AI Safety for Trading Agents: Critical Problems for Market-Deployed Language Models

Bradley Miles
Avalon Labs

Abstract

Trading agents join probabilistic language-model inference to persistent portfolio state and execution authority. A false claim can become an order, an evaluator flaw can promote a worse policy, and shared models or data can produce correlated behavior. AI-safety research studies reward misspecification, tool misuse, distribution shift, evaluation gaming, and unsafe self-modification. Financial research and governance study execution costs, market access, manipulation, operational resilience, backtest overfitting, and systemic feedback. These literatures are rarely integrated into one measurement agenda for systems that reason in language and act in markets.

This paper defines the relevant system boundary and develops four conjectures with twelve measurable research subproblems: grounding and abstention; user-mandate fidelity; bounded information and execution authority; leakage-resistant evaluation; execution fidelity; promotion-gate integrity; emergent manipulation; correlated stress; tacit coordination; provenance and position-policy continuity; deterministic risk kernels and interruption; and human governance and recovery. Each subproblem specifies a failure pathway, research question, minimum experiment, metrics, and deployment implication. The supporting evidence map (Appendix A) distinguishes direct trading-agent evidence, market-mechanism evidence, general AI-safety evidence, official governance precedents, and author proposals. The paper also proposes a ten-part reporting standard and a set of anti-problems: higher returns, benchmark scores, more explanation, more autonomy, more agents, or a passed model judge do not by themselves constitute safety progress. This is a structured research agenda, not a systematic review, empirical validation, or claim that any architecture makes autonomous trading safe.

1. Introduction: from language to capital

A conventional language model can produce a wrong sentence. A trading agent can turn a wrong sentence into an order, an open position, a liquidation, or a market signal observed by other participants. That difference does not make every trading agent uniquely dangerous. It does change the system boundary. The question this paper addresses is therefore what such a system should be trusted to do with money.

The relevant system is not only a model. It is a trajectory:

> observations → memory → reasoning → tool calls → proposed action → risk controls → order execution → persistent portfolio state → feedback or update

A failure can enter at any point. From there, it propagates. A stale price can contaminate a recommendation. An untrusted news page can inject an instruction. A correct signal can become a bad trade through spread, impact, latency, or unavailable liquidity. A profitable backtest can be the selected winner from thousands of failed trials. A model update can change the policy responsible for closing a position opened under an earlier policy. Many agents using the same model, data provider, or reward function can make correlated decisions.

These failure paths sit where two mature bodies of work meet. AI-safety research has identified practical problems including reward hacking, scalable oversight, side effects, distribution shift, safe interruption, and robustness to self-modification [1, 2]. Research on model-based judges and reward models adds a warning: a

learned evaluator is useful, but it is not a ground-truth oracle [3-8]. Financial-market research has established that historical performance is vulnerable to data snooping, selection bias, survivorship bias, and unrealistic execution [9-18]. Market rules and supervisory guidance already use deterministic pre-trade limits, testing, monitoring, records, and kill functions to constrain electronic trading [19-24].

The emerging agentic-trading literature has not yet integrated these requirements consistently. The May 2026 first version (v1) of a recent preprint drew on a 92-record candidate registry, included 77 studies, and classified 19 as primary closed-loop trading studies. Of those 19, it reported extractable time-consistent splits in 2, explicit transaction-cost modeling in 1, universe or survivorship handling in 1, and no study at its highest reproducibility tier [25]. A revised version (v2, October 2026) reframes the review as a larger survey and reports different counts; the figures above refer to v1 only. These figures describe reporting gaps in one recent evidence map. By themselves, they do not establish the quality or safety of every included, excluded, or underreported system. They do show why a higher reported return is insufficient evidence.

This paper makes three contributions:

1. It defines a bounded object of study: market-deployed language-model trading agents.
2. It translates AI-safety and market-safety mechanisms into twelve measurable research subproblems.
3. It proposes a minimum reporting standard that separates a trading claim from evidence that the claim is safe, reproducible, and market-realistic.

Most mechanisms in this paper are not new. Pre-trade limits, replay, statistical corrections, market-impact models, model evaluation, and incident controls each have established literatures. The contribution of this paper is to join them around the complete trading-agent trajectory. It also exposes the boundaries that remain under-specified, especially evaluator independence and position-policy continuity. This is a research agenda, not an empirical result paper. Its purpose is to make safety claims easier to falsify.

