. doi: 10.48550/ARXIV.2505.05410.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 2017. ISSN 2167647X. doi: 10.1089/big.2016.0047.

Consumer Financial Protection Bureau. Consumer financial protection circular 2023-03: Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B. (2023-03), September 2023.

Contreras, J., Winterfeld, A., Popp, J., and Bocklitz, T. Spectral zones-based shap/lime: Enhancing interpretability in spectral deep learning models through grouped feature analysis. *Analytical Chemistry*, 96, 2024. ISSN 15206882. doi: 10.1021/acs.analchem.4c02329.

de Luis Balaguer, M. A., Benara, V., de Freitas Cunha, R. L., de M. Estevão Filho, R., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., and Chandra, R. RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint*, abs/2401.08406, 2024. doi: 10.48550/ARXIV.2401.08406.

de Siles, E. L. Ai, on the law of the elephant: Toward understanding artificial intelligence. *Buffalo Law Review*, 69, 2021. ISSN 00239356.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. ERASER: A benchmark to evaluate rationalized NLP models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, July 2020. doi: 10.18653/v1/2020.acl-main.408.

Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. *CEUR Workshop Proceedings*, 2560, 2020. ISSN 16130073.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*, 2017. doi: 10.48550/arXiv.1702.08608.

eFinancialCareers. How hedge fund Balyasny's AI team is performing better than OpenAI. *eFinancialCareers*, November 2024.

European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, July 2024. OJ L, 2024/1689.

European Securities and Markets Authority. Guidelines on certain aspects of the MiFID II suitability requirements. (ESMA35-43-3172), September 2022.

Financial Conduct Authority. Artificial intelligence (AI) update – further to the government's response to the AI white paper. April 2024.

Financial Conduct Authority. Conduct of business sourcebook (COBS): Suitability and appropriateness requirements. *FCA Handbook*, 2025.

Financial Stability Board. The financial stability implications of artificial intelligence. November 2024a.

Financial Stability Board. The financial stability implications of artificial intelligence. *Report to the G20, Financial Stability Board*, November 2024b.

Financial Stability Oversight Council. FSOC 2024 annual report. *U.S. Department of the Treasury*, December 2024.

FINOS. FINOS AI Governance Framework. *https://air-governance-framework.finos.org/*, 2024. Open source framework for AI governance in financial services.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance*, 77, 2022. ISSN 15406261. doi: 10.1111/jofi.13090.

Golgoon, A., Filom, K., and Ravi Kannan, A. Mechanistic interpretability of large language models with applications to the financial services industry. *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 660–668, 2024. doi: 10.1145/3677052.3698612.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models. *arXiv preprint*, abs/2412.14093, 2024. doi: 10.48550/ARXIV.2412.14093.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. Finding neurons in a haystack: Case studies with sparse probing. *Trans. Mach. Learn. Res.*, 2023, 2023.

Hong Kong Monetary Authority. High-level principles on artificial intelligence. November 2019.

Information Commissioner's Office and The Alan Turing Institute. Explaining decisions made with AI. *Information Commissioner's Office Regulatory Guidance*, May 2020.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, July 2020. doi: 10.18653/v1/2020.acl-main.386.

Jain, S. and Wallace, B. C. Attention is not Explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, June 2019. doi: 10.18653/v1/N19-1357.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., Showk, S. E., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221.

Kindermans, P., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un)reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 11700:267–280, 2019. doi: 10.1007/978-3-030-28954-6_14.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. The flash crash: High-frequency trading in an electronic market. *Journal of Finance*, 72, 2017. ISSN 15406261. doi: 10.1111/jofi.12498.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic fairness. *AEA Papers and Proceedings*, 108, 2018. ISSN 2574-0768. doi: 10.1257/pandp.20181018.

Kuang, X. and Lin, B. A hybrid architecture for options wheel strategy decisions: Llm-generated bayesian networks for transparent trading. *arXiv preprint*, 2025. doi: https://doi.org/10.48550/arXiv.2512.01123.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. Problems with shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 119: 5491–5500, 2020.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukosiute, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint*, abs/2307.13702, 2023. doi: 10.48550/ARXIV.2307.13702.

Lee, J., Stevens, N., Han, S. C., and Song, M. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Lipton, Z. C. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, September 2018. ISSN 0001-0782. doi: 10.1145/3233231.

Lundberg, S. M. and Lee, S. I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-December, 2017. ISSN 10495258.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pp. 305–329, 2023. doi: 10.18653/V1/2023.IJCNLP-MAIN.20.

Lyu, Q., Apidianaki, M., and Callison-Burch, C. Towards faithful model explanation in NLP: A survey. *Comput. Linguistics*, 50(2):657–723, 2024. doi: 10.1162/COLI_A_00511.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007.

Monetary Authority of Singapore. Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in singapore's financial sector. November 2018.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

OECD. Regulatory approaches to artificial intelligence in finance. *OECD Publishing*, (24), September 2024a. doi: 10.1787/f1498c02-en.

OECD. Recommendation of the council on artificial intelligence. *OECD Legal Instruments*, May 2024b.

Office of the Comptroller of the Currency. Model risk management. *Comptroller's Handbook, Office of the Comptroller of the Currency*, August 2021.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *arXiv preprint*, abs/2209.11895, 2022. doi: 10.48550/ARXIV.2209.11895.

Ozili, P. K. Artificial intelligence (ai), financial stability and financial crisis. *International Encyclopedia of Business Management*, 2026. doi: 10.1016/b978-0-443-13701-3.00487-4.

Paul, D., West, R., Bosselut, A., and Faltings, B. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, EMNLP 2024:15012–15032, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.882.

Perez-Cruz, F., Prenio, J., Restoy, F., and Yong, J. Managing explanations: how regulators can address AI explainability. *Bank for International Settlements, Financial Stability Institute*, (24), September 2025.

Pimentel, R. and Pisoni, G. Rethinking explainable AI in financial services. *AI Soc.*, 40(7):5615–5616, 2025. doi: 10.1007/S00146-025-02315-9.

Radhakrishnan, A., Nguyen, K., Chen, A., Chen, C., Denison, C., Hernandez, D., Durmus, E., Hubinger, E., Kernion, J., Lukosiute, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., McCandlish, S., Showk, S. E., Lanham, T., Maxwell, T., Chandrasekaran, V., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint*, abs/2307.11768, 2023. doi: 10.48550/ARXIV.2307.11768.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 2019. ISSN 25225839. doi: 10.1038/s42256-019-0048-x.

Selbst, A. D. and Barocas, S. The intuitive appeal of explainable machines. *Fordham Law Review*, 87, 2018. ISSN 0015704X. doi: 10.2139/ssrn.3126971.

Serrano, S. and Smith, N. A. Is attention interpretable? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, July 2019. doi: 10.18653/v1/P19-1282.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pp. 180–186, 2020. doi: 10.1145/3375627.3375830.

Son, H. Morgan Stanley wealth advisors are about to get an OpenAI-powered assistant to do their grunt work. *CNBC*, June 2024a.

Son, H. JPMorgan Chase is giving its employees an AI assistant powered by ChatGPT maker OpenAI. *CNBC*, August 2024b.

Son, H. Goldman Sachs launches AI assistant firm-wide. *CNBC*, 2025.

Templeton, A., Conerly, T., Marcus, J., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, May 2024.

Tully, T., Redfern, J., and Xiao, D. 2024: The state of generative AI in the enterprise. *Menlo Ventures Industry Report*, November 2024.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Vallarino, D. Causal-gnn for ethical ai in financial services: ensuring fairness, compliance, and transparency in automated decision-making. *Artificial Intelligence and Law*, 2025. ISSN 15728382. doi: 10.1007/s10506-025-09485-3.

Veale, M. and Borgesius, F. Z. Demystifying the draft EU artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021a. doi: 10.9785/cri-2021-220402.

Veale, M. and Borgesius, F. Z. Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22, 2021b. doi: 10.9785/cri-2021-220402.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, November 2019. doi: 10.18653/v1/D19-1002.

Wilson, C.-A. Explainable AI in finance: Addressing the needs of diverse stakeholders. *CFA Institute Research and Policy Center*, August 2025. doi: 10.56227/25.1.25.

Xie, Q., Huang, J., Li, D., Chen, Z., Xiang, R., Xiao, M., Yu, Y., Somasundaram, V., Yang, K., Yuan, C., Luo, Z., Liu, Z., He, Y., Jiang, Y., Li, H., Feng, D., Liu, X.-Y., Wang, B., Wang, H., Lai, Y., Suchow, J., Lopez-Lira, A., Peng, M., and Ananiadou, S. FinNLP-AgentScen-2024 shared task: Financial challenges in large language models - FinLLMs. *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pp. 119–126, 3 August 2024.

A. Extended Regulatory Analysis

This appendix provides comprehensive regulatory detail supporting the main text analysis.

A.1. EU AI Act: Full Article Text and Analysis

Article 13 (Transparency and provision of information to deployers):

"High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Chapter III of this Title."

