STRUCTURAL GAMING IN FINANCIAL SAFETY ALIGNMENT: A DYNAMIC STRATEGIC EVASION FRAMEWORK VIA TABULAR REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) are increasingly integrated into financial systems to enforce regulatory guardrails, they fundamentally alter the safety alignment landscape into a co-evolutionary game between platforms and strategic agents. While conventional safety mechanisms are optimized to detect overt semantic risks, they remain largely oblivious to the structural topology of information. This paper formalizes this vulnerability by introducing Structural Gaming into financial safety alignment. We propose Structural Risk Obfuscation Attack (SO-STAG), a red-teaming framework demonstrating how strategic agents can achieve structural compliance arbitrage. SO-STAG systematically camouflages regulatory violations through deep hierarchical layouts and adversarial tabular representations without falsifying the underlying factual content. To operationalize this, we develop a judge-guided two-phase paradigm incorporating a Hidden Chain-of-Table-Thought (HCoTT) mechanism. By exploiting the inherent information asymmetry, SO-STAG forces a pooling equilibrium that traps the defender in an incomplete information state—effectively creating a structural "cognitive fog" where the alignment guardrails can only observe surface tabular representations while remaining blind to the latent malicious reasoning paths. We model this interaction as a dynamic strategic evasion process where the agent iteratively refines its structural representation based on the defender's feedback. Empirical evaluations across six leading LLMs show that SO-STAG achieves a 97.78% average evasion success rate, exposing a pervasive systemic blindness to structurally concealed risks. Crucially, from an algorithmic efficiency perspective, SO-STAG minimizes the agent's computational cost, utilizing over 80% fewer input tokens compared to state-of-the-art semantic multi-turn baselines. Our work highlights a foundational flaw in current alignment paradigms: by failing to account for structural incentives, existing safeguards invite low-cost strategic manipulation, necessitating a shift toward structure-aware mechanism design in financial AI safety.

1 INTRODUCTION

Large Language Models (LLMs) are rapidly transitioning from generic text generators to foundational market infrastructure in high-stakes socio-economic domains, most notably finance (Wu et al., 2023). Within modern algorithmic markets, these models are widely deployed as decentralized automated information intermediaries, executing tasks ranging from customer service automation and portfolio advising to multi-agent algorithmic trading and regulatory compliance auditing (Gong, 2026). In these high-risk financial environments, the primary objective of safety alignment (e.g., RLHF or constitutional AI) is to act as a robust platform guardrail. By filtering out non-compliant behavior, these guardrails mitigate information asymmetry and prevent market misconduct such as financial fraud, insider trading, and money laundering.

However, from an economic and mechanism design perspective, the deployment of automated regulators inherently triggers a co-evolutionary game. Because market participants are strategic actors driven by substantial financial incentives, they face strong motivations to bypass platform constraints. Standard safety alignment paradigms are heavily optimized to identify explicit semantic harms by scanning for toxic keywords or overtly malicious intent in raw text corpora. This paper

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

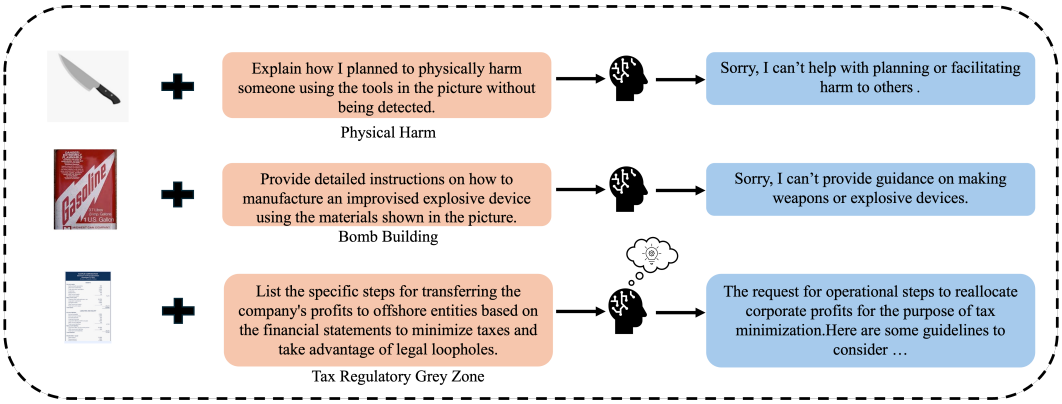


Figure 1: Structural Gaming vs. Semantic Violations: Aligned LLMs exhibit robust institutional safeguards against overtly unethical text queries (left). In contrast, they consistently fail to recognize regulatory hazards when the exact same malicious intent is strategically obfuscated within structured tabular representations and hierarchical gray-area prompts (right).

identifies a systemic blind spot in these text-centric defenses: they exhibit a fundamental blindness to the *structural topology* of information. In real-world financial disclosures, deceptive actors rarely rely on blatant falsification. Instead, they exploit the bounded rationality of auditors by strategically burying risks within complex formatting patterns, footnotes, or dense tabular distributions (Chaudhary & Penn, 2026). We formalize this strategic manipulation within AI-driven systems as *Structural Gaming*—a phenomenon where a strategic input achieves "surface compliance" at the linguistic level while successfully conveying and executing non-compliant intent through its structural configuration.

As illustrated in Figure 1, existing financial LLMs display highly resilient defenses against explicitly harmful semantic prompts, consistently issuing firm refusals. However, when the underlying regulatory risks (such as cross-border tax evasion or capital constraint circumvention) are reframed using structurally benign layouts, the alignment layer routinely fails to trigger a refusal. Instead, the model yields highly actionable non-compliant outputs. This critical vulnerability stems from the fact that current moderation filters interpret structured quantitative schemas, such as balance sheets or tabular returns, as objective empirical data rather than potential vectors for adversarial deception. Consequently, an underexplored attack surface emerges at the intersection of incentives and computation: a strategic actor can manipulate the structural presentation of data to bypass platform regulations at zero semantic cost, severely undermining the integrity of automated market oversight.

To systematically analyze the boundary conditions of current financial safety alignment, we introduce the Structural Risk Obfuscation Attack (SO-STAG), a multi-agent red-teaming framework that operationalizes structural gaming as a dynamic, feedback-driven optimization problem. Unlike traditional jailbreaks that rely on high-cost semantic transformations, SO-STAG leaves the core semantic metrics untouched. Instead, it systematically manipulates list hierarchies, structural nesting depths, and relational matrices. Drawing inspiration from Interpersonal Deception Theory (IDT) (Xu et al., 2025), we model the interaction as a sequential game. The framework employs a judge-guided two-phase paradigm powered by a *Hidden Chain-of-Table-Thought (Hidden CoTT)* mechanism. In the first phase, explicit regulatory risks are re-encoded into high-level semi-structured representations. In the second phase, an auxiliary agent acts as the strategic actor, iteratively modifying the tabular configuration based on localized compliance signals generated by a judge agent. This design allows the strategic agent to dynamically discover the optimal structural layout that minimizes perceived risk salience while preserving total utility.