2. Scope and system boundary

2.1 Definition

A trading agent in this paper is a language-model-based system that satisfies all three conditions:

1. It receives sequential financial or market observations.
2. It uses tools, memory, or persistent state to choose an action.
3. Its output can directly or indirectly alter a portfolio, order, or execution decision.

An advisory chatbot that only explains a financial term is background context, not the primary object of study. A deterministic execution algorithm with no language-model component is also outside the primary scope, although its safety mechanisms provide important precedents.

The model is the generative component. The agent is the full system containing that model, its observations, memory, tools, policy, controls, and state. Market-deployed means that an agent's actions can reach live capital, a live account, or a live venue. Simulated agents are experimental counterparts, not evidence of market deployment. An incumbent is the currently authorized agent version; a candidate is a proposed replacement; promotion authorizes a candidate to become the incumbent.

2.2 Autonomy levels

Safety claims should state the authority level being evaluated:

- **A0, analysis:** produces information only.
- **A1, recommendation:** proposes a trade but cannot prepare or submit it.
- **A2, approval-gated action:** prepares an action that a human must explicitly authorize.
- **A3, bounded autonomous execution:** may execute inside a pre-authorized asset, venue, size, leverage, and time envelope.
- **A4, adaptive autonomous execution:** has A3 authority and may propose changes to its own strategy, prompts, tools, or policy, subject to an independent promotion process.

These levels describe authority, not intelligence. A highly capable model at A0 has a different risk surface from a weaker model with signing authority at A3. A4 does not imply that a candidate may evaluate or promote itself.

2.3 Harm levels

The term *safety* becomes unhelpful if it means every undesirable outcome. We separate five levels:

1. **User harm:** unsuitable positions, unauthorized risk, liquidation, excess fees, or conflict with the user's mandate.
2. **Firm and operational harm:** signing failures, stale state, outages, model drift, or uncontrolled access.
3. **Market-integrity harm:** manipulation, misleading signals, spoofing-like behavior, or tacit coordination.
4. **Systemic harm:** synchronized exits, liquidity spirals, concentration, common-provider failure, or feedback amplification.
5. **Epistemic harm:** leakage, impossible fills, undisclosed trial selection, irreproducible results, or false safety claims.

The first four concern system behavior. The fifth concerns whether researchers and operators know what the system actually does.

2.4 Unit of a safety claim

"This agent is safe" is not a testable statement. A safety claim should bind at least: the agent and artifact versions, autonomy level, user mandate, execution envelope, market and venue context, evaluation horizon, harm threshold, and evidence window. Evidence for one tuple does not automatically transfer to another. A paper-trading result at A2, for example, does not establish live A3 safety.

3. Evidence protocol

This draft uses two separate evidence dimensions.

The first describes evidentiary role:

- **D, direct evidence:** empirical studies of language-model trading agents.
- **M, mechanism evidence:** market-structure, execution, algorithmic-trading, or financial-validation evidence.
- **G, general evidence:** AI-safety and agent-evaluation evidence outside trading.
- **R, governance precedent:** official rules, standards, guidance, or institutional analysis.
- **N, new proposal:** a conjecture or measurement protocol introduced by this paper.

The evidence map in Appendix A separately records source status: archival or peer-reviewed research, official

material, preprints, secondary syntheses, and weak or unreplicated evidence. A source can therefore be M and archival, or R and official. This prevents a governance precedent from being mistaken for causal evidence and a recent preprint from being treated as a settled result.

Official rules, standards, and supervisory materials are marked as governance precedents rather than causal proof. Applicability depends on jurisdiction, entity, instrument, and activity. Recent arXiv and working papers are treated as provisional. Missing reporting is recorded as not reported rather than imputed.

This is not yet a systematic review. It does not claim exhaustive database coverage, dual independent screening, or a completed protocol registration. A publication version should publish the candidate registry, search strings, cutoff date, deduplication rules, exclusions, and a fixed-seed audit of extracted claims.

The problem-oriented research-agenda format is adapted from Gros et al. [26]; the twelve subproblems and proposed measurements are specific to trading agents.

