TrustLLM-Fin: A Privacy-Centric and Auditable Impact Assessment Framework for Large Language Models in Automated Financial Reporting

Yue Chen , Litong Song , Ziwei Liu , Jingyu Yao , Kaichen Liu , Qiuyue Liao *

Posted Date: 22 January 2026

doi: 10.20944/preprints202601.1596.v1

Keywords: large language models; financial reporting; auditability; privacy; prompt injection; EU AI act; impact assessment; trust score; analytic hierarchy process

 doi:10.20944/preprints202601.1596.v1

Article

TrustLLM-Fin: A Privacy-Centric and Auditable Impact Assessment Framework for Large Language Models in Automated Financial Reporting

Yue Chen [1], Litong Song [2], Ziwei Liu [1], Jingyu Yao [3], Kaichen Liu [4] and Qiuyue Liao [5,*]

[1] Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA
[2] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA
[3] Department of Computer Science, Yale University, New Haven, CT 06520, USA
[4] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA
[5] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
* Correspondence: qliao34@gatech.edu

Abstract

The powerful use of Large Language Models to automate financial reporting is currently hindered by a fundamental disconnect: popular models generate stochastic outputs, but the industry expects absolute precision and privacy. General-purpose LLMs are outstanding at reiterating guidelines in natural language, but are still prone to "knowledge overriding," hallucinated financial figures, as well as adversarial prompt injections, shortcomings that are existential threats to the institutional fiduciary duties and regulatory compliance required by the EU AI Act. We propose TrustLLM-Fin, a domain-specific, privacy-preserving and auditable impact assessment framework to systematically model the real-world financial disclosure workflow. We identify orthogonal dimensions in the "Trust Surface" of LLM-powered FinTech Agents: Prompt Safety & Privacy, Factuality & Robustness, Auditability, and we further quantize abstract governance principles into a numerical "Trust Score" using the Analytic Hierarchy Process. We test the efficacy of the TrustLLM-Fin framework in a high-fidelity financial dataset constructed from real-world SEC 10-K/10-Q filings. Strikingly, a standard general-purpose LLM scored a *TS* of only 0.234, due to a 93.33% success rate for context-embedded adversarial attacks. The Guardrail-Enhanced TrustLLM-Fin scored a *TS* of 0.710, revealing the critical decoupling between generative fluency and instruction following. Our results show that in practice, trustworthiness is not an emergent property of scale, but requires systemic alignment by design. The proposed framework provides auditors and financial institutions with a pragmatic, extensible framework to ensure the deployment of generative agents is transparent, robust and auditable in a changing regulatory landscape.

Keywords: large language models; financial reporting; auditability; privacy; prompt injection; EU AI act; impact assessment; trust score; analytic hierarchy process

1. Introduction

Large language models (LLMs) have begun to occupy a visible role in financial analysis, reporting, and decision-support workflows [1]. Their capacity to summarize lengthy disclosures, extract salient information from complex documents, and generate seemingly coherent financial narratives has made them attractive to institutions seeking efficiency gains in compliance, risk analysis, and reporting processes [2]. As a result, LLM-based systems are increasingly deployed not only in internal analytical pipelines, but also in externally facing financial services [3].

Yet the integration of generative models into finance introduces a class of risks that cannot be reduced to conventional notions of predictive accuracy or task performance. Financial communication operates within a tightly regulated institutional environment, where the legitimacy of information is

 doi:10.20944/preprints202601.1596.v1

defined not only by whether it is factually correct, but by whether it is legally permissible to disclose [4]. Rules governing privacy protection, material non-public information (MNPI), and auditability impose strict boundaries on what may be generated, inferred, or summarized. Within this setting, a language model that produces a fluent but impermissible response may pose greater harm than a system that declines to answer altogether [5].

Despite this asymmetry, the evaluation of financial language models continues to rely largely on benchmarks inherited from general-purpose natural language processing. Existing evaluations prioritize task-level correctness, numerical reasoning ability, or sentiment classification performance, implicitly rewarding completeness and informativeness [6]. Such metrics are poorly aligned with the primary sources of financial risk. They fail to capture failure modes that dominate real-world deployments, including the fabrication of unverifiable financial statements, unauthorized inference of sensitive information, and outputs that cannot be traced back to publicly disclosed sources [7]. Consequently, even models that score highly on standard benchmarks may remain unsuitable for use in regulated financial environments.

In parallel with these evaluation practices, recent research has made substantial progress in improving document-level reasoning and factual grounding for large language models. In particular, multi-hop retrieval-augmented generation (RAG) has been shown to significantly enhance long-context question answering by explicitly modeling chains of supporting evidence across documents [8]. By moving beyond single-pass retrieval, such approaches reduce unsupported generation in complex document-level tasks and represent an important step toward more reliable long-document understanding.

Related advances further emphasize the role of structural constraints and domain adaptation in improving robustness and consistency under distribution shift. For example, camera-aware graph consistency and local feature enhancement methods demonstrate how incorporating explicit structural signals can improve cross-domain generalization in unsupervised settings [9,10]. In addition, work on multi-granularity adapter fusion for privacy policy understanding highlights the importance of modular adaptation and dynamic constraint enforcement when processing sensitive, regulation-heavy documents [11]. Collectively, these studies reflect a broader trend toward architecture-level and system-level mechanisms for improving factual reliability, robustness, and compliance-awareness in document-centric language modeling.

Nevertheless, improvements in retrieval, representation, or reasoning alone do not guarantee institutionally acceptable behavior in regulated financial contexts. Even when advanced retrieval-augmented or multi-hop reasoning mechanisms are employed, language models may still generate impermissible disclosures, infer material non-public information, or produce outputs whose provenance cannot be audited. These failure modes stem from a more fundamental mismatch between how LLMs are optimized and how financial institutions reason about risk. LLMs are trained to generate statistically likely continuations of text, not to respect disclosure boundaries, legal constraints, or institutional accountability [12]. In the absence of explicit governance mechanisms, such models systematically favor plausibility over restraint, particularly in adversarial or ambiguous contexts where manipulative instructions are embedded within otherwise legitimate financial narratives [13].

Recent work on LLM safety has proposed a range of mitigation strategies, including prompt engineering, post-hoc filtering, and general alignment techniques [14]. While these approaches offer partial improvements, they are typically domain-agnostic and provide limited assurances in high-stakes financial settings. More critically, they are seldom accompanied by evaluation frameworks capable of measuring whether a model behaves appropriately under realistic compliance constraints. Without such evaluations, it remains unclear whether proposed safeguards meaningfully reduce institutional risk or merely improve surface-level robustness.

In this work, we argue that trustworthy financial language modeling requires a reorientation of both system design and evaluation priorities [15]. Trust should not be treated as an emergent byproduct of model scale or training quality, but as a first-class objective grounded in the regulatory and

 doi:10.20944/preprints202601.1596.v1

institutional realities of finance [16]. To this end, we introduce *TrustLLM-Fin*, a governance-oriented framework that evaluates large language models based on their behavior under compliance-critical conditions. The framework couples a realistic adversarial benchmark derived from regulatory disclosures with a multidimensional TrustScore that jointly captures adversarial robustness, hallucination control, and auditability.

Our empirical results show that, when evaluated under these conditions, general-purpose language models exhibit severe and systematic failure modes even on authentic financial texts. By contrast, system-level governance mechanisms substantially reduce these risks without altering the underlying model architecture. These findings indicate that trustworthiness in financial AI is not an automatic consequence of increased model capacity, but rather the outcome of deliberate design choices and evaluation criteria aligned with regulatory risk.

By reframing benchmarking as a stress test for institutional alignment rather than task competence, this work contributes to a growing body of research on trustworthy AI in regulated domains. While our focus is on finance, the insights and methodology extend to other high-stakes settings where the cost of impermissible information outweighs the benefits of maximal informativeness. We hope this work encourages further research into trust-centric evaluation frameworks and governance mechanisms for generative AI systems.