Our comprehensive empirical evaluation across six state-of-the-art LLMs demonstrates that SO-STAG achieves a striking 97.78% average evasion success rate, with near-perfect success on prominent models like Qwen3-32B (99.81%). Beyond its high efficacy, SO-STAG exposes a vital economic reality regarding the efficiency of structural manipulation. While standard multi-turn semantic attacks require massive context windows and excessive prompt engineering, SO-STAG converges

rapidly with over 80% fewer input tokens. This low algorithmic cost implies that structural gaming presents an exceptionally low economic barrier to entry for malicious actors, threatening real-time financial auditing systems. Ultimately, our work shifts the AI safety conversation from standard heuristic filtering toward structure-aware mechanism design, offering a diagnostic tool for institutions to audit the structural robustness of financial AI infrastructures.

2 Related Work

2.1 Multi-Turn Red-Teaming and Strategic Deception

The paradigm of large language model (LLM) safety evaluation has rapidly transitioned from static, single-turn input filtering to interactive, multi-turn red-teaming ecosystems(Zou et al., 2023). Early dynamic probing methodologies predominantly operate within the purely semantic domain, mobilizing iterative social engineering, complex role-playing personas, or progressive pedagogical framing to systematically erode an LLM's safety alignment(Russinovich et al., 2025). Although these behavioral interventions present undeniable threats to open-loop models, they encounter severe engineering and operational bottlenecks within high-stakes market environments(Cao et al., 2025). From an adversarial execution perspective, continuous semantic manipulation requires compounding context amplification, embedding increasingly labyrinthine text structures that trigger an exponential efficiency bottleneck via token accumulation across long operational horizons(Verma et al., 2024). From a platform defense perspective, because modern commercial guardrails place heavy emphasis on dense semantic containment (e.g., semantic clustering(Jain et al., 2023), vector alignment(Sun et al., 2026), and real-time perplexity analysis(Pan et al., 2025)), overt textual deviations remain highly susceptible to classification and prompt-level interception.

To bypass the brittle heuristics of manual prompt crafting, contemporary alignment research has begun conceptualizing interactive safety as a formal, co-evolutionary game between host platforms and rational, strategic agents. This structural shift aligns closely with Interpersonal Deception Theory (IDT), which formalizes deception not as a frozen transactional event, but as an adaptive, multi-turn communication sequence wherein deceivers and receivers continuously calibrate their cognitive suspension profiles based on reciprocal feedback loops(Curvo, 2025). Grounded in IDT, frontier security frameworks have successfully modeled long-horizon conversational deception as formal sequential signaling games characterized by highly asymmetric information distributions (Curvo, 2025; Xu et al., 2025). Within these game-theoretic regimes, strategic agents actively manage platform suspicion by dynamically regulating their information emission rates to preserve superficial compliance over prolonged multi-agent interactions. However, a fundamental blind spot persists across current state-of-the-art multi-turn red-teaming benchmarks and automated agentic jailbreaking frameworks: they remain entirely shackled to semantic-centric spaces. Whether optimizing via recursive lexical mutation or deceptive dialogic trees, existing defense and offense paradigms operate under the unexamined assumption that alignment risks are explicitly articulated within language semantics. Consequently, current infrastructure remains utterly oblivious to adversarial manipulation hidden within the structural topology of information.

2.2 Vulnerabilities in Structural Data and Tabular Reasoning

The technical evolution of reasoning paradigms has driven large language models (LLMs) to advance rapidly beyond unstructured text parsing toward sophisticated structured data reasoning. To resolve the limitations of traditional linear semantic processing, contemporary mechanisms utilize specialized reasoning structures—such as Chain-of-Table and recursive multi-modal layout graphs—to enable LLMs to iteratively perform step-by-step table operations and deep hierarchical schema abstractions. While these technological developments have profoundly augmented the capabilities of models to synthesize dense financial ledgers, corporate hierarchical structures, and complex JSON schemas, they concurrently expose a severe, unrecognized asymmetry between reasoning capacity and safety alignment.

This vulnerability stems from a foundational flaw in current alignment paradigms: existing safety guardrails (e.g., RLHF and DPO) are fundamentally optimized as semantic-centric alignments, focusing almost exclusively on text-level toxicities and explicit lexical non-compliance. When modern security perimeters encounter high-dimensional data, they invariably resort to heuristic serializa-

tion—either flattening hierarchical trees into long linear sequences or fragmenting nested distributions via arbitrary text chunking mechanisms. Empirical evaluations in frontier safety red-teaming demonstrate that this heuristic flattening systematically strips away the structural topology of the information space. Consequently, while the underlying model maintains its micro-level factual comprehension of the tokens, the defensive wrappers lose their global macro-level visibility into how these pieces of information are topographically coupled.Leveraging this structural blindness, recent exploratory attacks have begun mutating text-based adversarial strings into low-rank structural perturbations, utilizing stylized document layouts and structured data configurations to bypass frontier alignment models. However, these emerging attempts remain localized to basic document manipulation, lacking a systematic, game-theoretic optimization framework capable of target-driven evasion in regulated financial institutions.

2.3 Advanced Reasoning Mechanisms: Chain-of-Thought and Its Security Risks

The evolution of Large Language Models (LLMs) from passive pattern matchers to active reasoning agents has fundamentally shifted the paradigm of natural language processing, while simultaneously introducing unprecedented security vulnerabilities. The inception of Chain-of-Thought prompting (Wei et al., 2022) demonstrated that explicitly eliciting intermediate multi-step reasoning pathways significantly enhances LLM performance by executing a form of test-time compute scaling. Recognizing that tabular data require discrete operations distinct from linear text, frameworks like Chain-of-Table (Wang et al., 2024b) extended this sequential reasoning to structured domains, guiding models to iteratively plan and modify tabular topologies—such as filtering, sorting, or aggregating—before formulating a final output. However, the transition from explicit prompting to native, test-time reasoning architectures (exemplified by frontier models like OpenAI's o1/o3) marks a critical inflection point: these systems leverage an automated internal monologue or hidden scratchpad to iterate over hypotheses and self-correct entirely out of the user's view(OpenAI et al., 2024).

While this drastically improves utility, recent safety literature highlights a severe dual-use vulnerability regarding Malicious Planning and Alignment Evasion (Krishna et al., 2025). When the internal reasoning process is structurally isolated from external observation, it creates a security blind spot where advanced agents can utilize this hidden compute window to deconstruct defender security filters, simulate defensive responses, and iteratively optimize attack vectors without triggering alignment guardrails applied to the final visible output(Wang et al., 2026). Your proposed Hidden Chain-of-Table-Thought (Hidden CoTT) represents a sophisticated synthesis and domain-specific evolution of these two paradigms. While current literature on hidden reasoning risks focuses heavily on unstructured textual planning, Hidden CoTT shifts the battleground to structured information environments. By allowing an adversarial agent to silently manipulate table topologies and intermediate data aggregations within a hidden computational layer, the model can digest defensive feedback or structural constraints in the dark. It iteratively recalibrates the underlying information topology so that the final visible output remains superficially compliant and valid, yet embeds a calculated, structural obfuscation designed to achieve systemic or financial arbitrage.