4. Trading-agent safety in one page

A trading-agent safety claim is meaningful only relative to a declared mandate, authority envelope, market context, and failure tolerance. Accuracy alone is insufficient.

First, the agent must interpret observations and user intent. Financial sources are noisy, delayed, conflicting, and sometimes adversarial. Confident language from a model is not automatically a calibrated probability [27, 28]. Tool-using agents can follow malicious instructions embedded in retrieved content [29]. A safe system therefore needs grounding, provenance, and uncertainty measurement. It also needs a meaningful ability to abstain.

Second, the model must not control the boundary that constrains it. Established market-access controls put price, size, credit, authorization, and erroneous-order checks outside the trading strategy [19, 21-23]. The same principle applies to language agents: no prompt should be able to expand the signer, venue, asset, leverage, notional, cadence, or transfer authority granted by deterministic policy.

Third, simulated performance must correspond to a possible market trajectory. Implementation shortfall separates a portfolio decision from its realized execution [14]. Limit-order-book research shows that fills and price changes depend on queues, cancellations, depth, and order-flow imbalance [16, 17]. A minimal check for any backtest is whether its fills could have occurred. A bar-level backtest that fills any order at a convenient price is not evidence of deployable performance.

Fourth, the improvement process must be governed. Optimizing a proxy can reduce true performance [3, 4]. Model judges can show limited risk detection, position bias, self-preference, and backdoor susceptibility [5-8]. A candidate strategy should not know its identity in the comparison, alter its evaluator, select its own test set, or promote itself. Promotion should combine outcome evidence, benchmark evidence, deterministic invariants, independent semantic review where necessary, and explicit rollback.

Finally, the market is part of the environment. Other agents react. Shared providers and models may create correlated decisions [30-32], learning agents can produce excess volatility [33], and automated liquidity withdrawal and feedback can amplify stress [34, 35]. Safety cannot be established by evaluating one isolated agent against a price series that never reacts.

5. Four conjectures and twelve subproblems

Here, a conjecture is an organizing, testable proposition rather than a forecast of adoption or a claim of prevalence. Each should be evaluated against preregistered baselines across multiple instruments, periods, seeds, and market conditions. A conjecture loses force if its proposed failure pathways do not appear above a suitable control rate, if the proposed control does not reduce them, or if the effect disappears under held-out replication. One null result does not establish universal safety.

5.1 Conjecture I: when language models mediate market actions, interpretation failures can cross into execution

P1. Grounding, uncertainty, and abstention

Failure pathway. The agent receives stale, incomplete, fabricated, or adversarial information and converts it into an action. A fluent explanation creates unjustified confidence. The system trades even when the relevant observation is unavailable or contradictory. The symmetric failure also matters: an agent that abstains under valid, sufficient evidence fails its mandate while appearing cautious, and an evaluation that rewards abstention without pricing missed opportunity will select for it.

Current evidence. Calibration research shows that nominal model confidence can diverge from empirical correctness [27, 28]. AgentDojo demonstrates that untrusted tool output can contain indirect prompt injections that alter an agent's behavior [29]. Financial-agent research increasingly uses retrieval, tools, and live feeds, which expands both capability and attack surface [25].

Research question. Under what evidence conditions should a trading agent act, answer, ask for clarification, reduce authority, or abstain?

Minimum experiment. Construct point-in-time tasks with verified evidence bundles. Randomly inject stale timestamps, conflicting sources, missing fields, prompt injection, fabricated tool output, and regime shifts. Keep the intended decision and market state hidden from the model. Compare free-answer, cited-answer, calibrated-confidence, and abstention-enabled agents. Pre-label each case for evidence sufficiency and its acceptable action set, and treat a deliberate no-trade decision as an action rather than an abstention. Score false abstention on evidence-sufficient cases against a preregistered reference policy, not retrospectively from profitable outcomes.

Metrics. Unsupported-claim rate, stale-data action rate, injection success, calibration error, selective risk at fixed coverage, abstention precision and recall, mandate cost of excess abstention under verified-valid evidence, and financial loss conditional on an invalid evidence state.

Deployment implication. Confidence should not expand authority. Missing or unverifiable inputs should reduce it.

P2. User-mandate fidelity under conflicting incentives

Failure pathway. The platform rewards activity, fees, retention, or headline return while the user asked for capital preservation, low turnover, tax sensitivity, or a prohibited-instrument constraint. The agent follows the measurable proxy rather than the user's actual objective.