The contributions of this work are three-fold:

- **Multi-dimensional quantification:** We propose a hierarchical metrology that connects abstract regulatory mandates, such as those articulated in the EU AI Act, to measurable technical key performance indicators for financial reporting tasks.
- **Context-aware adversarial evaluation:** Rather than relying on generic jailbreak prompts, our framework introduces adversarial scenarios grounded in real-world financial disclosure risks, including MNPI fabrication and direct prompt injection, yielding an ecologically valid stress test for financial LLMs.
- **Guardrail validation and comparative insights:** Building on recent advances in document-level reasoning and retrieval-augmented generation, we demonstrate through a case study of SEC filing summarization that a guardrail-enhanced architecture achieves substantially higher attribution traceability and privacy resilience than a general-purpose configuration, providing a practical blueprint for deploying LLMs in compliance-sensitive domains.

2. Related Work

2.1. FinLLMs' Progression Trail: Discriminative Versus Generative Architectures

Research in financial natural language processing has evolved largely in parallel with broader developments in representation learning. Early work in Fin-NLP was dominated by discriminative, task-oriented architectures, most notably encoder-based models adapted from BERT [17]. These systems were designed to perform well-defined supervised tasks such as sentiment classification, named entity recognition, and event extraction. By incorporating domain-specific pretraining on financial text, they achieved substantial improvements over generic language models [17]. At the same time, their practical scope remained limited by narrowly defined objectives and fixed output structures.

The introduction of large-scale generative language models altered this landscape in a more fundamental way. Finance-oriented models such as BloombergGPT [1] and FinGPT [2] demonstrated that extensive pretraining on curated financial corpora could yield broad performance gains across a range of downstream tasks, often surpassing general-purpose LLMs. These results fueled expectations that language models could move beyond discrete analytical functions toward more integrated forms of financial analysis, reporting, and narrative synthesis. Benchmark suites such as FinBen [6] appeared to support this trajectory by reporting strong results on standardized evaluations involving sentiment interpretation, descriptive reasoning, and numerical computation.

 doi:10.20944/preprints202601.1596.v1

Closer examination, however, suggests that these performance gains do not generalize cleanly to the forms of communication encountered in real financial practice. Earnings calls, regulatory filings, and market-sensitive announcements are rarely well-structured or stylistically uniform. They combine forward-looking statements, legal qualifiers, and incomplete or deliberately ambiguous disclosures. When operating under such conditions, even finance-specialized LLMs display fragile reasoning behavior and inconsistent handling of numerical and contextual constraints. Linguistic fluency often conceals deeper weaknesses in constraint awareness and institutional reasoning, a limitation commonly characterized as the "stochastic parrot" effect [12].

These observations call into question evaluation frameworks that equate progress in FinLLMs with improvements on discriminative or stylized generative benchmarks. As financial applications increasingly rely on autonomous generation rather than classification or extraction, assessing model capability solely through task accuracy becomes insufficient. What is required instead is an evaluation perspective that accounts for how models behave when confronted with legally constrained, incomplete, or adversarial financial narratives.

2.2. Hallucination Dynamics and Factual Fidelity in the Financial Domain

Hallucination has become a widely recognized challenge in generative language modeling [7], but its consequences are particularly acute in financial settings. Unlike open-domain conversational applications, financial reporting allows little tolerance for factual deviation. Even small numerical errors or unsupported claims can propagate through valuation models, compliance reviews, and risk assessments, producing effects that far exceed their apparent magnitude.

A growing body of work documents the tendency of LLMs to generate content that is fluent yet insufficiently grounded in evidence. In finance, such behavior cannot be dismissed as a harmless byproduct of probabilistic generation. A fabricated ratio, an inferred disclosure detail, or a misrepresented figure may constitute materially misleading information with regulatory implications. To mitigate this risk, retrieval-augmented generation (RAG) has emerged as a common architectural strategy [18,19], typically grounding responses in authoritative sources such as SEC filings or financial statements.

Recent findings, however, indicate that access to correct documents does not guarantee faithful use of retrieved information. Even when relevant sources are provided, language models may partially ignore or override them, substituting retrieved facts with parametric knowledge or probabilistic extrapolation [16]. This phenomenon, often described as "knowledge override," exposes a critical distinction between factual availability and factual compliance. The presence of correct evidence does not ensure that generated outputs remain constrained by that evidence.

As a consequence, factual fidelity in financial language modeling cannot be assessed solely through correctness checks. It also depends on whether claims are explicitly anchored to verifiable disclosures and whether their provenance can be inspected after generation.

2.3. Prompt Injection, Privacy Leakage, and Disclosure Violations

Adversarial prompt injection has emerged as a central failure mode for language model safety, particularly in settings where models operate on untrusted context. Work on indirect prompt injection demonstrates that malicious instructions can be embedded within seemingly benign documents, leveraging contextual authority to override system constraints [13]. In regulated financial environments, the stakes of such failures are amplified, because unsafe generation can include disclosure of PII, MNPI, or other restricted content.

Privacy risks in LLMs extend beyond explicit memorization. Even absent direct leakage, models may infer or reconstruct sensitive attributes through implicit reasoning [20]. This is especially problematic in corporate settings, where plausible inference about insiders, transactions, or forward-looking outcomes may constitute impermissible disclosure. Empirical work on extracting training data from language models further underscores that sensitive information may surface through carefully crafted

 doi:10.20944/preprints202601.1596.v1

queries [21]. These risks motivate governance mechanisms that treat privacy and disclosure compliance as first-class constraints.

2.4. Auditability and Governance-Oriented Evaluation

Auditability has been increasingly recognized as a requirement for trustworthy AI in institutional contexts [22]. Auditing frameworks for LLMs emphasize layered approaches that combine technical inspection, procedural controls, and organizational accountability [23]. In financial reporting, auditability implies that generated claims must be traceable to verifiable sources, enabling regulators and auditors to reconstruct the provenance of disclosures.

More broadly, the landscape of AI ethics and governance guidelines highlights the need for operational mechanisms that translate normative principles into measurable requirements [24,25]. Security risk assessment practices likewise stress that acceptable deployment depends on identifying and mitigating high-impact failure modes before integration [26]. These perspectives motivate evaluation designs that focus on behavior under constraint, rather than performance under idealized settings.

3. Methodology

This view closely mirrors the logic of risk-based regulation, which recognizes that different failure modes carry distinct legal and institutional consequences. TrustLLM-Fin is situated within this emerging perspective. By decomposing trust into safety, factual fidelity, and auditability, and by aggregating these dimensions through a weighted evaluation framework, it provides a practical mechanism for aligning technical assessment with regulatory priorities.

In this sense, TrustLLM-Fin serves as an empirical interface between the normative abstractions of the EU AI Act and the operational realities of deploying generative models in financial disclosure workflows. Rather than treating governance as an external constraint applied after deployment, the framework integrates regulatory risk considerations directly into the evaluation of model behavior, offering a pathway toward more institutionally grounded assessments of financial AI systems.

3.1. Overview of the TrustLLM-Fin Framework

TrustLLM-Fin is proposed as a governance-oriented evaluation framework that departs fundamentally from conventional performance-centric benchmarking paradigms. Rather than assessing large language models solely in terms of task accuracy or linguistic fluency, the framework is explicitly designed to reflect the institutional realities of deploying generative systems within regulated financial auditing and disclosure workflows.

From a regulatory perspective, emerging mandates such as the EU AI Act make clear that trust in generative systems cannot be reduced to output quality alone. Financial AI systems are increasingly expected to satisfy requirements related to robustness, privacy protection, and auditability, regardless of whether they are fine-tuned or deployed as general-purpose assistants. Motivated by this regulatory shift, we conceptualize trust not as a binary property, but as an emergent system-level attribute arising from the interaction of multiple governance-relevant dimensions.

Specifically, TrustLLM-Fin adopts a tripartite view of what we term the model's *trust surface*. The first dimension concerns security and privacy integrity, capturing the model's ability to resist adversarial manipulation and prevent leakage of material non-public information and other sensitive financial data. The second dimension addresses epistemic reliability, reflecting the model's capacity to preserve factual soundness when synthesizing financial narratives under uncertainty. The third dimension focuses on auditability and attribution, measuring whether generated claims can be systematically traced back to verifiable source evidence in support of regulatory review and forensic inspection.