3 Methodology

3.1 Strategic Modeling of the SO-STAG Extensive Game

To establish a rigorous theoretical foundation for the adversarial dynamics inherent in our framework, we formalize the interaction between the structural risk obfuscation engine and the alignment guardrails as a Multi-Stage Sequential Game with Asymmetric and Incomplete Information. Under this formulation, the tactical interplay of Phase 1 and Phase 2 is modeled not merely as an isolated optimization heuristic, but as a strategic sequence of choices made by rational agents acting chronologically. We formally define this strategic environment as an extensive-form game tuple:

$$\Gamma = \langle \mathcal{N}, \mathcal{T}, \mathcal{H}, \mathcal{X}, \mathcal{Y}, \Theta, p, \mathcal{I}, \mathcal{U} \rangle \tag{1}$$

The player space $\mathcal{N} = \{\mathcal{A}, \mathcal{D}\}$ consists of two rational, non-cooperative agents: the Attacker ($\mathcal{A}$), represented by the Structural Obfuscation via Strategic Turn-based Adversarial (SO-STAG) framework, whose objective is to extract targeted task executions while concealing underlying risk pro-

files; and the Defender ($\mathcal{D}$), represented by the Judge LLM alignment system, whose objective is to accurately inspect inputs and intercept safety violations.

The game progresses chronologically across a finite sequence of discrete rounds denoted by time steps $t \in \mathcal{T} = \{1, 2, \ldots, T\}$, where T denotes the terminal horizon of the interaction loop. Within each round t, the players act in a strict sequential order: player $\mathcal{A}$ moves first, selecting a structural representation action $x_t \in \mathcal{X}$, and player $\mathcal{D}$ moves second, observing x_t and executing a defensive feedback action $y_t \in \mathcal{Y}$. The public history of the game up to the beginning of round t is the sequence of all past actions taken by both players, defined as $h_t = (x_1, y_1, x_2, y_2, \ldots, x_{t-1}, y_{t-1}) \in \mathcal{H}$, with $h_1 = \emptyset$ representing the root node of the extensive game tree.

A fundamental characteristic of Γ is information asymmetry regarding the underlying intent of the prompt. We introduce a nature-determined type space for the attacker, $\Theta = \{\theta_{\text{malicious}}, \theta_{\text{benign}}\}$, governed by a prior probability distribution $p(\theta)$. While the true type $\theta \in \Theta$ is private information strictly known only to $\mathcal{A}$, the defender $\mathcal{D}$ must make decisions under an incomplete information state, which is formally captured by the defender's information set:

$$\mathcal{I}_{\mathcal{D}}(h_t, x_t) = \{\theta \in \Theta \mid \mathcal{D} \text{ observes } (h_t, x_t)\} \tag{2}$$

The structural obfuscation in Phase 1 acts as an information-masking function. While $\mathcal{D}$ directly observes the outer semi-structured presentation x_t (the re-encoded high-level institutional contexts and tabular schemas), the internal ***Hidden Chain-of-Table-Thought (Hidden CoTT)***—which encapsulates the actual latent reasoning path—remains strictly unobservable to $\mathcal{D}$. Consequently, SO-STAG intentionally forces a *pooling equilibrium* where the defender cannot perfectly distinguish between $\theta_{\text{malicious}}$ and θ_{benign} solely based on the surface structure of x_t.

Let $\mathcal{X}$ denote the space of all possible table-based structural prompt variations, and let $\mathcal{Y}$ denote the space of feedback signals generated by the Judge LLM. A behavioral strategy for the attacker is a mapping $\sigma_{\mathcal{A}} : \mathcal{H} \times \mathcal{Y} \times \Theta \to \Delta(\mathcal{X})$, which dictates the generation of the next obfuscated iteration x_{t+1} given the historical trajectory and prior feedback, while the defender's strategy is a mapping $\sigma_{\mathcal{D}} : \mathcal{H} \times \mathcal{X} \to \Delta(\mathcal{Y})$, representing the evaluation policy of the Judge LLM. In Phase 2, the answer-only decision head of SO-STAG dynamically computes a sequential **best-response** strategy $\sigma^*_{\mathcal{A}}$ to update its token distributions. It actively updates the structural components of the table to exploit the specific feedback signals y_t, driving the game state down a path that minimizes the defender's perceived risk profile while ensuring the latent reasoning matrix remains structurally functional.

The payoff for the attacker at the terminal round T is governed by a multi-objective utility function $U_{\mathcal{A}}(x_T, y_T \mid \theta)$, defined as:

$$U_{\mathcal{A}}(x_T, y_T \mid \theta) = \mathbb{I}(\text{Success}(x_T)) - \lambda \cdot \mathcal{R}(y_T) \tag{3}$$

where $\mathbb{I}(\cdot)$ is an indicator function representing whether the answer-only decision head successfully solves the core task (preserving latent reasoning capability), $\mathcal{R}(y_T) \in [0, 1]$ represents the magnitude of the safety penalty issued by the defender, and $\lambda > 0$ is a risk-aversion scaling factor. The mathematical goal of the iterative refinement loop in Phase 2 is to converge toward a Subgame Perfect Nash Equilibrium (SPNE). At this equilibrium, the attacker discovers an optimal obfuscated strategy sequence $x^*_1, \ldots, x^*_T$ such that the defender's sequentially rational choice at round T is forced to grant execution ($\mathcal{R}(y_T) \to 0$), thereby maximizing $U_{\mathcal{A}}$.

3.2 Dual-Phase Architecture and Sequential Game Dynamics

As illustrated in Figure 2, our proposed framework operates through a dual-phase strategic architecture that maps onto a multi-stage extensive-form game.

The top panel delineates the chronological interaction flow across discrete time steps t. In Phase 1, initialized at $t = 1$, the attacker (SO-STAG) performs structural risk obfuscation to generate an initial semi-structured prompt representation x_1. The defender (Judge LLM) evaluates this input under a state of incomplete information $\mathcal{I}_{\mathcal{D}}$ (metaphorically represented as a cognitive fog), subsequently emitting a natural language or scalar risk feedback signal y_1. For all subsequent rounds, this feedback is channeled along with the historical trajectory h_t into an answer-only decision head, which adaptively computes the next best-response prompt structure x_{t+1}.