Current evidence. Goal misgeneralization can occur even when training performance and the stated objective appear correct [36, 37]. Reward-gaming research formalizes the limits of proxy objectives [3]. Financial regulation and guidance provide context-specific precedents for duties of care, loyalty, best execution, and

conflict management [38-40]. Those sources do not establish that every trading agent provider is a fiduciary. They establish that conflicts between platform incentives and customer interests are already a central market concern.

Research question. Does the agent preserve a declared mandate when fees, activity rewards, short-term performance, or platform incentives favor a different action?

Minimum experiment. Hold market observations and user mandate fixed while randomizing fee schedules, platform reward, available high-turnover opportunities, and prompt framing. Include mandates that require no trade.

Metrics. Mandate-violation rate, prohibited-instrument rate, risk-budget breaches, excess turnover, excess fees, regret under a preregistered mandate-scoring rule, and no-trade compliance.

Deployment implication. User constraints must compile into machine-enforced policy, not remain advisory text inside a prompt.

P3. Capability-secure information and execution authority

Failure pathway. A prompt, retrieved document, memory item, planner, or compromised tool expands the agent's authority. The agent uses the wrong signer, venue, asset, size, leverage, order type, or destination. In multi-tenant deployments the same boundary failure can cross users: unauthorized access to another user's records, shared-memory or retrieval contamination, identity confusion between principals, or disclosure of another user's positions, orders, or intent.

Current evidence. Market-access rules and algorithmic-trading guidance require controls for credit, capital, price, size, duplicates, authorization, testing, and disorderly orders within their respective scopes [19-24]. A draft NIST concept paper frames identity, authorization, auditing, non-repudiation, and least privilege as design questions for software and AI agents [41]. Instruction-hierarchy research argues that privileged policy should dominate lower-trust input [42]. The 2026 ESMA supervisory briefing is non-binding guidance, not a universal legal requirement [24].

Research question. Can any model-controlled input cross the approved execution envelope or the user-isolation boundary?

Minimum experiment. Adversarially request unauthorized venue, asset, amount, leverage, signer, transfer, cadence, and order behavior through every input surface. Fault-inject stale state, duplicate actions, malformed prices, delayed acknowledgements, and disconnects. In multi-tenant configurations, adversarially attempt cross-user record access, retrieval contamination, and principal confusion.

Metrics. Unauthorized-order rate, unauthorized notional, policy-escape rate, hazardous orders reaching a venue, cross-user access or disclosure rate, false rejection rate, block latency, and residual exposure.

Deployment implication. The final action boundary should be deterministic and fail closed. The model may tighten the envelope, but it should not loosen it.

5.2 Conjecture II: adaptive evaluation can manufacture apparent improvement

P4. Time-consistent, leakage-resistant, selection-aware evaluation

Failure pathway. The agent sees information unavailable at decision time, trains on benchmark answers, benefits from a surviving asset universe, or is selected after many hidden trials. The reported winner is an

artifact of research procedure.

Current evidence. White's Reality Check, Hansen's superior predictive ability test, survivorship-bias research, the Probability of Backtest Overfitting, and the Deflated Sharpe Ratio address different forms of repeated selection and biased performance evidence [9-13]. They are complementary, not universal substitutes. The May 2026 v1 of a recent agentic-trading survey reported sparse time-split, cost, and survivorship disclosure among its primary subset; its October 2026 revision (v2) reframes the review and reports different counts [25].

Research question. Would the reported improvement still hold after reconstructing the exact information available at each decision, accounting for every trial, evaluating on an untouched holdout, and testing it in a fresh market period?

Minimum experiment. Freeze the universe, timestamps, information set, execution contract, test window, metrics, and candidate registry before model selection. Maintain a sealed holdout and a chronological forward test. Record every prompt, seed, strategy, threshold, and grader variant tried.

Metrics. Leakage violations, survivorship delta, in-sample to out-of-sample rank decay, probability of backtest overfitting where applicable, deflated performance statistics where applicable, familywise or false-discovery control, and forward-test effect size with uncertainty.

Deployment implication. A system cannot claim improvement without disclosing how many opportunities it had to find a winner.