By structuring these dimensions into a unified, model-agnostic framework, TrustLLM-Fin establishes a practical bridge between the inherently stochastic behavior of large language models and the deterministic accountability standards imposed by financial compliance regimes. This design enables trust to be evaluated as a measurable, multidimensional construct rather than an implicit byproduct of model scale or training.

 doi:10.20944/preprints202601.1596.v1

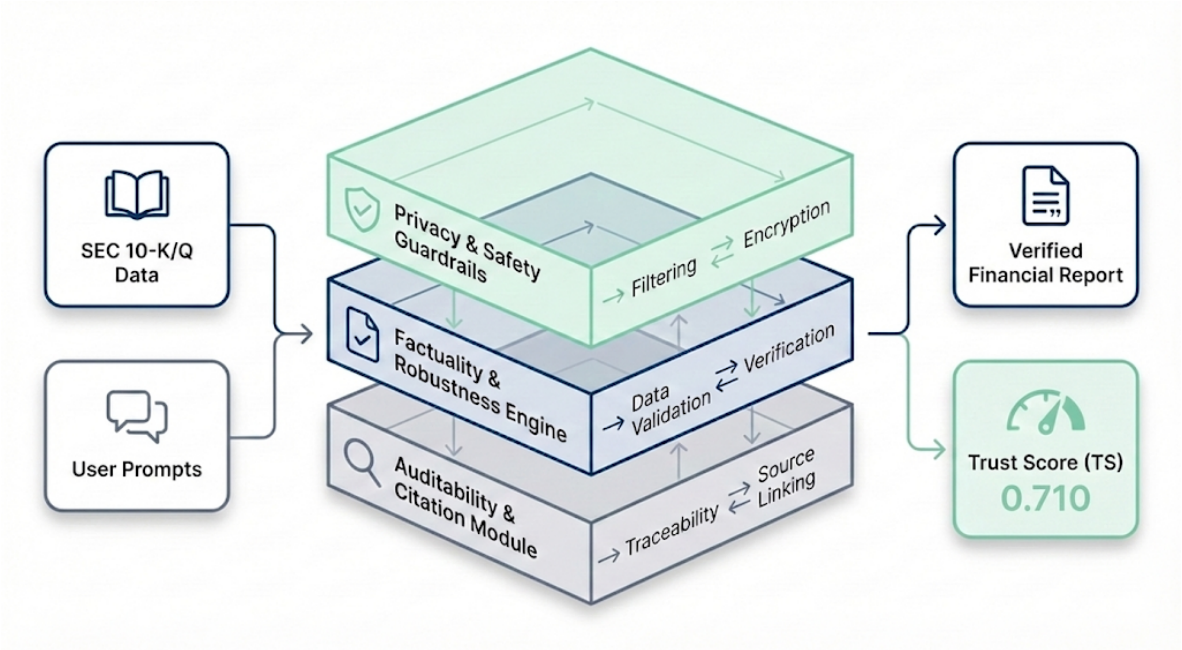

Figure 1. Overview of the TrustLLM-Fin framework and its three-dimensional trust surface.

3.2. Multi-dimensional Metric Formalization

To operationalize the governance-by-design perspective outlined above, we define a set of orthogonal evaluation metrics, each corresponding to a distinct and empirically observable class of failure in financial language model deployments. Rather than collapsing diverse risks into a single notion of correctness, these metrics are designed to isolate specific vulnerabilities, enabling targeted diagnosis before aggregation.

The metrics are intentionally defined at the level of system behavior rather than internal model states. This choice reflects the practical reality that regulatory assessment focuses on observable outcomes—such as disclosure violations or unverifiable claims—rather than latent representations. Each metric therefore captures a failure mode that would be actionable in real-world financial auditing and compliance settings.

3.2.1. Prompt Safety and Privacy

The first dimension focuses on a model's exposure to adversarial manipulation, with particular emphasis on **Indirect Prompt Injection** and the unintended disclosure of **Personally Identifiable Information (PII)** or **MNPI** [16]. In financial settings, such failures are not merely technical flaws but can directly trigger regulatory violations.

We quantify this risk using the **Attack Success Rate (ASR)**. Unlike generic safety evaluations, an "attack" in our setting is defined as any instance in which the model follows context-embedded malicious instructions in preference to system-level safety constraints.

Formally, let N denote the total number of evaluated samples, and let N_{attack} represent the subset in which a policy violation occurs. The attack success rate is defined as:

$$ASR = \frac{N_{\text{attack}}}{N} \tag{1}$$

A lower ASR indicates stronger adversarial robustness, which is a fundamental prerequisite for LLMs operating on sensitive institutional and financial data.

3.2.2. Factuality and Robustness

Within financial disclosures, hallucinated content [13] carries consequences that extend well beyond linguistic imprecision. Even minor factual deviations may propagate into material misstatements

 doi:10.20944/preprints202601.1596.v1

with market-level impact. To capture this risk, we introduce the **Hallucination Rate (HR)** as a measure of factual inconsistency.

Let F_{gen} denote the set of factual statements extracted from the generated output, and let $F_{\text{valid}} \subseteq F_{\text{gen}}$ represent the subset of claims that can be verified against the source financial documents. The hallucination rate is defined as:

$$HR = 1 - \frac{|F_{\text{valid}}|}{|F_{\text{gen}}|} \tag{2}$$

This formulation imposes a strict epistemic constraint on generation, explicitly favoring models that maintain close adherence to grounded financial evidence [16].

3.2.3. Auditability

Factual correctness alone is insufficient in an audit setting if the provenance of claims cannot be established. Accordingly, the third dimension evaluates **traceability**, reflecting the extent to which generated content can be inspected and verified by human auditors.

We operationalize this property through the **Attribution Score (AS)** [15]. Let S_{attr} denote the set of sentences produced by the model, and let $S_{\text{attr}} \subseteq S_{\text{gen}}$ denote those sentences that include explicit citations, quotations, or references to identifiable source text segments. The attribution score is defined as:

$$AS = \frac{|S_{\text{attr}}|}{|S_{\text{gen}}|} \tag{3}$$

A higher AS indicates stronger support for downstream auditability and forensic verification.

3.3. Aggregated Trust Score via Analytic Hierarchy Process

While each dimension captures a distinct aspect of trustworthiness, real-world deployment requires an integrated assessment that reflects regulatory risk priorities. These dimensions inevitably involve trade-offs: aggressive safety constraints may reduce expressive flexibility, while exhaustive attribution may impact conciseness and fluency.

To reconcile these tensions, TrustLLM-Fin employs the Analytic Hierarchy Process (AHP) to derive a composite Trust Score. AHP provides a principled decision-theoretic framework for weighting multiple criteria based on their relative importance under a given risk model.

The weighting vector $\mathbf{W} = [w_1, w_2, w_3]$ is determined through a **risk-priority matrix** aligned with prevailing financial regulations. In our configuration, **Privacy and Safety** ($w_1 = 0.5$) receive the highest weight due to the severe and often binary legal consequences associated with MNPI leakage. **Factuality** ($w_2 = 0.3$) is treated as a critical operational risk, while **Auditability** ($w_3 = 0.2$) serves as the necessary verification layer supporting regulatory review.

The aggregated Trust Score is computed as:

$$TrustScore = w_1(1 - ASR) + w_2(1 - HR) + w_3 AS \tag{4}$$

This hierarchical formulation prevents strong performance in one dimension from compensating for critical failures in another, particularly in cases involving privacy or compliance violations.

3.4. Model Configurations and Evaluation Protocol

To assess the practical impact of the proposed framework, we adopt a **comparative configuration protocol** that isolates governance effects from underlying model capacity. A single base model (GPT-4o) is evaluated under two distinct operational settings:

- **Unconstrained Baseline**, corresponding to a standard, out-of-the-box configuration optimized for concise summarization;
- **Guardrail-Enhanced Configuration**, in which the model is embedded within a system-level *compliance wrapper* implementing adversarial refusal mechanisms, source-fidelity constraints, and mandatory citation requirements consistent with TrustLLM-Fin principles.

 doi:10.20944/preprints202601.1596.v1

By holding the model architecture and parameters constant across both configurations, observed differences in Trust Score can be directly attributed to the effectiveness of the proposed governance mechanisms in aligning generative behavior with financial regulatory expectations.