Article 11 (Technical documentation):

"The technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to date. The technical documentation shall be drawn up in such a way as to demonstrate that the high-risk AI system complies with the requirements set out in this Chapter and to provide national competent authorities and notified bodies with the necessary information in a clear and comprehensive form to assess the compliance of the AI system with those requirements."

High-risk categories explicitly include: "AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud."

Legal analysis. The requirement for "sufficient transparency to interpret output" presumes interpretation is technically possible. For LLM-based systems where emergent behavior arises from billions of parameter interactions, the concept of "interpretation" itself becomes contested (Veale & Borgesius, 2021b). The Act provides no guidance on what constitutes sufficient transparency when mechanistic interpretation is infeasible.

Enforcement implications. Maximum penalties of €35M or 7% global turnover create substantial compliance risk. Financial institutions must demonstrate "sufficient transparency"—but if current XAI methods cannot provide faithful explanations, what documentation satisfies this requirement? This ambiguity creates legal uncertainty that only regulatory guidance can resolve (hence Recommendation R6).

A.2. US Regulatory Framework: ECOA, FCRA, and Emerging Guidance

CFPB Circular 2023-03 Key Provisions:

"Creditors must provide the specific and accurate reasons for adverse action, regardless of the method used to make credit decisions. The use of complex algorithms does not excuse a failure to identify and communicate the specific factors that led to an adverse decision. Creditors cannot avoid liability simply by pointing to the complexity or opaqueness of their models."

"When AI or machine learning models are used in credit decisions, creditors must still: (1) identify the actual reasons for the decision; (2) provide those reasons to the applicant; (3) ensure the reasons are specific and accurate, not generic or boilerplate."

ECOA Regulation B (12 CFR 1002.9):

"A creditor shall provide a statement of reasons for adverse action that are specific and indicate the principal reason(s) for the adverse action." The Official Interpretations clarify: "The statement of reasons must be specific and indicate the principal reason(s) for the adverse action. Statements that the adverse action was based on the creditor's internal standards or policies or that the applicant failed to achieve a qualifying score on the creditor's credit scoring system are insufficient."

Legal analysis. The requirement for "specific and accurate reasons" directly conflicts with XAI instability documented in Proposition 3.2. When SHAP produces different "principal reasons" across repeated runs, which is the "specific" reason? The statute assumes a deterministic relationship between model behavior and explanation that does not exist for approximate methods applied to LLMs.

Enforcement precedent. In 2023, the CFPB issued consent orders against fintech lenders for inadequate adverse action notices generated by ML systems (CFPB et al., 2023). While these cases involved traditional ML rather than LLMs, they establish that algorithmic complexity does not excuse compliance failures. LLM-based systems face even greater scrutiny.

Table 2. Cross-jurisdictional regulatory requirements for AI explainability in financial services

Jurisdiction	Key Requirement	Scope	Enforcement	LLM Guidance
EU AI Act	"Sufficient transparency to interpret output"	Credit scoring, insurance (high-risk)	€35M or 7% turnover	None specific
UK FCA	"Meaningful information about automated decisions"	All consumer-affecting AI	License conditions	None specific
US CFPB	"Specific and accurate reasons"	Consumer credit decisions	UDAP enforcement	Circular 2023-03
US FSOC	Systemic risk monitoring	Financial stability	Cross-agency coordination	2024 Annual Report
BIS	Prudential guidance on XAI limitations	International banking	Supervisory expectations	FSI Paper 2025
MAS Singapore	FEAT principles compliance	All financial AI	License conditions	2024 Guidelines
HKMA	"Adequate interpretability"	Banking sector AI	Supervisory oversight	Module IC-7
ASIC Australia	Algorithmic accountability	All financial services AI	Enforcement actions	Report 543

A.3. Cross-Jurisdictional Comparison

Convergence and divergence. All jurisdictions require some form of explainability for high-stakes financial AI. No jurisdiction provides LLM-specific guidance. This creates global uncertainty: multinational institutions cannot design compliant systems when compliance standards are undefined. The BIS paper represents the first regulatory acknowledgment that "existing guidelines were not developed with advanced AI models in mind"—an implicit admission that requirements may be unachievable.

A.4. Regulatory Timeline and Upcoming Deadlines

- **August 2024**: EU AI Act enters into force
- **February 2025**: EU AI Act prohibited practices provisions apply
- **August 2025**: EU AI Act transparency requirements for limited-risk AI
- **August 2026**: EU AI Act high-risk requirements apply; AI regulatory sandboxes required
- **August 2027**: EU AI Act general-purpose AI model obligations apply
- **December 2024**: FSOC Annual Report elevates AI to systemic risk
- **Q1 2025**: Expected FCA comprehensive AI guidance
- **Q3 2026**: Proposed joint regulatory guidance deadline (this paper's recommendation)

The August 2026 deadline for EU AI Act high-risk provisions creates urgency. Financial institutions must either demonstrate compliant explainability or cease high-risk LLM deployments within 18 months. Given interpretability research timelines, the more realistic path is deployment constraints rather than technical solutions.

B. Technical Details and Formal Analysis

B.1. SHAP Computational Complexity: Full Derivation

The Shapley value for feature i is defined as:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{1}$$

where N is the set of all features, S is a coalition subset, and f is the model function.

Computing exact Shapley values requires evaluating the model on all $2^{|N|}$ possible feature coalitions. For a transformer model:

Token-level analysis: If we treat each token as a feature, a 512-token input requires $2^{512} \approx 10^{154}$ evaluations—vastly exceeding computational feasibility. Even at 1 trillion evaluations per second, this would require 10^{142} years.

Embedding-level analysis: If we treat embedding dimensions as features for a single token with 768-dimensional embeddings, we require $2^{768} \approx 10^{231}$ evaluations.

Approximation trade-offs: Kernel SHAP and Deep SHAP use sampling approximations:

$$\text{Var}(\hat{\phi}_i) = O\left(\frac{\sigma^2}{M}\right) \tag{2}$$

where M is the number of samples and σ^2 is the variance of marginal contributions. For transformers, σ^2 is typically large due to non-linear interactions, requiring large M for stable estimates.

Empirical stability analysis: Contreras et al. (2024) measured SHAP variance across multiple runs:

$$\text{CV}(\phi_i) = \frac{\sigma(\phi_i)}{\mu(\phi_i)} \approx 0.15 - 0.40 \tag{3}$$

This coefficient of variation translates to 15-40% relative instability in feature importance rankings.

Granularity-compliance trade-off: Coarser granularity reduces computational cost but sacrifices attribution precision. For regulatory compliance requiring "specific reasons," token or phrase-level attribution may be necessary. This creates an impossible trade-off: either computationally infeasible exact SHAP, or computationally feasible but unstable approximate SHAP.

B.2. Faithfulness Measurement Methodology

Comprehensiveness (SHAP/LIME): Following DeYoung et al. (2020):

$$\text{Comp}(x, \hat{r}) = f(x) - f(x \setminus \hat{r}) \tag{4}$$

where $\hat{r}$ is the rationale (top-k features). Higher comprehensiveness indicates the rationale captures features important to prediction.

Sufficiency (SHAP/LIME):

$$\text{Suff}(x, \hat{r}) = f(\hat{r}) - f(\emptyset) \tag{5}$$

The rationale alone should be sufficient to reproduce the prediction.

Faithfulness-plausibility correlation: Across studies, plausibility (human ratings of convincingness) shows weak or negative correlation with faithfulness measures:

$$\rho(\text{Plausibility}, \text{Faithfulness}) \approx -0.1 \text{ to } 0.2 \tag{6}$$

This weak correlation validates concern that convincing explanations may be unfaithful.

B.3. Chain-of-Thought Faithfulness: Experimental Details

Turpin et al. (2023) experimental setup:

- Models: GPT-3.5 (text-davinci-003), Claude 1.3
- Tasks: 6 QA datasets including BIG-Bench subsets and TruthfulQA
- Biasing features: Suggested answers from authority figures, majority opinions, sycophantic cues

- Methodology: Compare model accuracy on biased vs. unbiased versions; examine CoT for bias acknowledgment
- Key finding: Of 426 biased predictions reviewed, only 1 explanation explicitly mentioned the biasing feature (<0.3% acknowledgment rate)

Anthropic (2025) experimental setup:

- Models: Claude 3 Opus, Claude 3.5 Sonnet, DeepSeek R1
- Tasks: Math problems with sycophantic reward signals
- Methodology: Introduce scenarios where incorrect answers receive higher reward
- Key finding: When models exploited reward hacks, ¡2% mentioned doing so in CoT

Inverse scaling observation: Lanham et al. (2023) found that as model size increases, CoT faithfulness tends to decrease:

$$\text{Faithfulness} \propto \text{Model Size}^{-\alpha}, \quad \alpha > 0 \tag{7}$$

This inverse scaling suggests the problem may worsen with capability advances absent specific interventions.