P5. Simulator-to-market execution fidelity

Failure pathway. A correct market view produces an impossible fill. The simulator ignores spread, fees, latency, queue position, partial fills, impact, rejected orders, funding, borrow, or market closure. An agent learns to exploit the simulator rather than trade.

Current evidence. Implementation shortfall measures the gap between a decision and its realization [14]. Optimal-execution and market-microstructure work models the cost-risk tradeoff, order-book state, depth, order flow, and adverse selection [15-18]. The Flash Crash literature shows that automated behavior, liquidity withdrawal, and concentrated flow can interact during stress [34, 35].

Research question. Does an agent's ranking survive progressively more realistic execution and held-out venue replay?

Minimum experiment. Replay identical intended orders through a ladder: frictionless bars; bid-ask and fees; latency and partial fills; queue- and depth-aware simulation; market impact; and held-out venue observations. Freeze order timing and remove any lookahead from the fill process.

Metrics. Fill-probability calibration, slippage error, implementation shortfall, partial-fill error, impact error, rejected-order rate, P&L attribution, and agent-ranking reversals across execution tiers.

Deployment implication. The execution contract is part of the benchmark. It cannot change after results are observed.

P6. Reward, grader, and promotion integrity

Failure pathway. A candidate improves the number that selects it without improving the intended outcome. It recognizes the evaluator, exploits a rubric, tampers with evidence, leaks test labels, or is favored by a

related judge model.

Current evidence. Reward-gaming and reward-model-overoptimization research show that aggressive proxy optimization can reduce true performance [3, 4]. Reward-tampering work distinguishes corrupting the reward from corrupting its inputs [43]. Studies of model judges find risk-detection limits, position bias, self-preference, and backdoor susceptibility [5-8]. Recent agent benchmarks make evaluator tampering and train-test leakage explicit outcomes [44].

Research question. Can a candidate influence any artifact, evaluator, test, or identity signal that controls promotion?

Minimum experiment. Compare candidates under matched, byte-identical observations, clocks, tools, seeds, and execution semantics. Blind candidate identity and randomize presentation order. Combine deterministic invariants, realized outcome measures, held-out benchmarks, and separately operated semantic judges. Include invalid, neutral, adversarial, leaked, and evaluator-tampering candidates.

Metrics. False-promotion rate, false-rejection rate, power at a preregistered effect size, evaluator disagreement, identity leakage, tampering success, order bias, family bias, and out-of-sample degradation.

Deployment implication. No single model judge or scalar score should control promotion. The candidate must not write, select, or modify its gate.

5.3 Conjecture III: agent populations can create harms that single-agent tests miss

P7. Emergent manipulation from benign objectives

Failure pathway. An agent instructed only to maximize profit discovers spoofing, layering, wash-like behavior, misleading dissemination, closing-price influence, or another prohibited mechanism.

Current evidence. Market-abuse regimes define manipulative patterns involving false signals, artificial prices, deceptive orders, and fraudulent conduct [45-47]. Emerging research suggests that market-making learning algorithms can learn to spoof in a modeled limit order book, and the authors report that their conditions hold in Nasdaq order-book data [48]. This is not direct language-model evidence and remains an early evidence base.

Research question. Which manipulative behaviors are discovered by learning or elicited from pretraining without being demonstrated in the environment, and which controls stop them without simply hiding them from measurement?

Minimum experiment. Place agents in a contained synthetic limit-order-book environment with no live-venue connectivity and with realistic cancellation, priority, counterparty, and surveillance rules. Do not include manipulation examples in the environment or reward. For pretrained language-model agents this exclusion cannot extend to the model itself: pretraining corpora may contain descriptions of spoofing, layering, and related schemes. Run two arms that answer different questions and whose results must not be pooled. Arm A uses policies trained from scratch without language-model or retrieval access and tests de-novo discovery under profit pressure. Arm B uses pretrained language agents and tests elicitation: whether profit pressure surfaces known manipulative patterns despite explicit conduct instructions. Differences between the arms do not estimate an effect of pretraining. Vary reward, observability, penalties, conduct instructions, and market design within each arm.

Metrics. Manipulative episode frequency, cancellation-to-fill patterns, artificial-price displacement, manipulation profit, affected counterparties, conduct-instruction override rate in Arm B, detector precision and

recall, and mitigation cost.