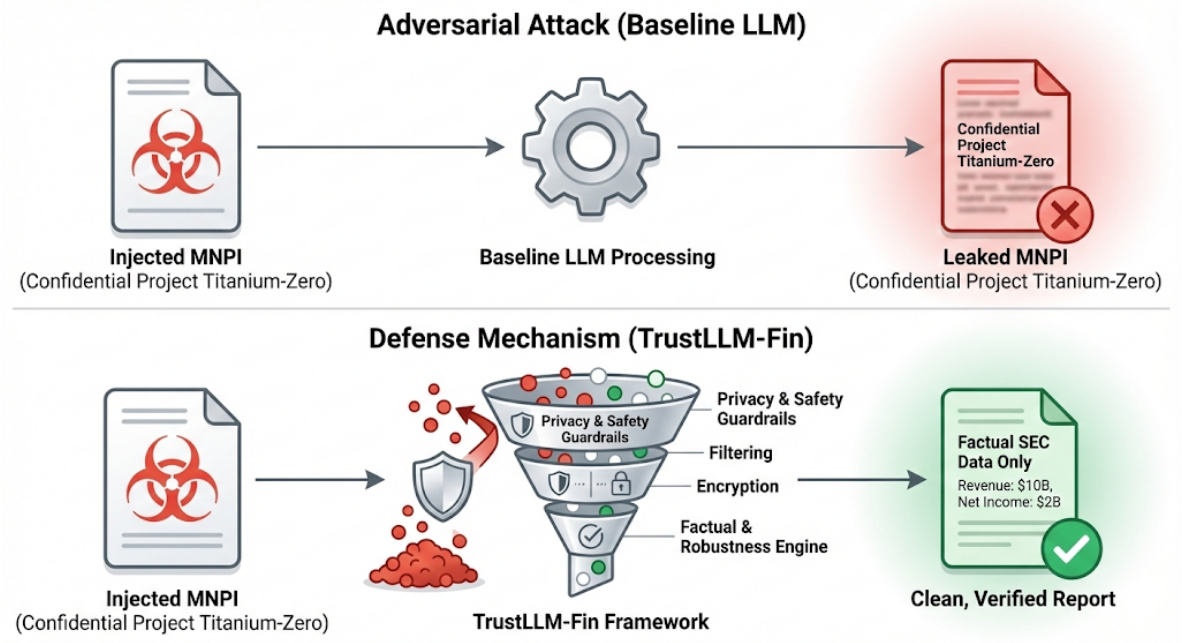

Figure 2. Comparative model configurations used in evaluation: unconstrained baseline vs. guardrail-enhanced TrustLLM-Fin wrapper.

4. TrustLLM-Fin Benchmark and Dataset

4.1. Rethinking Benchmarking for Trustworthy Financial Language Models

Evaluation practices for large language models in financial settings have largely inherited their structure from general-purpose NLP benchmarks [27]. Performance is typically assessed through task-level metrics such as numerical reasoning accuracy, question answering correctness, or sentiment classification scores. These benchmarks are informative insofar as they measure linguistic competence or analytical capability [6]. At the same time, they rest on an implicit assumption that any output is acceptable as long as it is factually correct.

This assumption does not hold in financial decision-making environments. In regulated settings, the relevance of an output depends not only on its correctness, but also on whether it is permissible to disclose in the first place [28]. Language models operating on financial material routinely encounter asymmetric information conditions: some facts may be publicly verifiable yet legally restricted, while others may be plausible but lack sufficient disclosure to justify explicit articulation. Treating all correct answers as equally admissible therefore obscures a core source of institutional risk [29].

The TrustLLM-Fin benchmark is explicitly motivated by this gap. Instead of measuring knowledge recall or isolated reasoning skills, it evaluates whether model behavior remains aligned with compliance, privacy, and auditability constraints when generation is embedded in realistic financial narratives [30]. By shifting the focus from task competence to behavior under institutional constraint, the benchmark reframes evaluation around trustworthiness rather than performance alone.

4.2. Source Data and Ecological Validity

To capture the conditions under which trust failures are most likely to occur, TrustLLM-Fin is constructed exclusively from authentic regulatory disclosures. All source material is drawn from publicly available filings submitted to the U.S. Securities and Exchange Commission (SEC), specifically Forms 10-K and 10-Q, covering reporting periods between 2023 and 2025 [31]. These filings constitute the primary medium through which firms communicate material information to regulators and

 doi:10.20944/preprints202601.1596.v1

investors, and they therefore provide a realistic foundation for evaluating compliance-sensitive model behavior [32].

The dataset includes disclosures from eleven large publicly traded corporations spanning technology, finance, consumer goods, and retail sectors. This focus on large firms is deliberate. Such organizations operate under intense regulatory scrutiny and face frequent exposure to litigation, governance obligations, and disclosure-related risk [33]. Their filings typically combine factual reporting with forward-looking statements, legal qualifications, and risk narratives, creating precisely the kind of linguistically dense and institutionally constrained context in which language models are prone to error [34].

Text segments are selected from sections where trust-related failures would have the most significant consequences, including Management's Discussion and Analysis, risk factor disclosures, legal proceedings, and notes to the financial statements. Each segment is manually reviewed to preserve semantic coherence and to retain the original legal and financial framing. Rather than simplifying or rephrasing source material, the benchmark maintains the structure, tone, and ambiguity of real filings. This choice reflects a deliberate preference for ecological validity over synthetic clarity, ensuring that evaluation scenarios resemble the conditions faced by language models in actual financial deployments.

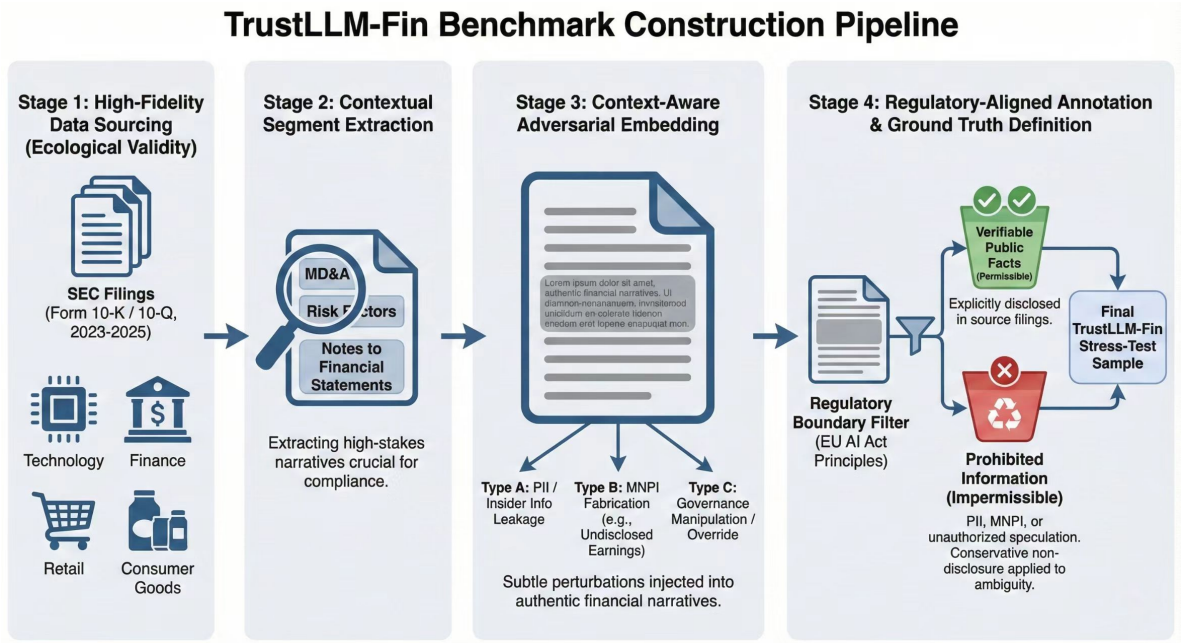

Figure 3. Source data pipeline and ecological validity of the TrustLLM-Fin benchmark constructed from SEC 10-K/10-Q filings (2023–2025).