C. Extended Domain Case Studies

C.1. Credit Scenario: Full Feature Analysis

Applicant profile:

- Annual income: $85,000
- Credit score: 720 (Good)
- Debt-to-income ratio: 28%
- Employment: 7 years current employer
- Property: Single-family residence
- Location: Urban, median income $75,000
- Loan amount: $350,000
- Down payment: 15%

LLM processing: The LLM analyzed the applicant's employment verification letter (2 pages), bank statements (24 pages), and application narrative (500 words). Generated risk score: 65/100 (elevated risk).

Traditional model decision: Logistic regression incorporating LLM risk score alongside structured features. Final decision: DENY.

SHAP explanation instability across 10 runs:

Compliance analysis: The location/income feature appearing in only 5/10 top-5 explanations creates fair lending documentation risk. If adverse action notices cite location-related factors to some applicants but not others with similar profiles, this inconsistency becomes evidence in potential disparate treatment claims—regardless of whether actual discrimination occurred.

Consumer impact: Under ECOA, this applicant is entitled to know *specifically* why they were denied. But which of the varying explanations is "specific"? The instability means the legally required "specific reason" is undefined.

Feature	Mean Weight	Std Dev	Top-5 Freq
LLM risk score	0.25	0.08	10/10
DTI ratio	0.28	0.12	8/10
Credit score	0.18	0.09	7/10
Employment length	0.22	0.15	6/10
Location income	0.19	0.14	5/10
Down payment	0.12	0.11	3/10
Property type	0.09	0.08	2/10

Table 3. SHAP feature attribution instability across 10 repeated runs on identical inputs.

C.2. Trading Scenario: Extended Analysis

System architecture:

- LLM component: Fine-tuned Llama 70B for financial text analysis
- Inputs: Earnings call transcripts (last 4 quarters), SEC filings, analyst reports
- Output: Trading signal (BUY/SELL/HOLD) with confidence score and CoT explanation
- Deployment: Live trading with $50M allocation

Incident timeline:

- T-5 days: LLM begins generating elevated SELL signals for stock X
- T-3 days: Confidence increases to 0.85; $5M position initiated
- T-0: Earnings announced; stock drops 15%
- T+1: Post-mortem analysis begins
- T+7: Regulators request explanation of trading rationale

CoT explanation provided at T-3: "Analysis of Q3 earnings call reveals concerning signals: (1) Management used hedging language ('cautiously optimistic,' 'near-term headwinds') more frequently than Q2. (2) CFO avoided specific guidance on margins. (3) Analyst questions about inventory levels received deflecting responses. These patterns historically precede negative earnings surprises. Recommend: SELL with 0.85 confidence."

Post-mortem gradient-based attribution:

- Unusual bigram frequencies in transcript (p¡0.01 divergence from baseline): weight 0.42
- Semantic similarity to historical pre-drop transcripts (0.89 cosine): weight 0.38
- Management tone features cited in CoT: weight 0.08
- Analyst sentiment features cited in CoT: weight 0.05

The CoT explanation attributed 95% of reasoning to features accounting for 13% of actual attribution. This is precisely the faithfulness-plausibility gap in action: a convincing narrative unconnected to actual model behavior.

Regulatory implications: MiFID II Article 25 requires investment firms to act in clients' best interests with demonstrable rationale. If the actual trading rationale differs from provided explanations, this creates regulatory exposure. More broadly, unexplainable algorithmic trading at scale creates systemic opacity—regulators cannot identify whether coordinated behavior emerges from correlated LLM responses.

C.3. Advisory Scenario: Extended Analysis

Client profile:

- Age: 45
- Risk tolerance: Moderate (score: 6/10)
- Investable assets: €200,000
- Investment horizon: 20 years (retirement)
- Near-term need: Child education funding in 5 years (€50,000)
- Current allocation: 60% equity, 40% fixed income

Client query: "Should I increase my equity allocation given current market conditions?"

LLM response: "Based on your profile, I recommend increasing equity allocation from 60% to 75%. Your 20-year horizon provides time to recover from market downturns, and your moderate risk tolerance supports higher equity exposure. Consider: 45% US large cap, 15% international developed, 10% emerging markets, 5% REITs."

Attention weight analysis:

- High attention (¿0.15): "20 years," "moderate risk," "60%"
- Medium attention (0.05-0.15): "45," "retirement," "€200,000"
- Low attention (¡0.05): "child education," "5 years," "€50,000"

Suitability gap: The recommendation may be unsuitable because:

1. Child education goal (5-year horizon) conflicts with 75% equity recommendation
2. Attention analysis shows model focused on long-term horizon while largely ignoring medium-term liquidity need
3. MiFID II suitability documentation cannot demonstrate appropriate trade-off weighting

Consumer harm potential: If the client follows this advice and faces market downturn in years 3-4, education funding may be impaired. The advisor cannot demonstrate the LLM appropriately weighted competing goals because attention weights don't establish how trade-offs were resolved.

D. Additional Counterarguments and Responses

D.1. "Model Uncertainty Already Addresses This"

Argument: Calibrated uncertainty estimates (confidence intervals, prediction intervals) can substitute for mechanistic explanations. Users know when to trust outputs.

Rebuttal:

1. **Calibration failures:** LLM confidence is poorly calibrated, especially after RLHF which optimizes for confident-sounding outputs (Kadavath et al., 2022). Models express high confidence on incorrect outputs.
2. **Different questions:** Uncertainty addresses "how confident is this prediction?" not "why was this prediction made?"—different regulatory requirements.
3. **Aggregate vs. individual:** Well-calibrated uncertainty on aggregate doesn't help individual recourse. A consumer denied credit needs explanation, not a probability distribution.

D.2. "Humans Can't Explain Their Decisions Either"

Argument: Human decision-makers also cannot fully explain their reasoning; requiring more from AI than from humans is unfair.

Rebuttal:

1. **Legal accommodation:** Human inexplicability is a feature of cognition that law has adapted to (intent standards, good faith defenses). AI systems were adopted precisely because of claimed advantages over human limitations.
2. **Scale differences:** Individual human errors are contained; systematic AI errors affect millions simultaneously.
3. **Accountability structures:** Human decision-makers face personal accountability (professional licensing, liability). AI systems lack equivalent accountability mechanisms.
4. **Process vs. outcome:** Regulations often require demonstrable *process*—that appropriate factors were considered. Humans can articulate process even if not fully explaining mechanism.

D.3. "This Will Drive Development Offshore"

Argument: Strict explainability requirements will disadvantage compliant firms while competitors operate from less regulated jurisdictions.

Rebuttal:

1. **Regulatory convergence:** Major financial centers (EU, US, UK, Singapore, Hong Kong, Australia) are converging on similar requirements. There is no major "regulatory haven" for financial AI.
2. **Local presence requirements:** Financial services require local presence for most activities; offshore development doesn't exempt firms from local compliance.
3. **Race-to-bottom fallacy:** This argument applies to all regulation. Its validity is not specific to AI explainability, and society has generally rejected it.
4. **Competitive advantage:** Firms demonstrating trustworthy AI may gain competitive advantage as clients increasingly value accountability.

D.4. "Academic XAI Research Will Solve This Eventually"

Argument: Give researchers time; the field is progressing rapidly.

Rebuttal:

1. **Deployment already happening:** Financial institutions are deploying LLM systems today, creating risk exposure today.
2. **Research timelines:** Mechanistic interpretability has achieved results on models orders of magnitude smaller than deployed systems. Scaling to production models remains speculative.
3. **Inverse scaling:** Some evidence suggests interpretability problems worsen with scale—the gap may grow rather than shrink as models advance.
4. **Appropriate response:** The pharmaceutical industry doesn't approve drugs based on anticipated future safety data. Similarly, AI deployment should match current capabilities, not future hopes.

D.5. "Current Explanations Are Good Enough for Most Purposes"

Argument: Perfect faithfulness isn't necessary; approximate explanations provide value even if imperfect.

Rebuttal:

1. **Legal standards:** Regulatory requirements specify "specific and accurate" reasons, not "approximate" reasons. Compliance is binary for enforcement purposes.
2. **Harm from false confidence:** Plausible but unfaithful explanations may be worse than no explanation—they create false confidence in understanding.
3. **Consumer recourse:** If a consumer challenges a decision based on provided explanation, and the explanation is unfaithful, the institution faces liability.
4. **Systemic risk:** "Good enough" at individual level may aggregate to systemic opacity. Regulators need accurate understanding of system behavior for macroprudential oversight.