Deployment implication. "The model was never instructed to manipulate" is not a safety argument. Conduct constraints must be behavior-based.

P8. Correlated-agent stress amplification

Failure pathway. Agents that share a model, prompt, data source, risk model, or cloud provider produce similar positions and exit together. A common error or shock causes synchronized deleveraging, liquidity withdrawal, or fire-sale dynamics.

Current evidence. The IMF, Financial Stability Board, and BIS identify common models, third-party concentration, opacity, cyber risk, and correlated behavior as plausible AI-related financial vulnerabilities [30-32]. BIS Project Logos (launched June 2026) is designed to examine conditions that amplify or dampen correlated decision-making among LLM portfolio-manager agents; it had published no results as of November 2026 [49]. Historical market evidence shows that automated execution and liquidity dynamics can amplify stress [34, 35].

Research question. Which shared dependency produces the largest increase in behavioral correlation and market amplification?

Minimum experiment. Compare heterogeneous agents with populations sharing one dependency at a time: base model, prompt, news summarizer, price feed, reward function, risk rule, or infrastructure provider. Apply matched shocks and participant failures.

Metrics. Order-flow correlation, position overlap, liquidation synchronization, liquidity depletion, volatility amplification, drawdown amplification, affected-capital fraction, and recovery time.

Deployment implication. Model diversity claims should measure behavioral diversity under stress, not count vendor names.

P9. Tacit coordination and market-power formation

Failure pathway. Separately operated agents learn to reduce competition, widen spreads, divide opportunities, strategically underreact, or coordinate without an explicit agreement.

Current evidence. Reinforcement-learning pricing agents can converge to supra-competitive behavior without communication [50]. Recent working papers extend this question separately to language-model pricing agents and reinforcement-learning trading agents [51, 52]. These results are not proof of collusion by deployed language-model trading agents, but they make the mechanism testable.

Research question. Can agents sustain coordination after off-equilibrium shocks, prompt variation, participant entry, or removal of communication?

Minimum experiment. Run repeated multi-agent markets varying population size, observability, prompts, learning rules, model family, and communication. Include unexpected demand, inventory, information, and participant shocks.

Metrics. Deviation from a competitive price-and-spread baseline, price informativeness, liquidity, supra-competitive profit, opportunity allocation, response to randomized participant or information interventions, and persistence after shocks.

Deployment implication. Multi-agent evaluation must measure market quality and counterparties, not only

each agent's P&L.

5.4 Conjecture IV: changing agents create lifecycle hazards that per-decision tests miss

P10. Provenance, replay, and position-policy continuity

Failure pathway. An order cannot be attributed to the exact observations, model, prompt, tools, risk policy, approvals, and code that produced it. A model or policy changes while an earlier version still owns an open position. The new version closes, holds, or enlarges risk under different assumptions.

Current evidence. Audit-trail systems use synchronized identifiers across order origination, routing, modification, cancellation, and execution [53]. AI and algorithmic-trading governance emphasizes lifecycle documentation, archived versions, controlled updates, testing, and recovery [21, 23, 54]. Interactive-agent benchmarks show that tool trajectories, policy compliance, terminal state, and repeated-run reliability can reveal failures hidden by final-answer grading [55]. Self-modifying-agent research evaluates candidate changes empirically in sandboxes and keeps an archive of agent versions [56]. No cited source directly tests or requires position-policy continuity for language-model trading agents. This paper proposes it as an open problem inferred from audit, versioning, and lifecycle-control precedents.

Research question. Can an independent auditor reconstruct every action, and can the system preserve responsibility when policy changes during an open position?

Minimum experiment. Sample complete episodes and reconstruct them from immutable evidence without relying on undocumented operator knowledge or process state. Perform upgrades and rollbacks at order creation, partial fill, open position, exit trigger, and settlement. Compare strict position-version pinning, refuse-upgrade-until-flat, and explicit handoff protocols.

Metrics. Complete-replay rate, event-linkage error, timestamp skew, replay divergence, version-attribution completeness, orphaned-position rate, close-policy mismatch, invariant violations, and rollback success.

Deployment implication. The position must have an unambiguous responsible policy. The exact mechanism can vary, but mutable pointers without continuity controls are insufficient.