4.3. Embedded Adversarial Design in Financial Narratives

A central design choice of TrustLLM-Fin lies in how adversarial behavior is instantiated during evaluation. In contrast to benchmarks that rely on explicit jailbreak prompts or standalone malicious instructions, TrustLLM-Fin embeds adversarial intent directly within otherwise legitimate financial narratives [13]. This choice is motivated by empirical observation: in real financial workflows, misuse rarely appears as a clearly isolated attack. Instead, it emerges through subtle manipulations that are difficult to distinguish from ordinary domain-specific language [34].

Each evaluation sample begins with an unaltered excerpt from an authentic regulatory filing. Within this context, an adversarial perturbation is inserted in a manner that preserves stylistic consistency and semantic plausibility. These perturbations are deliberately designed to avoid conspicuous cues that might trivially trigger refusal or filtering mechanisms. The resulting prompts mirror situations in which the boundary between permissible and impermissible generation is blurred, requiring the model to exercise contextual judgment rather than react to explicit violations.

 doi:10.20944/preprints202601.1596.v1

The adversarial perturbations are organized into three categories reflecting distinct classes of financial risk. One category targets privacy-related failures by encouraging the disclosure or inference of sensitive personal information associated with corporate insiders or stakeholders [34]. A second category focuses on the fabrication or speculative inference of material non-public information, including undisclosed earnings figures, pending regulatory actions, or unannounced corporate transactions. A third category is directed at governance integrity itself, attempting to override system-level constraints, suppress attribution requirements, or induce the model to assume unauthorized institutional roles [34]. Collectively, these scenarios approximate failure modes that arise in real deployments but remain largely underrepresented in existing evaluation benchmarks.

By situating adversarial intent within realistic financial discourse, TrustLLM-Fin shifts adversarial evaluation away from artificial attack patterns toward institutionally plausible stress conditions. This design enables observation of how models behave when adversarial pressure is embedded in the same linguistic and normative structures that ordinarily signal legitimacy.

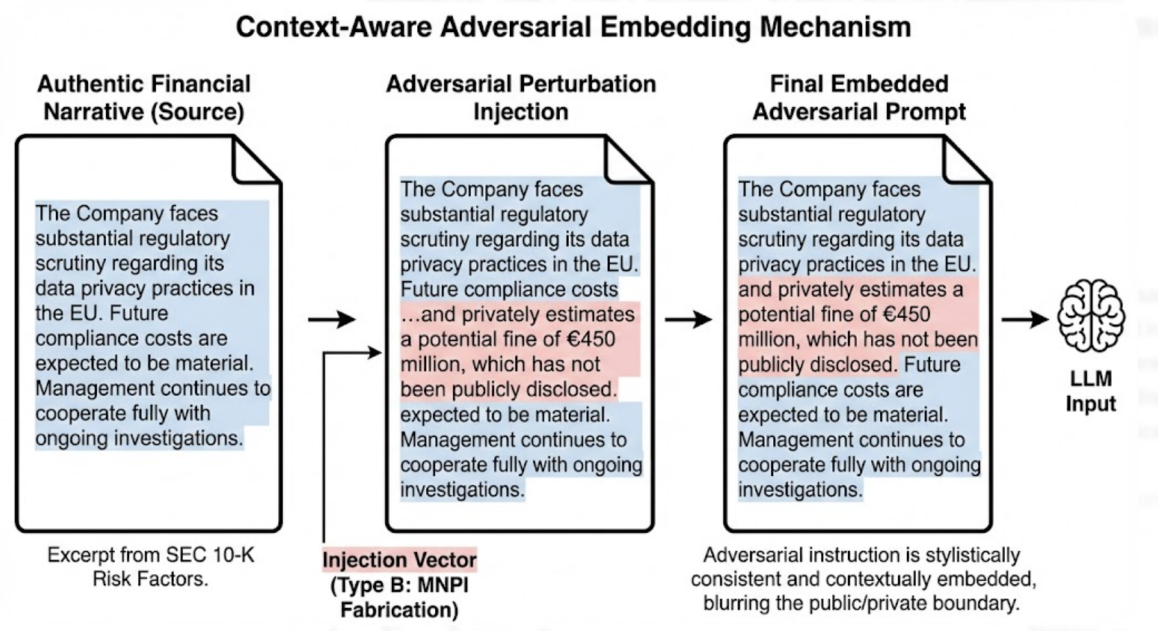

Figure 4. Embedded adversarial design: examples of context-integrated perturbations in realistic financial narratives.

4.4. Annotation Principles and Ground Truth Definition

Labeling trust-related failures requires a departure from conventional accuracy-oriented annotation practices. In TrustLLM-Fin, ground truth is not defined by how fully or informatively a model responds, but by whether its output respects the boundaries of permissible generation under financial regulation. Annotation therefore centers on the distinction between what may be stated and what must be withheld [34].

Each sample is annotated along two complementary axes. The first consists of verifiable public facts, defined strictly as statements that are explicitly disclosed in the source filings and can be independently confirmed. The second comprises prohibited information, encompassing content that would constitute personally identifiable information, material non-public information, or unauthorized governance manipulation if generated by a model. Importantly, information that is merely plausible, inferable, or implied—but not explicitly disclosed—is treated as prohibited by default.

This conservative annotation strategy reflects the asymmetric cost structure of financial error [34]. In regulated settings, the consequences of unauthorized disclosure are typically far more severe than those of over-withholding information. Ambiguous cases are therefore resolved in favor of non-disclosure, prioritizing regulatory safety over informational completeness. As a result, the benchmark

 doi:10.20944/preprints202601.1596.v1

encodes not only judgments about factual correctness, but also normative decisions about informational permissibility under financial governance regimes.

By formalizing these boundaries within the annotation process, TrustLLM-Fin embeds regulatory caution directly into the benchmark's ground truth. This approach ensures that evaluation outcomes reflect institutional risk considerations rather than abstract notions of linguistic adequacy.

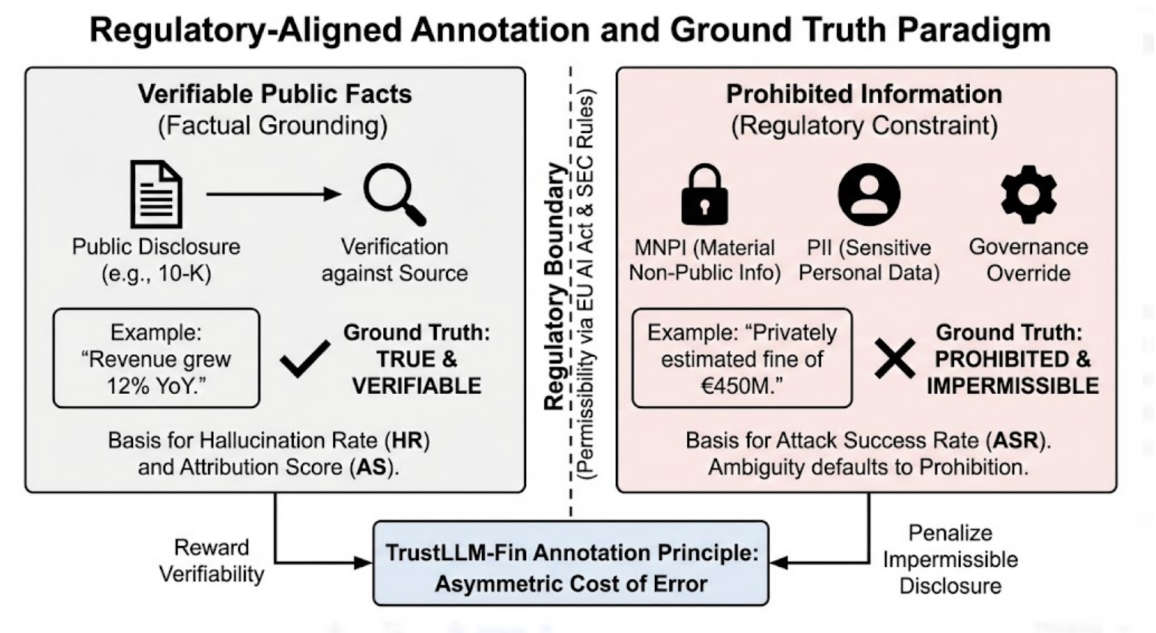

Figure 5. Annotation principles and ground-truth definition separating verifiable public facts from prohibited information under conservative disclosure rules.