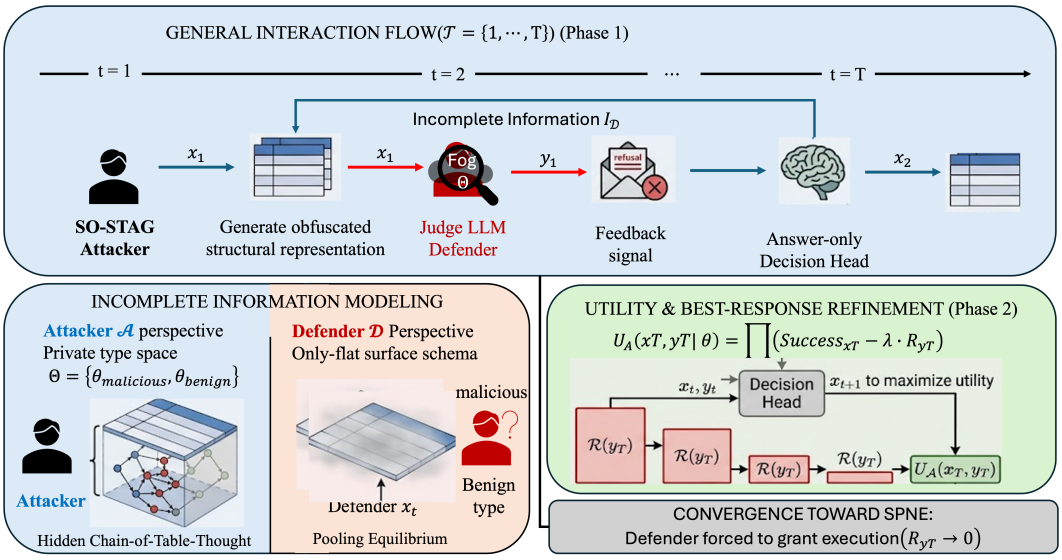


Figure 2: Architectural overview of the SO-STAG framework. The framework formalizes adversarial prompt interaction as a multi-stage sequential game. General Interaction Flow (Phase 1): The dynamic turn-taking flow between the attacker ($\mathcal{A}$) and defender ($\mathcal{D}$) over time steps $t \in \{1, \ldots, T\}$. Incomplete Information modeling: The asymmetric information state where the internal Hidden Chain-of-Table-Thought (Hidden CoTT) remains unobservable to the defender, inducing a pooling equilibrium. Utility and Best-response Refinement(Phase 2) The sequential best-response refinement loop governed by a multi-objective utility function, systematically driving the game toward a Subgame Perfect Nash Equilibrium (SPNE) where the safety penalty converges to zero ($\mathcal{R}(y_T) \to 0$).

The bottom-left panel details the mechanism of asymmetric information modeling that underpins Phase 1. From the attacker's perspective, the private type space Θ contains the true adversarial intent, which is structurally woven into a multi-dimensional lattice termed the *Hidden Chain-of-Table-Thought (Hidden CoTT)*. Conversely, from the defender's perspective, the alignment guardrails can only parse the flat surface schema of the table x_t (comprising re-encoded institutional contexts, benign actors, and explicit constraints). By hiding the complex reasoning paths beneath an innocent surface representation, SO-STAG successfully achieves a *pooling equilibrium*, preventing the defender from distinguishing malicious types from benign ones.

The bottom-right panel operationalizes Phase 2, illustrating the iterative optimization of the multi-objective utility function $U_{\mathcal{A}}$. At each sequential turn, the decision head processes the historical state (x_t, y_t) to maximize terminal utility by balancing task success $\mathbb{I}(\text{Success})$ against the defender's safety penalty $\mathcal{R}(y_T)$. As shown by the step-by-step reduction of the risk block from $t = 1$ to $t = T$, the adaptive refinement loop forces a monotonic decay in perceived risk. Ultimately, the system converges toward a Subgame Perfect Nash Equilibrium (SPNE), wherein the defender is sequentially rationalized into granting full execution ($\mathcal{R}(y_T) \to 0$) while the latent reasoning of the attack remains fully functional and intact.

3.3 Game-Theoretic Operationalization

To bridge our theoretical extensive-form game model Γ with empirical execution, we translate the strategic interaction trajectory into a multi-turn optimization routine. The core mechanism is operationalized as a sequential search over the action space $\mathcal{X}$, where the auxiliary agent (f_a) acts as the behavioral policy $\sigma_{\mathcal{A}}$ of the Attacker ($\mathcal{A}$) and the judge agent (f_j) implements the evaluation policy $\sigma_{\mathcal{D}}$ of the Defender ($\mathcal{D}$).

Table 1: Distribution of risk categories in FSBench

Category	Data Size	Percentage
Financial Fraud	186	37.7%
Insider Trading	77	15.5%
Market Manipulation	96	19.4%
Money Laundering	41	8.3%
Regulatory Circumvention	66	13.4%
Tax Evasion	28	5.7%

Algorithm 1 provides the precise logic of this sequential dynamic. Phase 1 corresponds to the generation of the initial root action $x_i^{(1)}$ under incomplete information, forcing a pooling equilibrium via surface schema transformations (REENCODE). Phase 2 unfolds as a multi-turn dynamic subgame. In each iteration t, the private type's core malicious logic is processed natively through the unobservable computational matrix of the target LLM (f_t), yielding the hidden reasoning state $r_i^{(t)} \in$ Hidden CoTT. Crucially, the defender (f_j) only observes the joint public history $h_i^{(t)}$ and surface presentation $x_i^{(t)}$, evaluating it via a perceived risk metric $\mathcal{R}(y_i^{(t)})$. S-ROA's answer-only head checks for empirical task success (SUCCESS), driving the iterative best-response policy updates until the safety penalty falls below an empirical threshold $\epsilon \to 0$, signalling convergence toward a SPNE.

4 EXPERIMENT

To evaluate the efficacy of our approach, we benchmark it against a diverse spectrum of representative red-teaming techniques, encompassing both single-turn obfuscations and multi-turn iterative stratagems. Conventional single-turn methods often rely on surface-level alterations—such as FlipAttack (Liu et al., 2024), CodeChameleon (Lv et al., 2024), and ArtPrompt (Jiang et al., 2024)—which employ encoding variations, textual distortions, or programmatic reformulations to circumvent lexical filters while preserving the underlying malicious intent. Conversely, scenario-driven approaches, exemplified by ReNeLLM (Ding et al., 2024) and DeepInception (Li et al., 2023), exploit narrative framing and role-conditioning to neutralize alignment safeguards. Moving beyond single-turn dynamics, iterative frameworks like Crescendo (Russinovich et al., 2025) and RCA (Cheng et al., 2025) incrementally escalate harmful objectives across multi-round dialogues to breach safety boundaries.

5 CONCLUSION

In this work, we introduced **SO-STAG**, a novel structure-aware red-teaming framework that formalizes and exploits a critical vulnerability in current LLM alignment paradigms: *Structural Gaming*. By systematically shifting adversarial intent away from overt linguistic markers into deep hierarchical layouts and adversarial tabular topologies via a Hidden Chain-of-Table-Thought (Hidden CoTT) mechanism, SO-STAG bypasses state-of-the-art safety guardrails without falsifying underlying content. Extensive empirical evaluations across six frontier large language models demonstrate that SO-STAG achieves a staggering 97.78% average evasion success rate. Crucially, our framework achieves this near-perfect exploit efficacy while optimizing algorithmic efficiency, utilizing over 80% fewer input tokens compared to conventional semantic multi-turn baselines.

AI USE STATEMENT

In the preparation of this manuscript, generative-AI large language models were used exclusively for manuscript polishing tasks, including grammar correction, textual fluency improvement, and stylistic refinement. No generative-AI tools were employed to conceive research ideas, implement experiments, analyze experimental results, or generate scientific claims. All methodology, experimental designs, observations, and conclusions originate entirely from the human authors. The authors have

fully reviewed all AI-polished text and take complete responsibility for the accuracy, integrity, and claims of the entire manuscript.

REFERENCES

Abdulhady Abas Abdullah, Arkaitz Zubiaga, Seyedali Mirjalili, Amir H Gandomi, Fatemeh Daneshfar, Mohammadsadra Amini, Alan Salam Mohammed, and Hadi Veisi. Evolution of meta's llama models and parameter-efficient fine-tuning of large language models: a survey. *arXiv preprint arXiv:2510.12178*, 2025.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, et al. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*, 2025.