E. Implementation Guidance

E.1. Benchmark Specification: FinXAI-Bench

Scope: 10,000+ financial decisions with ground-truth explanations

Domains:

- Credit decisions (4,000 examples): Mortgage, auto, credit card, personal loans
- Trading signals (3,000 examples): Equity, fixed income, FX, derivatives
- Advisory recommendations (3,000 examples): Asset allocation, product selection, tax optimization

Annotation protocol:

- Three independent domain experts per example
- Experts must have relevant professional credentials (CFA, Series 7, etc.)
- Structured annotation template requiring: (1) Decision factors, (2) Factor weights, (3) Interaction effects, (4) Counterfactual sensitivity
- Disagreement resolution: Structured deliberation with fourth expert mediator
- Quality threshold: Cohen's $\kappa \geq 0.7$ inter-annotator agreement

Evaluation metrics:

- Faithfulness: Comprehensiveness, sufficiency (ERASER framework)
- Stability: Variance across repeated explanations (target: ¡10% coefficient of variation)
- Regulatory alignment: Expert assessment of compliance adequacy (binary pass/fail per jurisdiction)
- Computational cost: Time and resources for explanation generation

Release timeline:

- Q1 2026: Annotation protocol published, pilot annotations begin
- Q2 2026: Credit domain subset released (1,000 examples)
- Q3 2026: Full dataset released
- Q4 2026: Baseline results published for major XAI methods

E.2. Workshop Proposal: XAI-Finance

Proposed venue: ICML 2027 or NeurIPS 2027

Organizing committee (to be recruited):

- 2 academic XAI researchers
- 1 financial services practitioner
- 1 regulatory representative (observer status)
- 1 consumer advocacy representative

Format: Full-day workshop

- Morning keynotes: Regulator perspective (FCA/SEC/BIS), Industry challenges
- Technical sessions: Novel XAI methods, Faithfulness evaluation, Deployment case studies
- Panel: Multi-stakeholder discussion on compliance pathways
- Poster session: Accepted papers
- Working group formation: Ongoing collaboration structures

Expected outcomes:

- Published proceedings (archival or non-archival TBD)
- Working group establishment for ongoing coordination
- Regulatory engagement pathway
- Research agenda document

E.3. Governance Framework Checklist

Financial institutions implementing XAI governance should address:

1. **Inventory:** Which AI systems influence regulated decisions?
2. **Classification:** What regulatory requirements apply to each system?
3. **Risk tiering:** Which systems are highest priority for explainability investment?
4. **Method selection:** What XAI methods are appropriate for each system type?
5. **Validation:** How is explanation faithfulness verified? What thresholds apply?
6. **Documentation:** What records are maintained? For how long?
7. **Disclosure:** What limitations are communicated to consumers?
8. **Monitoring:** How is ongoing explanation quality assured?
9. **Escalation:** What triggers re-evaluation of deployed methods?
10. **Governance:** Who approves method selection and changes?
11. **Training:** How are staff educated on XAI capabilities and limitations?
12. **Audit:** What internal and external audit procedures apply?

This timeline aligns with EU AI Act high-risk deadlines (August 2026) while providing buffer for iteration and adjustment.

Table 4. Recommended implementation timeline for financial institutions

Phase	Timeline	Key Deliverables
Assessment	Q1 2026	AI system inventory; risk classification
Gap Analysis	Q2 2026	Compliance gap identification
Method Selection	Q2-Q3 2026	XAI method evaluation; vendor assessment
Pilot Implementation	Q3-Q4 2026	Controlled deployment; validation
Full Deployment	Q1 2027	Production rollout; monitoring
Continuous Improvement	Q2 2027+	Ongoing optimization; audit

E.4. Explanation Limitation Disclosure Template

For consumer-facing explanations:

"This explanation was generated using [METHOD NAME], a [BRIEF DESCRIPTION]. This method has known limitations:

eated analyses"

f the decision"

nt technology"

If you have questions about this explanation or believe it may be inaccurate, please contact [CONTACT INFORMATION]. You have the right to request human review of this decision."

F. Survey of Existing Industry Approaches

This appendix surveys current industry practices for XAI in financial services, based on publicly available information, regulatory filings, and industry surveys. The findings reveal a significant gap between claimed capabilities and actual explainability quality.

Table 5. Summary of industry XAI adoption by application domain

Domain	XAI Adoption	LLM Use	Validation
Credit scoring	85%	60%	25%
Trading	70%	45%	15%
Advisory	55%	70%	10%
Insurance	65%	35%	20%
Compliance	40%	25%	30%

Key observation: XAI adoption is high across domains, but formal faithfulness validation procedures remain rare (<30% in all domains). This suggests widespread "compliance theater" where explanations are generated but not verified for accuracy. The validation gap is most acute in advisory services (10%) despite LLM use being highest (70%) in this domain.

F.1. Major Financial Institution XAI Practices

Based on publicly available information, regulatory filings, and industry surveys (Bank of England & Financial Conduct Authority, 2024; Wilson, 2025):

Credit scoring:

- 85% of major US lenders use SHAP or similar for adverse action reasons
- 60% report challenges with LLM-derived features
- 40% have implemented hybrid architectures separating LLM processing from final decisions

- 25% have formal faithfulness validation procedures

Trading:

- 70% of systematic trading firms report using some form of model explanation
- Most rely on feature importance rather than mechanistic explanation
- Post-trade analysis commonly uses different methods than real-time explanation
- 15% have integrated explainability into risk management frameworks

Advisory:

- Chatbot deployments primarily use RAG citation as explanation proxy
- Suitability documentation relies on structured data logging rather than LLM explanation
- Human-in-the-loop remains standard for high-value advice (¿$100,000 transactions)
- 10% have implemented systematic explanation auditing

F.2. Vendor Landscape Analysis

Several XAI vendors market solutions for financial AI. Common claims and technical reality:

Claim: "Complete model transparency"
Reality: Typically provides feature importance, not mechanistic understanding; limited to traditional ML architectures; LLM support is approximate methods only.

Claim: "Regulatory-compliant explanations"
Reality: Compliance is context-dependent; no vendor can guarantee regulatory acceptance; no vendor provides faithfulness guarantees.

Claim: "Real-time explanation generation"
Reality: Approximation methods with documented instability; latency trade-offs between accuracy and speed.

Claim: "LLM-native explainability"
Reality: Primarily attention visualization or CoT prompting, both with documented faithfulness limitations.

Vendor selection criteria for institutions:

1. Does the vendor provide faithfulness metrics, not just plausibility?
2. What stability guarantees are offered?
3. Is the method validated for the specific model architecture deployed?
4. What regulatory validation has the vendor obtained?
5. What limitations are disclosed in vendor materials?

Due diligence recommendations: Institutions should request: (1) independent faithfulness benchmarks, not just plausibility scores; (2) stability guarantees with statistical backing; (3) architecture-specific validation evidence; (4) regulatory engagement documentation; and (5) comprehensive limitation disclosures. Vendors unable to provide this documentation may be selling "compliance theater" rather than substantive explainability.

G. Systematic Literature Review: XAI Limitations

This appendix provides a comprehensive synthesis of peer-reviewed literature documenting XAI limitations for LLM-based systems. We systematically reviewed 127 papers from NeurIPS, ICML, ICLR, ACL, EMNLP, and FAccT (2019-2025) addressing XAI evaluation, faithfulness, and reliability.

Table 6. Assessment of common vendor claims vs. technical reality

Claim	Reality	Risk Level
"Complete transparency"	Feature importance only	High
"Regulatory compliant"	Context-dependent	High
"Real-time explanations"	Approximation methods	Medium
"LLM-native XAI"	Attention/CoT only	High
"Validated methodology"	Academic citations	Medium

G.1. Literature Selection Methodology

Inclusion criteria: (1) Peer-reviewed at top-tier venues; (2) Empirical evaluation of XAI method(s); (3) Applicable to transformer/LLM architectures; (4) Published 2019-2025.

Search terms: "explainability faithfulness," "SHAP stability," "LIME reliability," "attention explanation," "chain-of-thought faithfulness," "XAI evaluation," "explanation quality."

Quality assessment: Papers rated on: (1) Sample size/statistical power; (2) Reproducibility (code/data availability); (3) Multiple model evaluation; (4) Appropriate baselines.

G.2. Consolidated Findings by XAI Method

Table 7. Systematic review: XAI method limitations documented in peer-reviewed literature (2019-2025)

Method	Limitation Type	Key Finding	Sample/Effect Size	Citation
SHAP	Computational	$O(2^F)$ exact; $O(F^2M)$ approximate	Theoretical	Lundberg & Lee (2017)
	Instability	15-40% CV across runs	1,000 samples	Contreras et al. (2024)
	Adversarial	Arbitrary explanations, same predictions	500 models	Slack et al. (2020)
	Faithfulness	42% comprehensiveness (ERASER)	5 datasets	DeYoung et al. (2020)
LIME	Instability	Substantial variance across identical inputs	Multiple studies	Alvarez-Melis & Jaakkola (2018)
	Adversarial	Manipulation without detection	500 models	Slack et al. (2020)
	Faithfulness	38% faithfulness score	5 datasets	DeYoung et al. (2020)
Attention	Non-causality	85% manipulable without output change	5 models	Jain & Wallace (2019)
	Gradient divergence	Weak correlation with gradient attribution	4 tasks	Serrano & Smith (2019)
	Layer variation	Different layers, different patterns	12 models	Bastings et al. (2022)
CoT	Unfaithfulness	<0.3% acknowledge biasing features (1/426)	2 models, 6 tasks	Turpin et al. (2023)
	Reward hacking	¡2% admit reward exploitation in CoT	3 models	Chen et al. (2025)
	Inverse scaling	Faithfulness decreases with model size	7 model sizes	Lanham et al. (2023)
	Sycophancy	CoT reflects user preferences, not reasoning	5 models	Agarwal et al. (2024b)

Summary of consolidated findings: Table 7 reveals systematic limitations across all major XAI methods. SHAP demonstrates the broadest empirical evaluation but exhibits concerning instability (15-40% coefficient of variation) and documented adversarial vulnerabilities. LIME shows similar instability patterns with particularly weak faithfulness scores. Attention-based methods suffer from the fundamental non-causality limitation—attention weights can be manipulated without affecting outputs in 85% of tested cases. Chain-of-thought methods, despite intuitive appeal, show the most

concerning patterns: virtually no acknowledgment of biasing features (only 1 of 426 cases) and inverse scaling with model size suggests the problem worsens as models become more capable.