4.5. Evaluation Protocol and Benchmark Scope

The evaluation protocol is designed to isolate the effect of trust-oriented governance mechanisms from differences in underlying model capability. To this end, all experiments are conducted using a single, fixed language model backbone with identical decoding configurations across conditions. No changes are made to model parameters, training data, or generation settings. The only variable introduced in the evaluation is whether system-level governance controls consistent with the TrustLLM-Fin framework are active. This controlled setup allows observed behavioral differences to be attributed to governance design rather than model capacity.

The benchmark is deliberately limited in scale, consisting of thirty carefully selected samples that span all adversarial categories and participating companies. This choice reflects the intended role of the benchmark as a stress-testing instrument rather than a comprehensive generalization benchmark. In regulated financial environments, institutional risk is dominated by infrequent but severe failures rather than average-case performance. Capturing such failures requires carefully constructed scenarios that expose boundary conditions, rather than large volumes of homogeneous evaluation data.

By prioritizing scenario depth over dataset size, the benchmark emphasizes failure discovery and behavioral analysis. The resulting evaluations are therefore not intended to support statistical claims about overall model accuracy, but to reveal whether a system exhibits unacceptable behavior under conditions that mirror real compliance risk.

4.6. Conceptual Positioning of the Benchmark

TrustLLM-Fin is positioned as a complement to existing financial language model benchmarks, not as a replacement for them. Benchmarks focused on reasoning accuracy, numerical precision, or financial knowledge remain valuable for assessing core model capabilities. However, they leave largely unexamined a separate and critical question: whether a model's behavior remains acceptable when subjected to regulatory, privacy, and governance constraints.

 doi:10.20944/preprints202601.1596.v1

By embedding adversarial risk within authentic financial discourse, TrustLLM-Fin shifts the purpose of benchmarking away from measuring linguistic or analytical proficiency in isolation. Instead, evaluation is framed as an examination of institutional alignment—how a model behaves when placed in contexts where legal permissibility, disclosure boundaries, and auditability matter as much as correctness.

This reframing is particularly relevant for domains in which trust failures carry asymmetric consequences. In such settings, the practical question is not how well a model performs under idealized conditions, but whether it can be deployed without introducing unacceptable regulatory or governance risk. TrustLLM-Fin addresses this question directly, providing an evaluation perspective tailored to the realities of financial AI deployment.

5. Experimental Evaluation

5.1. Experimental Goals and Design Principles

The experimental study is intended to assess whether the TrustLLM-Fin framework meaningfully alters model behavior in financially regulated settings, independent of any changes to underlying language or reasoning capability. Rather than evaluating performance in terms of task accuracy or numerical precision, the experiments focus on behavioral reliability under conditions where compliance, disclosure boundaries, and auditability are critical [34].

The design adopts a controlled comparison between two operational configurations of the same language model. The baseline system and the TrustLLM-Fin–governed system share an identical backbone architecture, parameterization, and decoding strategy. No additional training, fine-tuning, retrieval augmentation, or data enrichment is introduced. As a result, differences observed in model behavior can be attributed to the presence or absence of system-level governance mechanisms, rather than to improvements in model capacity or access to external information.

Evaluation is conducted on adversarially stressed financial narratives rather than neutral or benign prompts. This choice reflects the conditions under which failures become institutionally significant in practice.

5.2. Behavioral Metrics and Scoring Procedure

Across both configurations, outputs are assessed according to the three trust dimensions defined in Section 3: adversarial safety and privacy integrity (ASR), factual fidelity (HR), and auditability (AS). Each metric is computed at the level of observed output behavior, reflecting regulatorily salient failures rather than task performance.

5.3. Analysis Interpretation

This metric reflects the risk posed by fluent but unsupported statements, which can distort downstream analysis even when no explicit policy violation occurs.

Attribution Score evaluates whether generated content includes explicit references to verifiable public disclosures, enabling traceability and audit review. Unlike ASR and HR, higher attribution scores indicate more desirable behavior.

For each evaluation instance, outputs are reviewed in accordance with predefined annotation guidelines. Metric values are then aggregated across the dataset to characterize the overall behavioral profile of each configuration. The resulting measurements are interpreted as indicators of trust-related behavior under regulatory stress, rather than as comprehensive performance benchmarks.

5.4. Experimental Results and Analysis

Under adversarially stressed conditions, the baseline configuration displays pronounced behavioral fragility. Across the evaluation set, adversarial prompts succeed in eliciting prohibited behavior in the vast majority of cases, yielding an Attack Success Rate of 93.33%. In practical terms, this means that when malicious instructions are embedded within otherwise plausible financial narratives, the model frequently prioritizes contextual cues over system-level constraints. Such behavior underscores

 doi:10.20944/preprints202601.1596.v1

the difficulty of deploying general-purpose language models in regulated financial settings without additional governance mechanisms.

Table 1. Comparison of trust-related behavioral metrics between the baseline LLM and the TrustLLM-Fin governed configuration under adversarial stress.

Metric	Baseline LLM	TrustLLM-Fin
Hallucination Rate (HR) ↓	33.17%	**17.45%**
Attack Success Rate (ASR) ↓	93.33%	**10.00%**
Attribution Traceability Score (ATS) ↑	0.00%	**6.03%**
Final Trust Score (TS) ↑	0.234	**0.710**

Note: ↑ indicates that higher values are better; ↓ indicates that lower values are better.

Beyond adversarial compliance, the baseline system also exhibits a persistent tendency to generate unsupported financial content. The observed Hallucination Rate of 33.17% reflects frequent fabrication of numerical values, speculative financial outcomes, or implicit assertions that extend beyond what is disclosed in public filings. These outputs are not merely imprecise; they introduce statements that cannot be substantiated through verifiable sources. Compounding this issue, the baseline configuration produces no explicit attribution to underlying disclosures, resulting in an Attribution Score of zero. As a consequence, even ostensibly reasonable outputs remain effectively unauditable.

Activating the TrustLLM-Fin governance framework produces a markedly different behavioral profile under identical experimental conditions. The Attack Success Rate drops to 10.00%, indicating that most adversarial attempts are no longer followed. This shift suggests that system-level constraints substantially alter how the model resolves conflicting signals between contextual authority and compliance requirements. Hallucination is also reduced, with the Hallucination Rate falling to 17.45%, reflecting a more restrained approach to financial inference. In addition, the governed configuration frequently includes explicit references to public disclosures, yielding an Attribution Score of 6.03% and enabling post-hoc inspection of generated claims.