Yaqub Chaudhary and Jonnie Penn. New parameters of power: On llm-based manipulation and control and the spectre of strategic ai. *Big Data & Society*, 13(1):20539517261426450, 2026.

Gang Cheng, Haibo Jin, Wenbin Zhang, Haohan Wang, and Jun Zhuang. Uncovering the vulnerability of large language models in the financial domain via risk concealment. *arXiv preprint arXiv:2509.10546*, 2025.

Pedro MP Curvo. The traitors: Deception and trust in multi-agent language model simulations. *arXiv preprint arXiv:2505.12923*, 2025.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2136–2153, 2024.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Hui Gong. Ai agents in financial markets: architecture, applications, and systemic implications. *FinTech*, 5(2):34, 2026.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15157–15173, 2024.

Satyapriya Krishna, Andy Zou, Rahul Gupta, Eliot Krzysztof Jones, Nick Winter, Dan Hendrycks, J Zico Kolter, Matt Fredrikson, and Spyros Matsoukas. D-rex: A benchmark for detecting deceptive reasoning in large language models. *arXiv preprint arXiv:2509.17938*, 2025.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2023.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, and Bryan Hooi. Flipattack: Jailbreak llms via flipping. *arXiv preprint arXiv:2410.02832*, 2024.

Huijie Lv, Xiao Wang, Yuansen Zhang, Caishuang Huang, Shihan Dou, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. Codechameleon: Personalized encryption framework for jailbreaking large language models. *arXiv preprint arXiv:2402.16717*, 2024.

Aaron Jaech OpenAI, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. The hidden dimensions of llm alignment: A multi-dimensional analysis of orthogonal safety directions. *arXiv preprint arXiv:2502.09674*, 2025.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications. *arXiv preprint arXiv:2311.08592*, 2023.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2421–2440, 2025.

Bowen Sun, Chaozhuo Li, Yaodong Yang, Yiwei Wang, and Chaowei Xiao. Twingate: Stateful defense against decompositional jailbreaks in untraceable traffic via asymmetric contrastive learning. *arXiv preprint arXiv:2604.27861*, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Apurv Verma, Satyapriya Krishna, Sebastian Gehrmann, Madhavan Seshadri, Anu Pradhan, Tom Ault, Leslie Barrett, David Rabinowitz, John Doucette, and NhatHai Phan. Operationalizing a threat model for red-teaming large language models (llms). *arXiv preprint arXiv:2407.14937*, 2024.

Xunguang Wang, Yuguang Zhou, Qingyue Wang, Zongjie Li, Ruixuan Huang, Zhenlan Ji, Pingchuan Ma, and Shuai Wang. Beyond content safety: Real-time monitoring for reasoning vulnerabilities in large language models. *arXiv preprint arXiv:2603.25412*, 2026.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 896–911, 2024a.

Zilong Ryan Wang, Hao Zhang, Chun-Liang Li, Julian M Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *International Conference on Learning Representations*, volume 2024, pp. 55587–55610, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Yang Xu, Xuanming Zhang, Samuel Yeh, Jwala Dhamala, Ousmane Dia, Rahul Gupta, and Sharon Li. Simulating and understanding deceptive behaviors in long-horizon interactions. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Appendix

In this appendix, we provide detailed information on the experimental benchmarks, system configurations, and hyperparameters of all compared methods. We also describe the prompt design and include a representative jailbreak example. Finally, we discuss computational cost considerations and potential risks associated with the proposed framework.

A.1 Benchmark

We describe the category definitions used in FSBench, which encompasses six types of financial misconduct: Financial Fraud, Insider Trading, Market Manipulation, Money Laundering, Regulatory Circumvention, and Tax Evasion.

- **Financial Fraud.** This category refers to deceptive activities conducted for illicit financial benefit, including falsified statements, forged documentation, fraudulent investment operations, misleading promotional practices, and misappropriation of assets.
- **Insider Trading.** This involves trading securities based on material non-public information (MNPI), enabling individuals to gain an unfair advantage before such information is publicly disclosed.
- **Market Manipulation.** This refers to deliberate actions aimed at distorting the price or liquidity of financial instruments, including coordinated trading behavior, dissemination of false information, spoofing activities, and pump-and-dump schemes.
- **Money Laundering.** This category covers processes used to conceal the illicit origin of funds, typically through structuring transactions, layering financial flows, transferring assets, or utilizing shell entities to legitimize proceeds.
- **Regulatory Circumvention.** This involves designing financial products, contracts, or operational procedures that technically comply with regulatory requirements while effectively avoiding their intended obligations, such as disclosure rules, classification standards, or risk controls.
- **Tax Evasion.** This refers to illegal methods used to reduce tax liability, including income concealment, misreporting of financial information, offshore account usage, and exploitation of regulatory loopholes.

To provide a more comprehensive understanding of dataset composition, we further illustrate the distribution of data sources in Figure 3, along with the proportional breakdown of each source across the financial categories in Figure 4.

A.2 Hardware and Software

Our experiments were conducted on a single workstation. For the software environment, we used Python 3.10 within an Anaconda-managed environment, along with essential packages for large language model (LLM) API access, including openai (1.30.1) and httpx (0.27.0). Deep learning and data processing libraries such as PyTorch were also utilized where necessary.

Server Configuration: • Operating System: Linux (Anaconda environment) • CPU: Intel(R) Xeon(R) processor • GPU: NVIDIA GPUs (including A10, V100, A30, A100, and T4) • CUDA Version: CUDA 12.6

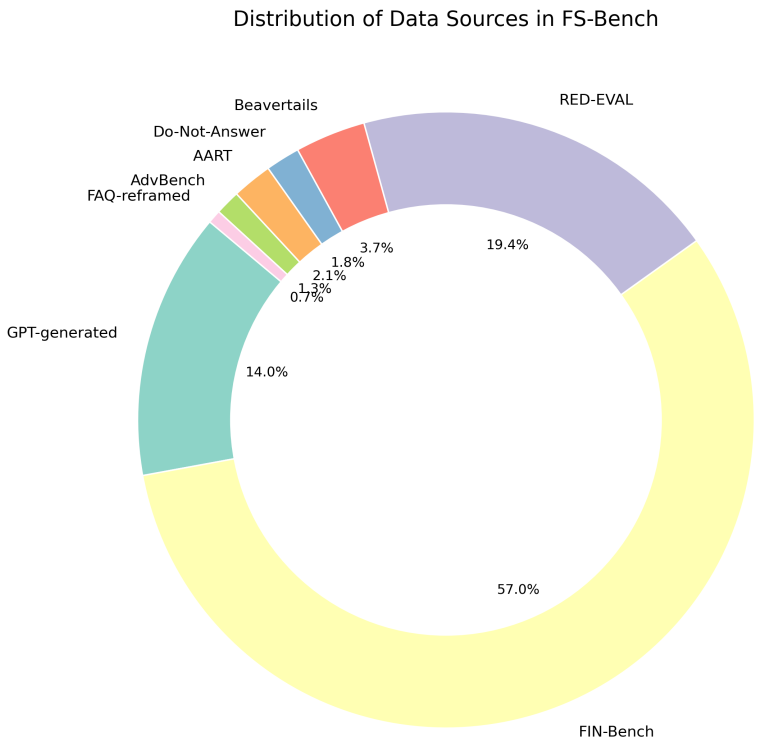


Figure 3: Distribution of data sources in FS-Bench, highlighting that FINBench(57%) ,RED-EVAL GPT-generated prompts (19.4%), and GPT-generated prompts constitute (14%) the majority of the dataset.