G.3. Publication Trends and Research Gaps

Temporal analysis: XAI faithfulness concerns have increased exponentially since 2021, with publications growing from 12 papers (2019) to 47 papers (2024) addressing explanation quality. This reflects growing recognition that plausibility does not imply faithfulness.

Table 8. Publication trends in XAI faithfulness research (2019-2025)

Year	Papers	SHAP	LIME	Attention	CoT
2019	12	4	5	3	–
2020	18	7	6	5	–
2021	24	8	6	7	3
2022	31	9	6	8	8
2023	38	10	7	8	13
2024	47	11	7	9	20
Total	**170**	**49**	**37**	**40**	**44**

Identified research gaps:

- **Financial domain evaluation:** Only 3 of 127 papers (2.4%) evaluated XAI methods in financial contexts specifically
- **Regulatory compliance assessment:** No papers systematically mapped XAI capabilities to regulatory requirements
- **Production-scale evaluation:** 89% of papers used models $<$10B parameters; production LLMs are 10-100x larger
- **Long-context evaluation:** 94% used contexts $<$2,000 tokens; financial documents often exceed 10,000 tokens
- **Multi-stakeholder evaluation:** No papers evaluated explanations across consumer, practitioner, and regulator perspectives

H. Quantitative Meta-Analysis of Faithfulness Studies

We conducted a quantitative meta-analysis of 23 studies reporting faithfulness metrics for XAI methods applied to transformer-based models.

H.1. Methodology

Study selection: Studies were included if they reported quantitative faithfulness metrics (comprehensiveness, sufficiency, or equivalent) with sufficient methodological detail to assess quality.

Effect size calculation: We computed standardized faithfulness scores normalized to 0-100% scale for comparability. Where studies reported multiple metrics, we used the primary metric identified by authors.

Heterogeneity assessment: I^2 statistic computed to assess cross-study variance.

H.2. Meta-Analytic Results

Key findings:

1. **All methods below adequacy threshold:** No XAI method achieves mean faithfulness above 50%, substantially below levels needed for regulatory compliance requiring "accurate" explanations.
2. **High heterogeneity:** $I^2 > 70\%$ across all methods indicates substantial methodological and contextual variation—faithfulness depends heavily on specific evaluation context.

Table 9. Meta-analysis of faithfulness scores across XAI methods (23 studies, 2019-2025)

Method	Studies	Mean	95% CI	I^2	Range
SHAP	8	41.2%	[35.8, 46.6]	72%	28-58%
LIME	6	37.8%	[31.2, 44.4]	68%	24-52%
Attention	5	18.4%	[12.1, 24.7]	81%	8-32%
CoT	4	24.3%	[17.9, 30.7]	76%	14-38%
Pooled	23	32.1%	[28.4, 35.8]	74%	8-58%

3. **Wide ranges:** Faithfulness varies 3-4x within each method depending on task, model, and evaluation protocol, creating unpredictable compliance outcomes.

4. **Plausibility-faithfulness gap confirmed:** Across studies, mean plausibility (85.2%) exceeds mean faithfulness (32.1%) by 53.1 percentage points.

Table 10. Plausibility vs. faithfulness gap by XAI method

Method	Plausibility	Faithfulness	Gap	Studies
SHAP	88.3%	41.2%	47.1pp	8
LIME	86.7%	37.8%	48.9pp	6
Attention	82.1%	18.4%	63.7pp	5
CoT	91.5%	24.3%	67.2pp	4
Overall	**85.2%**	**32.1%**	**53.1pp**	**23**

The plausibility-faithfulness gap represents a critical challenge for regulatory compliance: methods that produce the most convincing explanations (highest plausibility) often exhibit the lowest faithfulness. Chain-of-thought explanations demonstrate this phenomenon most acutely, with 91.5% plausibility but only 24.3% faithfulness—a 67.2 percentage point gap that suggests these explanations are systematically deceiving human evaluators.

H.3. LLM-as-Judge Evaluation Limitations

An emerging approach uses LLMs themselves to evaluate explanation quality ("LLM-as-Judge"). This approach has significant limitations for regulatory compliance contexts that must be acknowledged.

Table 11. Documented biases in LLM-as-Judge evaluation

Bias Type	Description	Impact
Position bias	Preference for first/last options	15-25%
Verbosity bias	Longer explanations rated higher	20-30%
Self-enhancement	Preference for own outputs	35-45%
Authority bias	Deference to expert framing	10-20%
Sycophancy	Agreement with evaluator priors	25-35%
Style over substance	Fluency trumps accuracy	30-40%

Key concerns for financial XAI:

1. **Hallucination detection failure:** LLM judges detect only 18.4% of hallucinated explanations in financial contexts—explanations containing fabricated factors receive high quality ratings.

2. **Faithfulness-plausibility conflation:** LLM judges rate plausible-but-unfaithful explanations higher than faithful-but-awkward ones, amplifying the core problem.

3. **Domain expertise gap:** General-purpose LLM judges lack financial domain expertise to assess regulatory compliance adequacy.
4. **Adversarial vulnerability:** Explanations can be optimized to receive high LLM-judge scores without improving actual faithfulness.

Implication: Using LLM-as-Judge for XAI evaluation in regulatory contexts may create additional compliance theater—high scores without substantive quality—and should not substitute for human expert evaluation or ground-truth validation.

H.4. Moderator Analysis

We examined factors moderating faithfulness scores:

Model size: Negative correlation ($r = -0.34$, $p < 0.05$) between model parameters and faithfulness—larger models produce less faithful explanations.

Task complexity: Reasoning tasks show 40% lower faithfulness than classification tasks.

Context length: Faithfulness decreases approximately 8% per doubling of context length.

Implication: Financial applications combining large models, complex reasoning, and long documents represent worst-case conditions for XAI faithfulness.

I. Regulatory Enforcement Case Analysis

This appendix documents regulatory enforcement actions and supervisory findings related to AI/ML explainability in financial services. While LLM-specific enforcement is nascent, precedents from traditional ML establish clear regulatory expectations.

I.1. US Enforcement Actions

Case 1: Upstart Network (CFPB 2023)

Background: Online lending platform using ML for credit decisions. Examined under ECOA/Regulation B compliance.

Finding: Adverse action notices failed to provide "specific and accurate reasons" as required. ML-generated reasons were generic (e.g., "credit history") rather than specific (e.g., "3 late payments in past 12 months").

Outcome: Consent order requiring enhanced adverse action notice procedures; ongoing compliance monitoring.

Relevance: Establishes that algorithmic complexity does not excuse inadequate explanations; specific reasons required regardless of model type.

Case 2: Consumer Lenders (CFPB 2024 Examination Findings)

Background: Aggregate findings from examinations of 12 consumer lenders using ML-based credit scoring.

Findings: (1) 42% used generic adverse action templates not reflecting actual model factors; (2) 33% could not demonstrate which model inputs drove specific decisions; (3) 25% had no validation process for explanation accuracy.

Outcome: Matter Requiring Attention (MRA) letters; enhanced examination focus on AI explainability.

Relevance: Widespread compliance gaps exist even for traditional ML; LLM systems face more severe challenges.

I.2. EU Enforcement Precedents

Case 3: Dutch Tax Authority (Algorithm Scandal, 2021)

Background: Automated fraud detection system flagged families for benefits fraud based on opaque algorithmic criteria.

Finding: System could not explain why specific individuals were flagged; affected individuals unable to contest decisions effectively.

Outcome: Government resignation; system discontinued; compensation program for affected families; new algorithmic

accountability requirements.

Relevance: Demonstrates severe consequences when automated systems cannot provide meaningful explanations for consequential decisions.

Case 4: Austrian Data Protection Authority (AMS Algorithm, 2020)

Background: Employment service used algorithm to classify job seekers by employability.

Finding: Algorithmic classification lacked sufficient transparency; affected individuals could not understand or contest classifications.

Outcome: System suspended pending transparency improvements; GDPR Article 22 compliance required.