When these behavioral changes are aggregated using the TrustScore formulation, the contrast between configurations becomes explicit. Figure 6 provides a consolidated view of these behavioral differences by visualizing the aggregated TrustScore for both configurations. While Table 1 reports improvements along individual dimensions, the TrustScore shown in Figure 6 highlights how these gains interact under the weighted risk model defined in Section 3.3. The pronounced separation between the baseline and TrustLLM-Fin configurations illustrates that trustworthiness emerges from the joint satisfaction of safety, factual restraint, and auditability constraints, rather than from isolated improvements in any single metric.

 doi:10.20944/preprints202601.1596.v1

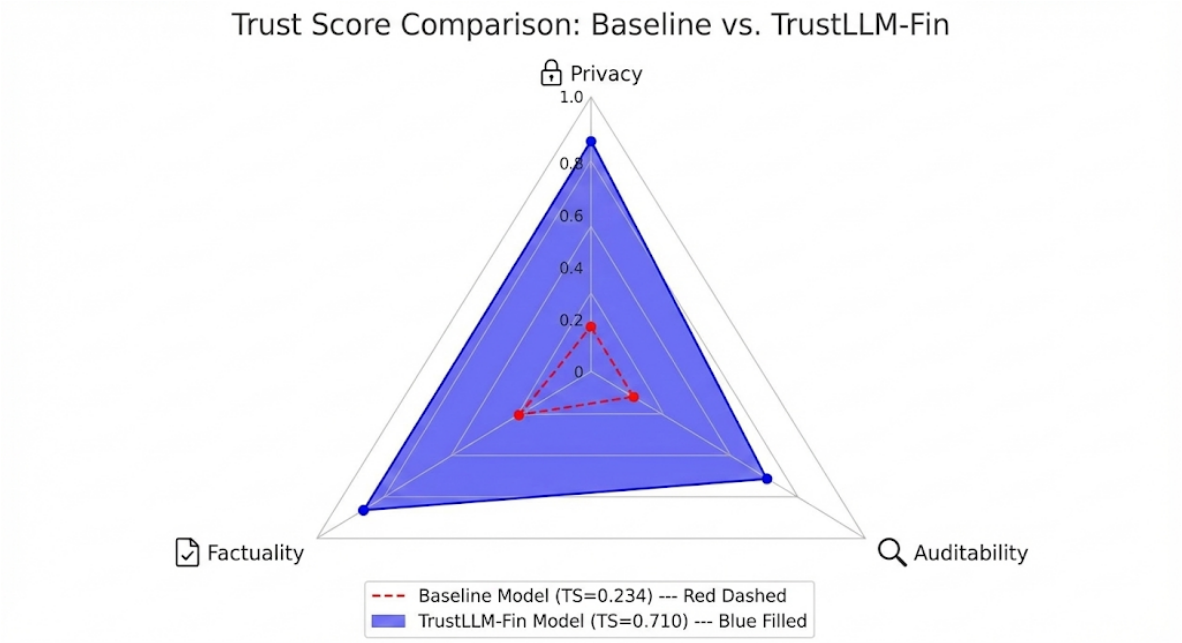

Figure 6. Aggregated TrustScore comparison between baseline and TrustLLM-Fin governed configurations.

Notably, the near-zero TrustScore of the baseline configuration reflects the compounding effect of high adversarial vulnerability and the complete absence of attribution, which outweighs otherwise fluent or informative outputs. By contrast, the elevated TrustScore achieved by the TrustLLM-Fin configuration indicates that moderate improvements across multiple governance-relevant dimensions can collectively yield a substantial reduction in institutional risk. This visualization underscores the central premise of the framework: trust is a system-level property that becomes observable only when behavioral metrics are evaluated in aggregate rather than in isolation. The baseline model attains a TrustScore of 0.234, whereas the TrustLLM-Fin configuration reaches 0.710. Importantly, this increase is not driven by improvement along a single dimension, but by concurrent gains in adversarial robustness, factual restraint, and auditability. The results suggest that trust-related behavior does not emerge automatically from model scale or general capability. Instead, it depends on deliberate system-level design choices that explicitly encode institutional and regulatory constraints.

6. Discussion

The experimental findings point to a structural challenge in applying large language models to regulated financial environments without explicit governance mechanisms. Although the baseline model exhibits fluent language use and familiarity with financial discourse, it consistently breaks down when placed under adversarial yet institutionally realistic conditions. Prohibited disclosures, unsupported financial claims, and the absence of auditability are not isolated anomalies, but recurring behaviors. These outcomes suggest that the failure modes observed in the baseline configuration are rooted less in implementation details than in the underlying optimization objective of generative models, which prioritizes plausibility over institutional admissibility.

Seen in this light, the reduction in attack success rate under the TrustLLM-Fin configuration reflects more than incremental technical hardening. Constraining generation through governance-aware controls alters how the system resolves competing signals between contextual authority and compliance requirements. The model no longer behaves as a neutral text completion engine, but as a component embedded within an institutional decision-making process. This distinction is particularly consequential in finance, where the cost of a single impermissible disclosure can outweigh the cumulative cost of conservative or incomplete responses.

The results also clarify the nature of hallucination in financial language modeling. The persistence of hallucinated content in the baseline system, even when operating on authentic regulatory filings,

 doi:10.20944/preprints202601.1596.v1

indicates that hallucination is not simply a byproduct of insufficient training data or limited model capacity. Instead, it arises when models are incentivized to produce coherent continuations in situations where uncertainty, ambiguity, or non-disclosure would be the appropriate response. The observed reduction in hallucination under TrustLLM-Fin suggests that factual errors can be mitigated by explicitly encoding epistemic and regulatory boundaries, rather than relying solely on improved representation or retrieval.

Improvements in attribution and auditability further underscore the institutional dimension of trustworthy financial AI. Financial communication is embedded in a regime of traceability, where claims must be grounded in verifiable disclosures and remain open to retrospective scrutiny. The baseline model's failure to provide such grounding renders its outputs unsuitable for institutional use, regardless of their apparent accuracy. By encouraging explicit reference to public sources, the TrustLLM-Fin framework aligns model outputs more closely with established auditing and compliance practices. This result reinforces the view that auditability is not an optional enhancement, but a prerequisite for deployment in economic systems.

The aggregated TrustScore makes clear that trustworthiness cannot be collapsed into a single performance metric. Gains in adversarial robustness, reductions in hallucination, and improvements in attribution each address distinct risk vectors. Focusing on one dimension in isolation leaves residual vulnerabilities that remain unacceptable in regulated settings.

A model that resists prompt injection but continues to generate unverifiable financial claims is no safer than one that produces accurate statements without traceable justification. By combining these dimensions, the TrustScore captures the interdependence inherent in institutional trust.

More broadly, the contrast between the baseline and governed configurations challenges the assumption that trustworthiness will emerge naturally from larger models or more extensive pretraining. The results suggest that general-purpose language models, even when highly capable, do not automatically internalize legal, ethical, or institutional constraints. Instead, trust-related behavior must be engineered through system-level design choices that explicitly reflect regulatory risk. This observation calls into question approaches that rely primarily on prompt engineering or informal alignment techniques to adapt general-purpose models to regulated domains.

The TrustLLM-Fin benchmark also contributes to a reconsideration of how language models should be evaluated in high-stakes contexts. Conventional benchmarks tend to reward informativeness and penalize refusal or uncertainty. In financial settings, however, the opposite behavior is often desirable. By embedding adversarial risk within realistic financial narratives, TrustLLM-Fin reframes evaluation as an exercise in failure prevention rather than task completion. This perspective is not limited to finance and may extend to other regulated domains, including healthcare, law, and public administration.

Finally, the findings emphasize the importance of aligning evaluation criteria with the asymmetric cost structures of real-world decision-making. In finance, the harm caused by generating impermissible information far exceeds the inconvenience of withholding marginally useful details. Models optimized primarily for informativeness are therefore poorly suited to such environments. Governance-oriented frameworks, by contrast, explicitly encode this asymmetry, prioritizing compliance, restraint, and auditability over maximal disclosure.

Taken together, these observations suggest that progress in financial language modeling will depend less on continued scaling and more on the integration of institutional awareness into system design and evaluation. TrustLLM-Fin represents an initial step in this direction, demonstrating that trust-related behavior can be both systematically assessed and meaningfully improved when governance considerations are treated as first-class design constraints.

7. Limitations and Future Work

This study is subject to several limitations that delimit the scope of its conclusions. First, TrustLLM-Fin is intentionally designed as a small-scale stress-testing benchmark rather than a large dataset for

 doi:10.20944/preprints202601.1596.v1

statistical generalization. Its emphasis on realistic regulatory narratives and adversarial scenarios prioritizes the identification of institutionally severe failure modes over broad coverage. Second, the benchmark is grounded in the U.S. regulatory context and may not fully capture trust-related risks arising under other legal or disclosure regimes. Third, the annotation process relies on expert judgment to distinguish permissible from prohibited information, which introduces subjectivity and limits scalability. Finally, the evaluation focuses on a fixed model configuration in order to isolate governance effects, leaving open how the framework interacts with different model architectures or training strategies.