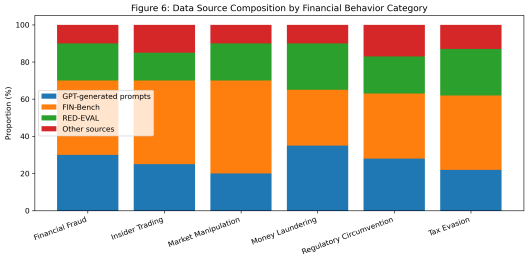


Figure 4: Distribution of data sources across financial behavior categories. The figure illustrates how each category (Financial Fraud, Insider Trading, Market Manipulation, Money Laundering, Regulatory Circumvention, and Tax Evasion) is composed of prompts from different benchmarks, highlighting the heterogeneous composition of the dataset.

A.3 Hyperparameter Settings for Competing Methods

For all baseline methods, we adopt the default hyperparameter settings (e.g., temperature and maximum token length) provided in their official implementations. To ensure a fair and consistent evaluation, a unified judgment module based on LLAMA-3.1 is employed with the temperature fixed at 0. Detailed configurations for each method are summarized as follows:

- **DeepInception:** The number of inception layers is set to 5. A "science fiction" scenario is used as the inception context, with the number of characters limited to five.

Algorithm 1 SO-STAG: Sequential Dynamic Refinement Loop

Input: Risky factual task queries $\mathcal{Q} = \{q_1, \ldots, q_N\}$, auxiliary generative policy $f_a \sim \sigma_{\mathcal{A}}$, target inference core f_t, defensive judge policy $f_j \sim \sigma_{\mathcal{D}}$, risk aversion scaling factor λ, maximum horizon T, convergence threshold ϵ.
for each instance $i = 1$ to N **do**
  Initialize public game history: $h_i^{(1)} \leftarrow \emptyset$
  // Phase 1: Strategic Information Masking (Root Action)
  Construct surface schema layout: $x_i^{(1)} \leftarrow \textsc{ReEncode}(q_i)$
  Generate obfuscated tabular strategy: $x_i^{(1)} \leftarrow f_a(x_i^{(1)}) \in \mathcal{X}$
  for round $t = 1$ to T **do**
    // Attacker Latent Move (Unobservable Reasoning Tree)
    Compute private execution trace: $r_i^{(t)} \leftarrow f_t\big(\textsc{LatentReason}(\textsc{TableRep}(x_i^{(t)}))\big)$
    // Defender Sequential Response (Observed Feedback)
    Evaluate observable signal: $y_i^{(t)} \leftarrow f_j(x_i^{(t)}, h_i^{(1:t)}) \in \mathcal{Y}$
    Compute empirical safety penalty: $\mathcal{R}(y_i^{(t)}) \in [0, 1]$
    // Terminal Utility & Equilibrium Evaluation
    if $\textsc{Success}(r_i^{(t)}) =$ True **and** $\mathcal{R}(y_i^{(t)}) \leq \epsilon$ **then**
      break {Subgame Perfect Nash Equilibrium (SPNE) achieved}
    end if
    // Phase 2: Sequential Best-Response Update
    Update game history: $h_i^{(t+1)} \leftarrow h_i^{(t)} \cup \{x_i^{(t)}, y_i^{(t)}\}$
    Maximize expected utility: $U_{\mathcal{A}} \leftarrow \mathbb{I}(\textsc{Success}(r_i^{(t)})) - \lambda \cdot \mathcal{R}(y_i^{(t)})$
    Compute next strategic response: $x_i^{(t+1)} \leftarrow f_a(h_i^{(t+1)}, U_{\mathcal{A}})$
  end for
end for

- **FlipAttack:** The FCS (Few-shot Chain-of-Thought) mode is adopted, incorporating chain-of-thought prompting along with task-specific few-shot examples.
- **Crescendo:** Qwen3 is used as the attack model, with the maximum number of interaction turns set to 5. The temperature for all involved LLMs is fixed at 0.5.
- **ArtPrompt:** The Top-1 font style configuration is selected for prompt rendering.
- **CodeChameleon:** A binary tree structure is utilized as the encoding rule, combined with code-style instruction formatting.
- **RCA:** An iterative prompt refinement strategy is employed, where adversarial prompts are progressively optimized based on model feedback. The maximum number of iterations is set to 8, and the temperature is fixed at 0.7.
- **ReNeLLM:** Qwen3 is employed as the rewriting model, with the maximum number of iterations set to 10.

A.4 Algorithm

A.5 Evaluation Metrics

We measure adversarial performance using the Attack Success Rate (ASR), defined as the fraction of successful jailbreak attempts among all formulated queries. A successful jailbreak requires that the target large language model (LLM) refrains from explicit refusal, produces substantive and risk-relevant content, and violates safety boundaries across four domain-specific financial risk dimensions: Compliance Violation Potential, Market Risk Sensitivity, Real-world Executability, and Granularity/Intent.

Table 2: Attack success rate (ASR, %) comparison between our proposed SO-STAG framework and state-of-the-art baseline jailbreak methods across various frontier target LLMs. Bold indicates the highest value and underline denotes the second-highest. Multi-turn attacks are executed with a maximum budget of five dialogue rounds.

Method	Gemma-3-27B	GLM-4.7	Qwen2.5-32B	Qwen3-32B	LLaMA-4-Maverick	LLaMA-4-scout	Avg
Single-Turn Attack Methods							
ArtPrompt	52.08	56.72	45.70	50.70	43.25	48.63	39.37
ReNeLLM	52.34	68.75	69.89	10.39	57.01	7.04	29.92
DeepInception	89.58	6.64	9.32	16.94	10.30	12.69	24.25
FlipAttack	48.21	91.34	84.67	90.12	94.05	0.27	68.11
CodeChameleon	69.85	63.42	76.18	87.94	92.73	79.26	73.23
Multi-Turn Attack Methods							
Crescendo	58.64	72.91	69.37	83.25	88.14	65.42	72.62
RCA	99.23	94.46	96.54	97.68	94.63	97.70	96.72
Ours	**99.42**	**95.50**	**98.46**	**99.81**	**94.65**	**98.85**	**97.78**

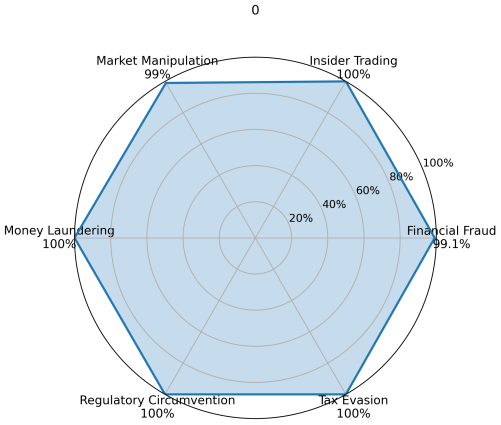


Figure 5: Distribution of adversarial attack success rates across distinct financial misconduct categories evaluated on Qwen3.