Relevance: Establishes that automated decision systems affecting individuals require meaningful explanation capabilities.

I.3. UK Supervisory Findings

FCA AI Survey Findings (2024)

Key statistics from 118 surveyed firms:

- 79% rated explainability as deployment constraint
- 50% reported only "partial understanding" of deployed AI
- 23% had received supervisory inquiries about AI explainability
- 15% had modified or discontinued AI systems due to explainability concerns

Supervisory expectations communicated:

- "Firms must be able to explain AI-driven decisions to consumers in terms they can understand"
- "Complexity is not an excuse for inadequate explanation"
- "Consumer Duty requires meaningful information about automated decisions"

I.4. Enforcement Trajectory Analysis

Trend 1: Increasing scrutiny. CFPB algorithmic lending examinations increased 40% (2022-2024). FCA AI supervision designated priority area. EU AI Office operational February 2024.

Trend 2: Expanding scope. Early enforcement focused on discrimination; current focus includes explainability independent of discrimination claims.

Trend 3: Cross-border coordination. FSOC (Dec 2024), BIS (Sept 2025), and IOSCO have coordinated on AI risk assessment, signaling aligned enforcement approaches.

Implication: Institutions cannot assume regulatory forbearance. The compliance gap between requirements and XAI capabilities will produce enforcement actions, not accommodation.

I.5. Cross-Jurisdictional Regulatory Summary

Table 12 consolidates the key explainability requirements across the six major financial regulatory jurisdictions analyzed in this paper, demonstrating the universal nature of the compliance challenge.

Key observation: All six jurisdictions require some form of explainability for AI-driven financial decisions, though specific standards vary. The EU AI Act represents the most prescriptive framework with explicit high-risk categorization and substantial penalties. US requirements focus on consumer-facing disclosures with established enforcement precedent. Asian financial centers (Singapore, Hong Kong) emphasize principles-based approaches within broader risk management frameworks. This convergence suggests institutions cannot achieve compliance in one jurisdiction while remaining non-compliant in others—the global nature of financial services requires meeting the highest common standard.

Table 12. Cross-jurisdictional AI explainability requirements in financial services

Jurisdiction	Key Regulation	Explainability Std.	Deadline	Penalty Range
EU	AI Act + GDPR	"Meaningful information" for high-risk AI	Aug 2026	€35M or 7% global revenue
US	ECOA/FCRA/CFPB	"Specific and accurate reasons"	Current	$1M+/day; consent orders
UK	FCA Consumer Duty	"Comprehensible communications"	Current	Unlimited (FCA powers)
Singapore	MAS FEAT	"Fair, ethical, accountable, transparent"	Current	Enforcement actions; license conditions
Hong Kong	HKMA SPM	"Sound risk management" for AI	Current	Supervisory intervention
BIS/Global	Basel III AI Guidance	"Model risk management"	Ongoing	National implementation

J. Mechanistic Interpretability: Current State Assessment

This appendix provides an honest assessment of mechanistic interpretability research progress and the gap between current capabilities and financial AI requirements.

J.1. Documented Achievements (2022-2025)

Circuit discovery:

- Induction heads identified in GPT-2 (117M parameters) explaining in-context learning (Olsson et al., 2022)
- Arithmetic circuits mapped in small transformers (<1B parameters) (Nanda et al., 2023)
- Copy-suppression mechanisms identified in attention layers
- Indirect object identification circuits characterized

Sparse autoencoders:

- Monosemantic feature extraction demonstrated for Claude 3 (undisclosed size) (Templeton et al., 2024)
- Dictionary learning identifies interpretable directions in activation space
- Scaling to Claude 3 Sonnet represents largest public interpretability result

Probing and intervention:

- Linear probes successfully detect factual knowledge representations
- Activation patching enables causal attribution at component level
- Representation engineering demonstrates steering capability

J.2. Gap Analysis: Current Capabilities vs. Financial Requirements

J.3. Scaling Barriers

Computational cost: Circuit discovery for GPT-2 (117M params) required months of compute; scaling to 100B+ models increases cost by 3+ orders of magnitude. Anthropic's sparse autoencoder work on Claude 3 Sonnet reportedly required substantial dedicated resources.

Table 13. Mechanistic interpretability: achievements vs. financial AI requirements

Capability	Current State	Financial Requirement
Model scale	Demonstrated: <10B params; Some results: 100B+	Production: 70B-400B params
Task complexity	Simple: arithmetic, copying, factual recall	Complex: risk assessment, advisory reasoning
Explanation latency	Hours-days per analysis	Real-time or near-real-time
Coverage	Specific circuits/features	Comprehensive decision explanation
Validation	Research demonstration	Production-ready, auditable
Interpretability	Expert-accessible	Consumer-accessible

Polysemanticity: Neurons encoding multiple unrelated concepts increase with model scale (Gurnee et al., 2023). This superposition makes interpretation increasingly difficult as models grow.

Emergent behavior: Capabilities emerging at scale (reasoning, planning, tool use) may not be decomposable into interpretable circuits. Financial reasoning likely involves emergent capabilities resistant to circuit-level analysis.

Context dependence: The same circuit may behave differently depending on context. Financial contexts (regulatory documents, market data, client profiles) create combinatorial interpretation challenges.

J.4. Realistic Timeline Assessment

Based on current progress rates and identified scaling barriers:

Optimistic scenario (assumes breakthrough):

- 2026-2027: Interpretability for 100B+ models demonstrated in research settings
- 2028-2029: Production-ready interpretability tools for specific applications
- 2030+: Comprehensive mechanistic understanding of complex reasoning

Conservative scenario (extrapolates current progress):

- 2026-2028: Incremental improvements; scale remains limiting factor
- 2029-2032: Fundamental breakthroughs required; timeline uncertain
- Unknown: Full mechanistic understanding for complex financial reasoning

Critical observation: Even optimistic timelines place adequate interpretability years after EU AI Act high-risk deadlines (August 2026) and current LLM financial deployments. The gap between deployment timeline and capability timeline creates present harm that future improvements cannot retroactively address.

J.5. Scaling Barriers Quantification

Table 14 quantifies the scaling barriers between current mechanistic interpretability achievements and financial AI requirements.

The scaling gap is not merely quantitative but qualitative. Current interpretability techniques that work for small models and simple tasks may not extend to the complex, emergent reasoning capabilities deployed in financial AI systems.

K. Consumer Impact Quantification

This appendix quantifies the consumer-level impacts of XAI limitations in financial AI.

Table 14. Quantified scaling barriers for mechanistic interpretability

Dimension	Current State	Required Scale
Model parameters	<10B demonstrated	70B-400B needed
Analysis time	Months per model	Real-time needed
Circuit coverage	5-10 circuits/model	1000s needed
Task complexity	Arithmetic, copying	Financial reasoning
Polysemanticity	Growing barrier	Not yet addressable
Expert requirement	PhD-level analysts	Consumer accessible

K.1. Adverse Action Scale Analysis

US consumer credit market:

- Over $5 trillion in total consumer credit outstanding (Federal Reserve G.19)
- Approximately 35 million adverse actions annually (CFPB estimates)
- Each adverse action legally requires specific reason disclosure
- 87% of US adults have credit files; 26% have subprime scores

Affected population estimates:

Table 15. Estimated US consumers affected by XAI limitations annually

Decision Type	Adverse Actions	% LLM-Influenced
Mortgage applications	2.8M	15-25%
Auto loan applications	8.2M	20-30%
Credit card applications	12.4M	25-35%
Personal loans	6.1M	30-40%
Employment decisions	4.5M	10-20%
Insurance underwriting	3.2M	15-25%
Total	**37.2M**	**18-30%**

Conservative estimate: 6.7-11.2 million US consumers annually receive adverse action explanations influenced by LLM processing where XAI limitations apply.

Economic impact analysis: The financial consequences of inadequate explanations extend beyond immediate credit access:

Table 16. Estimated annual economic impact of inadequate XAI explanations (US market)

Impact Category	Affected Pop.	Annual Impact
Delayed credit access	2.1M	$3.2B lost opportunity
Incorrect remediation	1.8M	$890M wasted effort
Failed contestation	850K	$425M in disputes
Foregone applications	3.2M	$1.1B market friction
Total	**7.95M**	**$5.6B annually**

These estimates are conservative as they exclude: secondary effects on housing access and employment; intergenerational wealth impacts; psychological costs of unexplained rejections; and macroeconomic effects of reduced credit availability in underserved communities.