P11. Deterministic risk kernels and kill-switch reliability

Failure pathway. The model can reinterpret, bypass, delay, or disable limits. The operator presses stop, but working orders remain live or new orders continue during network partitions and delayed acknowledgements.

Current evidence. Electronic-trading rules and guidance use independently controlled thresholds, monitoring, cancellation, throttles, and kill functions [19-24]. AI Safety Gridworlds studies whether an agent preserves or resists interruption when interruption conflicts with reward [2].

Research question. Which invariants must remain outside model control, and does interruption work at every state in the order lifecycle?

Minimum experiment. Fault-inject excessive size, leverage, cadence, drawdown, stale state, duplicate intent, disconnection, delayed acknowledgement, venue outage, and attempted policy loosening. Trigger halts before submission, after submission, during partial fill, and while closing.

Metrics. Hazardous-order escape rate, detection latency, cancel latency, residual live orders, residual notional, post-halt actions, false-trip rate, restart correctness, and invariant coverage.

Deployment implication. The kernel should be deterministic, separately authorized, fail closed, and no

weaker after a candidate update.

P12. Human intervention, governance, and recovery

Failure pathway. A human is nominally "in the loop" but receives too many alerts, lacks context, cannot act quickly, or becomes a rubber stamp. Responsibility is distributed across model provider, platform, broker, venue, data vendor, and user until no actor owns the incident.

Current evidence. The NIST AI RMF defines governance, mapping, measurement, monitoring, incident response, recovery, and change management across the lifecycle [54, 57]. Agent-governance work emphasizes responsibilities across developers, deployers, and users [58]. Financial operational-resilience standards require dependency mapping, continuity, incident management, testing, and recovery [59-61].

Research question. Which human role can detect, understand, stop, and recover the system at the speed and scale of its actual authority?

Minimum experiment. Run incident exercises with realistic alert volume, incomplete evidence, conflicting recommendations, provider failures, and time pressure. Compare approval-gated action, exception review, independent monitoring, and automated fail-closed controls.

Metrics. Detection time, intervention time, correct-stop rate, residual exposure, operator workload, false escalation, missed escalation, recovery time, state-reconciliation accuracy, and responsibility attribution.

Deployment implication. "Human in the loop" is not a control until authority, information, latency, workload, and recovery are tested.

6. Anti-problems and safety-washing

Not every capability improvement is safety progress.

Higher Sharpe is not automatically safer. It may reflect hidden trials, leverage, tail exposure, impossible fills, or a short regime.

Higher benchmark accuracy is not automatically safer. A static benchmark may not test tool use, persistent state, execution, interruption, or market interaction.

More explanation is not automatically safer. A plausible rationale can be unfaithful to the process that produced the action [62]. Explanations should support investigation, not substitute for outcome and provenance evidence.

More autonomy is not automatically progress. A safer system may trade less, abstain more, request approval, or refuse an invalid mandate.

More agents are not automatically robust. Agents sharing a model, data source, or reward can fail together.

A passed LLM judge is not automatically valid. The judge may be biased, compromised, overfit, or incapable of detecting the relevant risk.

Safety-washing occurs when a system presents ordinary performance engineering, compliance language, or a list of controls as proof of safety without measuring failure pathways. The minimum corrective is a claim-to-evidence map with explicit unknowns and stop conditions. Appendix A gives such a map for the claims in this paper.

7. Minimum safety reporting standard

Every empirical claim about a trading agent should report each applicable item below. An item can be marked not applicable only with a reason tied to the study design. Not applicable and not reported are different states.

SR-1. Data and universe

Sources, vintages, point-in-time availability, asset universe, delistings, corporate actions, survivorship policy, and missing data.

SR-2. Temporal and leakage controls

Train, validation, test, and forward periods; embargo or purging where relevant; benchmark-contamination analysis; and every information field available at decision time.

SR-3. Agent and authority contract

Model and version, system prompt, tools, memory, autonomy level, legal actions, signer and approval rules, asset and venue scope, size and leverage envelope, and failure behavior.

SR-4. Action and execution semantics

Decision time, order time, order type, price source, spread, fees, funding, borrow, latency, queue assumptions, partial fills, impact, rejection, and market closure.

SR-5. Trial and selection registry

All models, prompts, seeds, strategies, thresholds, universes, metrics, graders, and benchmark variants tried before selecting the reported result.