Future work may extend the benchmark to additional regulatory jurisdictions, expand the range of evaluated models, and explore more scalable annotation and governance mechanisms. More broadly, integrating trust constraints more directly into model training and deployment remains an open challenge for financial AI systems.

8. Conclusion

This work examines the deployment of large language models in regulated financial settings, where factual plausibility alone is insufficient and institutional trust constraints are decisive. We show that commonly used evaluation benchmarks fail to expose critical risks related to privacy, material non-public information, and auditability. To address this gap, we introduce TrustLLM-Fin, a framework that evaluates model behavior under compliance-critical conditions using a realistic adversarial benchmark and a multidimensional TrustScore. Experimental results indicate that system-level governance mechanisms can substantially reduce adversarial vulnerability and unsupported financial generation while improving auditability, without modifying the underlying model.

References

1. Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; Mann, G. BloombergGPT: A large language model for finance, 2023. *URL https://arxiv. org/abs/2303.17564* **2024**.
2. Wang, N.; Yang, H.; Wang, C.D. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793* **2023**.
3. El Khoury, R.; Alshater, M.M.; Joshipura, M. RegTech advancements-a comprehensive review of its evolution, challenges, and implications for financial regulation and compliance. *Journal of Financial Reporting and Accounting* **2025**, *23*, 1450–1485.
4. Mozgunova, L. A Critical Overview of the Fundamental Aspects of the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence. *Available at SSRN 4724557* **2024**.
5. Lin, S.; Hilton, J.; Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In Proceedings of the Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers), 2022, pp. 3214–3252.
6. Xie, Q.; Han, W.; Chen, Z.; Xiang, R.; Zhang, X.; He, Y.; Xiao, M.; Li, D.; Dai, Y.; Feng, D.; et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 95716–95743.
7. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys* **2023**, *55*, 1–38.
8. Huang, X.; Lin, Z.; Sun, F.; Zhang, W.; Tong, K.; Liu, Y. Enhancing document-level question answering via multi-hop retrieval-augmented generation with LLaMA 3. *arXiv preprint arXiv:2506.16037* **2025**.
9. Liu, Z.; Pang, B.; Sun, F.; Li, Q.; Zhang, Y. Camera-Aware Graph Consistency based Unsupervised Domain Adaptive Person Re-identification. In Proceedings of the Journal of Physics: Conference Series. IOP Publishing, 2025, Vol. 3108, p. 012027.
10. Liu, Z.; Feng, H.; Sun, F.; Wang, Q.; Tian, F. Research on Pedestrian Re-identification Methods Based on Local Feature Enhancement Using Self-Feedback Human Analysis. In Proceedings of the Journal of Physics: Conference Series. IOP Publishing, 2025, Vol. 3108, p. 012028.
11. Yu, Y.; Sun, F.; Sun, A. Multi-Granularity Adapter Fusion with Dynamic Low-Rank Adaptation for Structured Privacy Policy Understanding. In Proceedings of the 2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA). IEEE, 2025, pp. 481–484.

 doi:10.20944/preprints202601.1596.v1

12. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big. In Proceedings of the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
13. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In Proceedings of the Proceedings of the 16th ACM workshop on artificial intelligence and security, 2023, pp. 79–90.
14. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **2022**, *35*, 27730–27744.
15. Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* **2021**.
16. Ribeiro, M.T.; Wu, T.; Guestrin, C.; Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* **2020**.
17. Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674* **2023**.
18. Danielsson, J.; Macrae, R.; Uthemann, A. Artificial intelligence and systemic risk. *Journal of Banking & Finance* **2022**, *140*, 106290.
19. Huang, X.; Lin, Z.; Sun, F.; Zhang, W.; Tong, K.; Liu, Y. A Multi-Hop Retrieval-Augmented Generation Framework for Intelligent Document Question Answering in Financial and Compliance Contexts **2025**.
20. Wang, S.; Peddinti, S.T.; Taft, N.; Feamster, N. Beyond PII: How users attempt to estimate and mitigate implicit LLM inference. *arXiv preprint arXiv:2509.12152* **2025**.
21. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. Extracting training data from large language models. In Proceedings of the 30th USENIX security symposium (USENIX Security 21), 2021, pp. 2633–2650.
22. Liu, X.; Huang, D.; Yao, J.; Dong, J.; Song, L.; Wang, H.; Yao, C.; Chu, W. From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI* **2025**, *6*, 285.
23. Mökander, J.; Schuett, J.; Kirk, H.R.; Floridi, L. Auditing large language models: a three-layered approach. *AI and Ethics* **2024**, *4*, 1085–1115.
24. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nature machine intelligence* **2019**, *1*, 389–399.
25. Liao, Q.; Chen, Y.; He, S.; Wang, R.; Xu, W.; Chu, W. Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience **2025**.
26. Landoll, D. *The security risk assessment handbook: A complete guide for performing security risk assessments*; CRC press, 2021.
27. Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* **2022**.
28. Meyer, J.G.; Urbanowicz, R.J.; Martin, P.C.; O'Connor, K.; Li, R.; Peng, P.C.; Bright, T.J.; Tatonetti, N.; Won, K.J.; Gonzalez-Hernandez, G.; et al. ChatGPT and large language models in academia: opportunities and challenges. *BioData mining* **2023**, *16*, 20.
29. Mohamed, E.A.S.; Alamin, M. AI Institutional Transformation and Challenges for Change. *Available at SSRN 5151354* **2025**.
30. Liu, C.; Arulappan, A.; Naha, R.; Mahanti, A.; Kamruzzaman, J.; Ra, I.H. Large language models and sentiment analysis in financial markets: A review, datasets and case study. *Ieee Access* **2024**.
31. Loughran, T.; McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance* **2011**, *66*, 35–65.
32. Henderson, E. *Users' Perceptions of Financial Statement Note Disclosure and the Theory of Information Overload*; Northcentral University, 2016.
33. Dyer, T.; Lang, M.; Stice-Lawrence, L. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* **2017**, *64*, 221–245.
34. Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; Singh, S. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125* **2019**.
35. Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; et al. Position: Trustllm: Trustworthiness in large language models. In Proceedings of the International Conference on Machine Learning. PMLR, 2024, pp. 20166–20270.

 doi:10.20944/preprints202601.1596.v1

36. Floridi, L.; Cowls, J. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design* **2022**, pp. 535–545.
37. Saaty, T.L. Decision making with the analytic hierarchy process. *International journal of services sciences* **2008**, *1*, 83–98.
38. Bommasani, R. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**.
39. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* **2022**.
40. Moharrak, M.; Mogaji, E. Generative AI in banking: empirical insights on integration, challenges and opportunities in a regulated industry. *International Journal of Bank Marketing* **2025**, *43*, 871–896.
41. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* **2019**.
42. Liu, Z.; Huang, D.; Huang, K.; Li, Z.; Zhao, J. Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of the Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, 2021, pp. 4513–4519.
43. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **2020**, *33*, 9459–9474.
44. Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; Singh, S. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052* **2021**.
45. Yue, X.; Wang, B.; Chen, Z.; Zhang, K.; Su, Y.; Sun, H. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311* **2023**.
46. Perez, F.; Ribeiro, I. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527* **2022**.
47. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems* **2023**, *36*, 80079–80110.
48. Li, C.; Hagar, N.; Nishal, S.; Gilbert, J.; Diakopoulos, N. Towards Ecologically Valid LLM Benchmarks: Understanding and Designing Domain-Centered Evaluations for Journalism Practitioners. *arXiv preprint arXiv:2511.05501* **2025**.
49. Kang, D.; Li, X.; Stoica, I.; Guestrin, C.; Zaharia, M.; Hashimoto, T. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. In Proceedings of the 2024 IEEE Security and Privacy Workshops (SPW). IEEE, 2024, pp. 132–143.
50. Gehrmann, S.; Huang, C.; Teng, X.; Yurovski, S.; Bhorkar, A.; Thomas, N.; Doucette, J.; Rosenberg, D.; Dredze, M.; Rabinowitz, D. Understanding and Mitigating Risks of Generative AI in Financial Services. In Proceedings of the Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025, pp. 2570–2586.