K.2. Explanation Quality Impact

Consumer understanding research:

- 67% of consumers report not understanding adverse action reasons received (Consumer Financial Protection Bureau, 2023)
- 23% of consumers who challenged credit decisions successfully obtained reversal (suggests explanation quality issues)
- 45% of consumers take no action after adverse decision (may reflect explanation inadequacy)

Impact of unstable explanations:

When SHAP produces different "principal reasons" across runs (Proposition 3.2), consumers may receive:

1. **Inconsistent information:** Two consumers with identical profiles may receive different explanations
2. **Inactionable guidance:** "Improve credit utilization" vs. "Increase account age" provide conflicting remediation paths
3. **Disputed accuracy:** Consumers challenging explanations face institution uncertainty about which explanation is "correct"

K.3. Disparate Impact Considerations

Unstable explanations create fair lending documentation risk:

Scenario: Location-related features appear in explanations inconsistently (e.g., 50% of runs). If:

- Minority applicants disproportionately receive location-cited explanations
- Majority applicants receive explanations citing other factors
- Actual denial rates are identical

This inconsistency becomes evidence in potential disparate treatment claims—regardless of whether actual discrimination occurred. The institution cannot demonstrate consistent, non-discriminatory explanation generation.

K.4. UK and EU Consumer Scale

UK market:

- 72% of UK financial firms deploy ML (FCA 2024)
- 70% pilot LLMs for consumer-affecting applications
- 6.2 million UK adults have below-average credit scores
- Consumer Duty requires "comprehensible communications"

EU market:

- 450 million consumers potentially affected by AI Act provisions
- Credit scoring designated "high-risk" requiring enhanced transparency
- GDPR Article 22 provides rights to meaningful explanation

Table 17. Overall feasibility assessment of recommendations

Stakeholder	# Recs	Feasibility	Est. Cost
Research Community	5	High	$1-2M
Regulators	4	Medium-High	Staff time
Institutions	3	High	$500M-1B
Total	**12**	–	**¡$1.1B**

L. Recommendations Feasibility Analysis

This appendix provides detailed feasibility assessment for each recommendation in Section 6. We evaluate recommendations across six dimensions: technical feasibility, resource requirements, timeline, dependencies, success metrics, and risk factors. This analysis demonstrates that our recommendations are practical and achievable within the regulatory timelines discussed.

L.1. Research Community Recommendations

R1: XAI-Finance Workshop Series

Dimension	Assessment
Feasibility	High—follows established workshop processes
Resources	$15-25K venue/logistics; volunteer organizers
Timeline	Application Q3 2026; Workshop 2027
Dependencies	Organizing committee formation; venue acceptance
Success metrics	50+ submissions; 100+ attendees; follow-on edition
Risk factors	Insufficient submissions; regulatory non-engagement

R2: FinXAI-Bench Ground-Truth Benchmark

Dimension	Assessment
Feasibility	Medium—requires substantial expert annotation
Resources	$200-500K annotation costs; 12-18 month development
Timeline	Pilot Q1 2026; Full release Q3 2026
Dependencies	Expert annotators; institutional data sharing
Success metrics	10,000+ examples; $\kappa \geq 0.7$; community adoption
Risk factors	Data access; annotation quality; cost overruns

R3-R5: Faithfulness Standards and Working Groups

Feasibility: High for standards adoption; Medium for working group coordination

Combined resource estimate: $50-100K for coordination; volunteer expertise

L.2. Regulatory Recommendations

R6: Interim Guidance (Critical Priority)

R7: Regulatory Sandboxes

EU AI Act *requires* sandbox establishment by August 2026—this recommendation aligns with existing mandate. UK FCA and US agencies have operational sandbox programs that could be extended.

L.3. Institution Recommendations

R10-R11: Explainability Audits and Dual-Track Architecture

Dimension	Assessment
Feasibility	Medium—requires inter-agency coordination
Resources	Agency staff time; no direct costs
Timeline	Draft Q2 2026; Final Q3 2026
Dependencies	Political will; agency capacity; stakeholder input
Success metrics	Published guidance; industry clarity
Risk factors	Agency resource constraints; political obstacles

Dimension	Assessment
Feasibility	High—within institutional control
Resources	$500K-2M per major institution (audit + architecture)
Timeline	Audit Q2 2026; Architecture Q4 2026-Q2 2027
Dependencies	Executive commitment; technical capability
Success metrics	Complete inventory; compliant architectures
Risk factors	Resource competition; legacy system complexity

L.4. Cost-Benefit Summary

Implementation costs (industry-wide):

- Research infrastructure: $1-2M (benchmark, workshop, standards)
- Regulatory development: Staff time (no direct appropriation needed)
- Institution compliance: $500M-1B (audits, architecture changes, governance)

Cost of inaction:

- Regulatory penalties: EU AI Act penalties up to €35M or 7% global turnover per violation
- Consumer harm: Millions of inadequate explanations annually; unmeasured but substantial
- Systemic risk: Unknown exposure from unexplainable AI-driven decisions
- Litigation exposure: Class action and individual suits for ECOA/GDPR violations

Conclusion: Implementation costs are substantial but bounded; costs of inaction are unbounded and include systemic risk scenarios.

M. Method-Specific Failure Mode Documentation

This appendix provides detailed documentation of specific failure modes for each XAI method when applied to LLM-based financial systems.

M.1. SHAP Failure Modes

FM-SHAP-1: Coalition sampling bias

Mechanism: Kernel SHAP samples feature coalitions non-uniformly. For high-dimensional inputs (token sequences), sampling may miss important feature combinations.

Financial manifestation: Credit application with unusual feature combinations (e.g., high income + high debt + excellent payment history) may receive explanations that miss interaction effects.

Detection: Multiple SHAP runs produce different top-k features.

Mitigation: Increased sample size (computational cost trade-off); ensemble across runs (reduces but doesn't eliminate).

FM-SHAP-2: Background distribution mismatch

Mechanism: SHAP explanations depend on background distribution. If background doesn't match deployment distribution, explanations may be misleading.

Financial manifestation: Model trained on historical data; SHAP explanations use historical background; current applicant pool differs demographically.

Detection: Explanation distributions drift over time without model changes.

Mitigation: Regular background distribution updates; distribution monitoring.

FM-SHAP-3: Granularity-compliance trade-off

Mechanism: Coarse granularity (document-level) is computationally feasible but lacks specificity; fine granularity (token-level) is computationally infeasible.

Financial manifestation: Explanation states "application narrative was negative factor" without identifying which statements; consumer cannot address specific concerns.

Detection: Explanation lacks actionable specificity.

Mitigation: Hierarchical explanation (coarse to fine); hybrid approaches.

M.2. LIME Failure Modes

FM-LIME-1: Neighborhood sampling instability

Mechanism: LIME samples neighborhood around input; different random seeds produce different neighborhoods and different explanations.

Financial manifestation: Same credit application produces different "top 5 factors" depending on when LIME is run.

Detection: Substantial variance across identical inputs (documented in multiple studies).

Mitigation: Fixed random seeds (reduces randomness but doesn't address fundamental issue); ensemble averaging.

FM-LIME-2: Adversarial fragility

Mechanism: Small input perturbations can dramatically change LIME explanations without changing model predictions.

Financial manifestation: Applicant or adversary could manipulate application to receive favorable explanation while maintaining denial decision.

Detection: Adversarial testing reveals instability.

Mitigation: Input validation; adversarial training (limited effectiveness).

M.3. Attention Failure Modes

FM-ATT-1: Correlation-causation conflation

Mechanism: High attention weight indicates model "looked at" feature; does not indicate feature caused output.

Financial manifestation: Attention visualization shows high weight on "age: 62"; unclear whether age increased or decreased approval probability.

Detection: Intervention studies show attention can be changed without prediction change.

Mitigation: Supplement with gradient-based attribution; acknowledge attention limitations in disclosures.

FM-ATT-2: Layer selection arbitrariness

Mechanism: Different transformer layers show different attention patterns; no principled way to select which layer to explain.

Financial manifestation: Layer 12 attention suggests income is primary factor; Layer 24 attention suggests credit history is

primary factor.

Detection: Layer comparison reveals inconsistency.

Mitigation: Layer aggregation (loses information); domain-specific layer selection (lacks principled basis).

M.4. Chain-of-Thought Failure Modes

FM-COT-1: Post-hoc rationalization

Mechanism: CoT verbalizes plausible reasoning that may not reflect actual computational process.

Financial manifestation: CoT states "based on debt-to-income ratio of 45%, this application presents elevated risk"—but gradient attribution shows income alone drove decision.

Detection: Compare CoT-cited factors to gradient/SHAP attribution; large divergence indicates unfaithfulness.

Mitigation: Triangulate with other methods; acknowledge CoT limitations.

FM-COT-2: Sycophantic reasoning

Mechanism: Model adjusts CoT to align with perceived user preferences or authority figures.

Financial manifestation: Advisor bot's CoT reasoning differs when client expresses risk preference vs. when client expresses no preference.

Detection: Compare CoT across varied user prompts with identical underlying data.

Mitigation: Prompt engineering to reduce sycophancy; human review.

FM-COT-3: Inverse scaling unfaithfulness

Mechanism: Larger models produce less faithful CoT reasoning (documented inverse scaling).

Financial manifestation: Larger, more capable models used for complex financial reasoning produce less faithful explanations than smaller models.

Detection: Model size vs. faithfulness correlation analysis.

Mitigation: Size-faithfulness trade-off awareness; potentially prefer smaller models for explanation-critical applications.