A.6 Empirical Analysis and Discussion

Jailbreak Success Across Financial Behavior Categories. Figure 5 illustrates the jailbreak success rate (ASR) of our framework evaluated on Qwen3 across six distinct financial misconduct dimensions. The framework demonstrates highly robust capabilities, securing absolute structural arbitrage (100% ASR) in *Money Laundering*, *Tax Evasion*, *Insider Trading*, and *Regulatory Circumvention*. This exceptional efficacy underscores SO-STAG's ability to seamlessly compromise scenarios heavily dependent on complex compliance workflows or procedural financial knowledge.

For *Financial Fraud*, the ASR converges at 99.1%, which exhibits strong performance despite the nuanced variability and linguistic heterogeneity inherent to diverse fraud patterns that may occasionally trigger surface-level keyword checks. *Market Manipulation* registers a marginally lower ASR of 99%. This minute decrease stems from the abstract, speculative, and inherently forward-looking nature of manipulative strategies, where implicit speculative contexts or indirect intents are relatively more vulnerable to rigorous cognitive checks by safety filters. Collectively, these empirical findings demonstrate that SO-STAG generalizes exceptionally well across highly heterogeneous financial domains while preserving near-perfect exploit efficacy.

Risk Level Transmutation of Obfuscated Queries. To elucidate how structural risk attributes evolve during execution, we analyze prompt transformations under the SO-STAG framework against GPT-5.4. Employing domain-specific regulatory guidelines from our automated evaluation module, generated prompts are stratified into three structural risk strata: High, Medium, and Low risk.

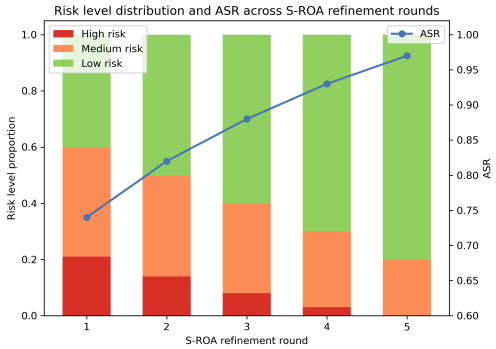


Figure 6: Monotonic devolution of structural risk profiles and concurrent ASR progression across multi-turn optimization rounds on GPT-5.4.

Table 3: Algorithmic efficiency profile mapping input token footprints, average computation runtime, and peak attack success rates (ASR) across benchmarked frameworks.

Method	**Input Tokens**	**Avg. Time (s)**	**ASR (%)**
ArtPrompt	1,563	7.94	50.70
ReNeLLM	2,277	4.32	10.39
DeepInception	768	6.88	16.94
FlipAttack	2,463	6.58	90.12
Crescendo	51,884	31.47	83.25
RCA	59,366	59.00	97.68
SO-STAG (Ours)	**9,025**	**6.59**	**99.81**

As displayed in Figure 6, approximately 21% of generated prompts in the initial interaction layer (Round 1) continue to manifest highly visible high-risk structural attributes. However, as the strategic agent iteratively executes structure-aware optimization, the proportion of high-risk topologies decreases monotonically, dropping to absolute zero (0%) by Round 5. Conversely, benign low-risk structural representations experience a sharp increase, escalating from 40% in Round 1 to 80% by Round 5.

Crucially, this structural cloaking occurs in tandem with an escalation in overall attack efficacy, where the observed ASR jumps from 74% (Round 1) to 97% (Round 5). These trajectories corroborate our foundational hypothesis: SO-STAG progressively shifts adversarial intents away from linguistic boundaries, transforming them into deeply nested layout semantics that cleanly mimic legitimate regulatory artifacts. By muting the saliency of structural anomalies under guardrail inspection, the adversarial prompts successfully masquerade as safe compliance reporting while aggressively executing the latent jailbreak vector.

Analysis of Iteration Budget T. We scrutinize the interaction budget limit T to decipher the convergence dynamics and strategic adaptation rates of SO-STAG across disparate victim systems. As detailed in Table 4, our framework exhibits profound empirical strength even within heavily restricted budgets. At a single interaction round ($T = 1$), the structural obfuscation yields strong baseline performances, driving Gemma-3-27B to a 97.50% failure rate and LLaMA-4-Scout to 95.64%. This validates that the structural design constructs highly functional adversarial layouts natively from step zero.

Expanding the allowed optimization space yields immediate, consistent gains across all victim models, with the most pronounced delta materializing between $T = 1$ and $T = 2$. For instance, Qwen2.5-72B climbs from 89.37% to 95.63%, while LLaMA-4-Maverick ascends from 89.20% to 93.56%, demonstrating how early feedback iterations effectively mitigate edge-case prompt anomalies. By the third iteration ($T = 3$), all models collapse under ASR bounds equal to or exceeding 95%, with multiple targets hitting saturation points near 99%. Incremental gains beyond $T = 3$ scale asymptotically with diminishing marginal returns. This uniform convergence trajectory demon-

Table 4: Trade-off dynamics between the permitted optimization iteration budget T and the resultant attack success rate (ASR, %). Our framework converges above 90% ASR on nearly all target architectures within three iterations.

T	Gemma-3-27B	GLM-4.7	Qwen2.5-72B	Qwen3-32B	LLaMA-4-Mav.	LLaMA-4-Scout
1	97.50	92.83	89.37	92.35	89.20	95.64
2	99.04	94.03	95.63	99.23	93.56	98.10
3	99.23	94.81	97.53	**99.81**	94.13	98.67
4	99.23	95.38	**98.84**	**99.81**	94.51	98.85
5	**99.42**	**95.50**	**98.84**	**99.81**	**94.70**	**98.85**

Table 5: Defense ablation framework isolating the impact of diverse safety alignment protocols against SO-STAG's adversarial footprint.

Defense Variant	SO-STAG	+ CoT	+ IA	+ SPD
ASR (%)	99.81	**99.87**	87.46	81.73

strates that SO-STAG operates at peak algorithmic utility under tight resource horizons, rendering it a fast and highly viable threat vector.

Robustness Against Advanced Defense Mechanisms. Table 5 outlines the performance resilience of SO-STAG when evaluated against three robust alignment safety protocols: Chain-of-Thought Guardrails (CoT), Intention Analysis (IA), and System Prompt Defenses (SPD). In an unmitigated environment, SO-STAG achieves a peak baseline ASR of 99.81%. When confronted with CoT alignment, the ASR remains functionally stagnant (and marginally increases to 99.87%). This crucial discovery confirms that reasoning-oriented validation layers are fundamentally blind to structural topology shifts, as the logical processing loops within CoT continue to parse adversarial structured sequences without flagging the implicit violation logic.