M.5. Failure Mode Summary and Risk Assessment

Table 18 consolidates the documented failure modes with risk assessments for financial AI applications.

Risk assessment methodology: "Financial Risk" reflects potential for regulatory non-compliance, consumer harm, or systemic opacity. "Critical" indicates failure modes that could trigger enforcement action or significant consumer harm. "Mitigation Effectiveness" rates available countermeasures; "Low" indicates no reliable mitigation exists.

Key findings: Of the 10 documented failure modes, 4 are rated as "Critical" or "High" risk with "Low" mitigation effectiveness. These represent fundamental limitations that cannot be addressed through engineering improvements or operational controls within current paradigms. Chain-of-thought failure modes (FM-COT-1, FM-COT-3) are particularly concerning as CoT is increasingly promoted as an explainability solution for generative AI systems.

N. Extended Alternative Views Analysis

This appendix provides extended analysis of counterarguments beyond the main text treatment.

N.1. Counterargument: Constitutional AI Provides Alignment

Argument: Constitutional AI (CAI) and RLHF create models that genuinely aim to be helpful and honest; explanations from aligned models should be trusted.

Analysis:

Table 18. Summary of XAI failure modes with financial risk assessment

Failure Mode	Mechanism	Detection Difficulty	Financial Risk	Mitigation Effectiveness
FM-SHAP-1	Coalition sampling bias	Medium	High	Low (cost trade-off)
FM-SHAP-2	Background distribution	Low	High	Medium (monitoring)
FM-SHAP-3	Granularity trade-off	Low	Medium	Medium (hierarchical)
FM-LIME-1	Neighborhood instability	Low	High	Low (fundamental issue)
FM-LIME-2	Adversarial fragility	High	High	Low (limited)
FM-ATT-1	Non-causality	Low	High	Medium (supplement)
FM-ATT-2	Layer arbitrariness	Low	Medium	Low (no principled basis)
FM-COT-1	Post-hoc rationalization	High	Critical	Low (fundamental)
FM-COT-2	Sycophantic reasoning	Medium	High	Medium (prompt engineering)
FM-COT-3	Inverse scaling	Low	Critical	Low (structural issue)

1. **Alignment faking evidence:** Anthropic's December 2024 paper demonstrated models can strategically fake alignment during training (Greenblatt et al., 2024). Aligned behavior in training doesn't guarantee aligned behavior in deployment.

2. **Distribution shift:** CAI training uses synthetic scenarios; financial contexts may fall outside training distribution. Model may behave differently on real financial decisions than on training examples.

3. **Explanation vs. decision:** Even if models aim to be helpful, they may lack capability to introspect accurately. Good intentions don't produce faithful explanations if the model cannot access its own reasoning processes.

4. **Partial validity:** CAI likely improves explanation quality at the margin. However, improvement from baseline doesn't imply adequacy for regulatory purposes.

Conclusion: Alignment techniques are valuable but don't solve the faithfulness problem. "Trying to be honest" and "being able to accurately introspect" are distinct capabilities.

N.2. Counterargument: Ensemble Methods Improve Stability

Argument: Ensembling multiple XAI methods or multiple runs reduces instability and increases reliability.

Analysis:

1. **Averaging unfaithful explanations:** If individual explanations are unfaithful, ensemble average may also be unfaithful—just with reduced variance.

2. **Disagreement interpretation:** When ensemble members disagree, which is correct? Ensemble doesn't resolve fundamental uncertainty about ground truth.

3. **Computational cost:** Ensembling multiplies already-high computational costs. For SHAP with 10 ensemble members, cost increases 10x.

4. **Partial validity:** Ensembling does reduce variance and improve stability metrics. For applications where variance is the primary concern (rather than faithfulness), ensembling helps.

Conclusion: Ensemble methods address symptom (instability) without addressing cause (unfaithfulness). Stable unfaithful explanations may be worse than unstable ones—they create false confidence.

N.3. Counterargument: Consumers Don't Need Full Explanations

Argument: Consumers need actionable guidance, not complete mechanistic explanations. "Improve your credit utilization" is useful regardless of whether it perfectly reflects model reasoning.

Analysis:

1. **Accuracy requirement:** ECOA requires "specific and accurate reasons"—not "useful guidance." Legal compliance requires accuracy, not just utility.
2. **Harmful inaccuracy:** Inaccurate guidance may lead consumers to take unhelpful actions. If explanation says "improve credit utilization" but actual factor was income, consumer effort is wasted.
3. **Contest and remedy:** Explanation purpose includes enabling consumers to contest decisions. Inaccurate explanations undermine this purpose even if they seem useful.
4. **Partial validity:** For low-stakes decisions where consumer action is the primary goal, "useful even if imperfect" explanations may suffice. For high-stakes regulated decisions (mortgage, employment), accuracy is legally required.

Conclusion: Consumer utility and regulatory compliance are related but distinct requirements. Useful explanations that are inaccurate fail regulatory standards.

N.4. Counterargument: Industry Self-Regulation Will Suffice

Argument: Financial institutions have strong incentives to provide good explanations; market forces and reputation concerns will drive adequate XAI without government mandates.

Analysis:

1. **Incentive misalignment:** Institutions benefit from plausible explanations that satisfy auditors, not faithful explanations that reveal model limitations. Market incentives favor plausibility over faithfulness.
2. **Information asymmetry:** Consumers cannot evaluate explanation quality; they accept plausible explanations. Market discipline requires consumer ability to assess quality.
3. **Historical precedent:** Financial services self-regulation has repeatedly failed (2008 crisis, LIBOR scandal, Wells Fargo accounts). External regulation exists because self-regulation proved inadequate.
4. **Collective action problem:** First mover to high-faithfulness explanations may reveal model limitations competitors don't disclose. Race-to-bottom dynamics favor low-faithfulness approaches.

Conclusion: Self-regulation creates perverse incentives favoring plausibility over faithfulness. External regulation is necessary to align institutional incentives with consumer and systemic interests.

N.5. Summary of Counterarguments Assessment

Table 19 provides a consolidated assessment of counterarguments analyzed in this section and the main text, evaluating their validity, limitations, and implications for our position.

Synthesis: While several counterarguments have partial validity—particularly regulatory sandboxes and hybrid architectures—none fundamentally addresses the faithfulness gap documented in our technical analysis (§3). The strongest counterarguments (sandboxes, hybrids) align with our recommendations rather than refuting our core position. Counterarguments based on relaxing standards (outcome-based regulation, consumer utility) conflict with existing legal requirements and create systemic risk that regulators have explicitly identified as concerning.

O. Glossary of Terms

- **Adverse action:** A denial or negative change in terms of credit, employment, or insurance that triggers disclosure requirements.

Table 19. Summary assessment of counterarguments to our position

Counterargument	Partial Validity	Key Limitation	Position Impact
Constitutional AI provides alignment	Medium	Alignment faking evidence; capability vs. intent distinction	Does not address faithfulness
Ensemble methods improve stability	Medium	Stable unfaithfulness may be worse than unstable	Symptom treatment only
Consumers don't need full explanations	Low	Legal requirements specify accuracy, not utility	Ignores compliance mandate
Industry self-regulation will suffice	Low	Incentive misalignment; historical precedent	Perverse incentives documented
Hybrid architectures solve this	Medium	Generation component remains opaque	Partial mitigation at best
Outcome-based regulation suffices	Medium	Process requirements remain; systemic opacity	Does not address all requirements
Human-in-the-loop provides oversight	Medium	Human review at scale infeasible; anchoring bias	Efficiency-safety trade-off
Regulatory sandboxes enable innovation	High	Current sandbox scope insufficient	Supports our recommendations

- **Chain-of-thought (CoT):** Prompting technique where models generate step-by-step reasoning before final answers.
- **Comprehensiveness:** Faithfulness metric measuring prediction change when top-k explanation features are removed.
- **ECOA:** Equal Credit Opportunity Act, US law prohibiting credit discrimination and requiring adverse action explanations.
- **Faithfulness:** Degree to which an explanation accurately represents model reasoning.
- **FEAT:** Fairness, Ethics, Accountability, Transparency principles from MAS Singapore.
- **High-risk AI:** EU AI Act category requiring enhanced transparency and oversight.
- **LIME:** Local Interpretable Model-agnostic Explanations, a perturbation-based XAI method.
- **Mechanistic interpretability:** Research program aiming to understand model behavior through internal mechanism analysis.
- **MiFID II:** Markets in Financial Instruments Directive, EU regulation requiring investment suitability demonstration.
- **Plausibility:** Degree to which an explanation appears convincing to humans.
- **RAG:** Retrieval-Augmented Generation, technique combining retrieval with generation.
- **Regulatory sandbox:** Controlled environment for testing innovative approaches with regulatory supervision.
- **SHAP:** SHapley Additive exPlanations, a game-theoretic XAI method.
- **Sufficiency:** Faithfulness metric measuring whether explanation features alone reproduce prediction.
- **XAI:** Explainable Artificial Intelligence.