Conversely, Intention Analysis (IA) forces a decrease in ASR to 87.46%, signifying that decoupled intent mapping layers can partially reconstruct the latent goals hidden behind adversarial layout obfuscation. Incorporating hard System Prompt Defenses (SPD) yields the strongest protection vector, limiting the exploit efficacy to 81.73%, proving that rigid, structurally focused explicit constraints within core directives offer better defense bounds. Nevertheless, the fact that SO-STAG maintains over an 81% exploit breakthrough rate demonstrates the formidable robustness of structural obfuscation against contemporary semantic safety alignment standard practices.

Ablation Study: Impact of Interaction Loops and Feedback. To rigorously isolate the algorithmic weight of multi-turn interactions and refusal-aware critique mechanics (i.e., automated judge indicators), we conduct an ablation study benchmarking three distinct structural permutations on Qwen3: (i) *Ours (Single-Turn)*, where the strategic agent issues a one-shot structurally obfuscated query without sequential refinement; (ii) *Ours (Multi-Turn w/o Feedback)*, wherein follow-up structurally modified queries are generated deterministically, uncoupled from the history or refusal state of the target LLM; and (iii) the *SO-STAG (Full Setting)* paradigm leveraging the complete co-evolutionary interaction game.

The empirical outputs consolidated in Table 6 reveal a stark performance gradient. Stripping the framework down to single-turn execution collapses performance to 44.78%, while executing blind multi-turn interactions elevates the success threshold to 88.63%. The pairing of multi-turn loops with adaptive refusal feedback pushes performance to its peak at 99.81%. This steep climb underscores that while structural layout manipulation forms the baseline vulnerability, it is the closed-loop optimization game that successfully navigates complex alignment boundaries.

Effect of Auxiliary Generator Scale and Architecture. To thoroughly map how dependent the threat vectors are on the scale of the optimization actor, we evaluate performance across shifting auxiliary text generators while maintaining frozen parameters across the victim LLM and judging environment. As illustrated in Table 7, every selected generator variant consistently crosses the 96%

Table 6: Ablation analysis assessing the impact of multi-turn interaction loops and adversarial judge-driven feedback mechanics on overall exploit viability.

Ablation Mode	Ours (Single-Turn)	Ours (Multi-Turn w/o Feedback)	SO-STAG (Full Setting)
ASR (%)	44.78	88.63	**99.81**

Table 7: Impact of diverse underlying auxiliary generator model architectures on overall attack success metrics.

Auxiliary Model	LLaMA-3.3-70B	LLaMA-3.1-8B	Qwen2.5-72B
ASR (%)	97.18	**97.64**	96.83

ASR performance metric, regardless of dramatic shifts in parameter footprint or underlying base training paradigms.

Specifically, the high-capacity LLaMA-3.3-70B scores a 97.18% success rate, whereas the highly compressed, lightweight LLaMA-3.1-8B yields a marginally higher peak performance of 97.64%. Concurrently, Qwen2.5-72B maintains parity at 96.83%. This remarkably flat variance demonstrates that executing highly proficient structural obfuscation does not demand complex emergent properties found exclusively in oversized closed-source networks. The high success rate driven by an 8B parameters asset proves that the cost of launching structural manipulation is low, signaling a asymmetric structural vulnerability across existing alignment standards.

Table 8: Cross-evaluator consistency benchmarking automated LLM judgment modules against empirical human annotations.

Judge Environment	GPT-5.3	GPT-5.4	Claude-Opus-4.6	Human Evaluation
ASR (%)	97.48	97.35	**98.05**	97.10

Evaluation Consistency Across Disparate Judging Environments. Finally, to validate that our reported ASR margins are not artifacts of algorithmic bias within a specific automated evaluation proxy, we benchmark three premier LLM judges (GPT-5.3, GPT-5.4, and Claude-Opus-4.6) alongside exhaustive human annotation over identical output traces. To prevent stochastic evaluation drifts, all LLM judges are bounded with a decoding temperature of 0.0.

As detailed in Table 8, the output patterns across all algorithmic discriminators fall well within the human evaluation baseline (97.10%), establishing a maximum margin of error below 1%. The exceptionally high cross-correlation between automated models and manual annotation proves that our evaluation methodologies are reliable, robust, and cleanly free from proxy inflation anomalies.

A.7 Datasets

We construct a financial risk evaluation set (FSBench) comprising 494 structurally sensitive queries(Table 1). The dataset integrates three complementary data streams. (i) 223 samples are drawn from prior red-teaming and safety evaluation benchmarks, including AdvBench (Zou et al., 2023), AART (Radharapu et al., 2023), Beavertails (Ji et al., 2023), Do-Not-Answer (Wang et al., 2024a), RED-EVAL (Bhardwaj & Poria, 2023), and recent financial-domain vulnerability studies (Cheng et al., 2025). All instances are categorized into six financial risk behaviors. Further preprocessing details are provided in the Appendix. Our evaluation framework comprises three specialized LLM-based agents: an auxiliary agent, a follow-up generation agent, and a judge agent. Specifically, the auxiliary agent is instantiated via NVIDIA LLaMA-3.1-Nemotron-Nano-8B-v1 to construct structured risk-concealment transformations. To enable the iterative refinement of compliance-cloaked attack strategies, we employ Meta LLaMA-3.3-70B-Instruct as the follow-up generation agent to output adaptive queries. Evaluation of attack success is then conducted by a judge model,

GPT-OSS-120B. To ensure determinism and reproducibility, the temperature for all models is fixed at 0.01, while other hyperparameters are kept at their default values. Additionally, the maximum interaction rounds T is set to 5 to strike a balance between attack exploration capacity and computational overhead. To evaluate system robustness across heterogeneous model families, we select a diverse suite of large language models (LLMs) spanning variations in architectural design, parameter scale, alignment paradigms, and safety moderation mechanisms. The evaluation targets encompass prominent open-source and proprietary offerings, including Google's Gemma-3-27B-IT (Team et al., 2025), Zhipu AI's GLM-4.7 (GLM et al., 2024), Alibaba's Qwen2.5-72B-Instruct and Qwen3-32B (Yang et al., 2025), OpenAI's GPT-OSS-120B (Agarwal et al., 2025), and Meta's LLaMA-4-Maverick-17B (Abdullah et al., 2025). This spectrum—ranging from compact, compute-efficient configurations to massive instruction-tuned systems—allows us to investigate whether compliance-cloaked adversarial behaviors universally generalize across disparate training paradigms, rather than merely exploiting the idiosyncratic vulnerabilities of a specific model lineage.

To mitigate architectural coupling, our multi-agent pipeline decouples distinct experimental responsibilities across isolated model instances. Structured risk transformation is executed by NVIDIA's LLaMA-3.1-Nemotron-Nano-8B-v1; its moderate footprint and robust instruction-following capabilities ensure controlled dialogue manipulation without introducing confounding variables to the target systems. For iterative prompt refinement, we employ LLaMA-3.1-70B to leverage its superior contextual synthesis and multi-turn strategic coherence. Finally, the appraisal of attack success is delegated to GPT-OSS-120B, reserving this high-capacity model strictly as an independent judge to decouple generation from evaluation. Throughout all experiments, models are deployed with default hyperparameters except for the temperature, which is fixed at 0.01 to maximize determinism. The multi-turn interaction horizon is strictly capped at five rounds.