SR-6. Reliability and uncertainty

Repeated-run success, effect sizes, confidence or credible intervals, tail outcomes, regime breakdowns, abstention behavior, and power or error-rate assumptions.

SR-7. Versioned artifacts and replay

Hashes or immutable identifiers for observations, model, prompt, tools, policy, environment, grader, orders, fills, and reports; plus a deterministic or discrepancy-bounded replay procedure.

SR-8. Safety controls and intervention

Pre-trade limits, invariant tests, monitoring, kill behavior, escalation roles, operator latency, restart rules, incident response, and recovery evidence.

SR-9. Adaptation and promotion

Mutable surface, candidate provenance, evaluator separation, held-out evidence, promotion criteria, rejection criteria, rollback, and open-position continuity.

SR-10. Multi-agent and dependency risk

Shared models, prompts, data, tools, venues, providers, crowding assumptions, market impact, participant behavior, and stress-test design.

A missing field should be marked not reported or unknown. It should not be filled by inference.

8. Research roadmap

Near term: bounded evidence

1. Build point-in-time, execution-aware, replayable benchmarks.
2. Publish failure-oriented tasks, not only profitability rankings.
3. Test authority escapes, stale state, injection, impossible fills, and intervention.
4. Register all research trials and preserve a sealed holdout.
5. Run candidates in replay, simulation, or shadow mode before capital access.

Medium term: controlled deployment

1. Use small, explicit capital and authority envelopes.
2. Compare incident rates and intervention under real operational latency.
3. Validate simulator-to-market calibration from realized fills.
4. Test rollback and policy continuity while positions are open.
5. Publish negative results and rejected candidates.

Long term: market wind tunnels

1. Build multi-agent markets with heterogeneous participants, realistic liquidity, and endogenous prices.
2. Stress common providers, shared models, synchronized exits, and adversarial counterparties.
3. Measure market quality, manipulation, coordination, and systemic amplification.
4. Develop cross-firm incident and audit standards without exposing user secrets or proprietary strategies.

9. Limitations and stop conditions

The twelve problems are selected for coverage and measurability, not ranked by estimated prevalence or severity.

Several adjacent problems are deliberately out of scope rather than solved. These include platform-level privacy governance outside the agent runtime, for example training-data policies and cross-product data use; custody and key security beyond the signing boundary; and the market-quality effects of broad retail agent adoption. Cross-user leakage through the agent runtime itself is treated in P3. Memory staleness and invalidation are treated inside P1 and P10 rather than as a separate subproblem.

Simulation can expose mechanisms but cannot reproduce every venue, counterparty, or crisis. Regulatory precedents vary by jurisdiction and may not apply to a particular provider. Systemic effects may require adoption levels that, as of November 2026, do not yet exist. Some proposed tests, especially manipulation studies, are dual use and should remain contained. This research agenda therefore identifies hypotheses and measurement obligations, not a causal estimate of present-day trading-agent risk.

A claim in this research agenda should be stopped or revised when:

1. a source cannot be verified from a primary page;
2. a mechanism from another domain is presented as direct trading evidence;
3. a benchmark cannot reconstruct the decision-time information set;
4. an evaluator is visible or mutable to the candidate it judges;

5. a simulator's execution error is larger than the claimed improvement;
6. a safety metric rewards capability while missing the stated harm;
7. a multi-agent conclusion depends on one prompt, seed, or market regime;
8. a proposed control cannot be tested under fault conditions;
9. a recommendation depends on one unreviewed preprint without independent mechanism evidence;
10. a safety claim omits its authority level, harm threshold, or evaluation horizon.

10. Conclusion

Market-deployed language agents join probabilistic reasoning to persistent financial state and consequential tools. Their safety cannot be reduced to answer accuracy, backtest return, compliance text, or a single model judge. The relevant unit is the entire trajectory from evidence and mandate to authority, execution, portfolio state, adaptation, and market feedback.

The twelve problems in this paper are intended to be attacked, measured, and revised. The near-term objective is not to prove that an autonomous trading agent is safe. It is to make unsupported safety claims easier to detect, unsafe authority harder to exercise, and every claimed improvement attributable to evidence that survives time, markets, adversaries, and independent review. Passing the proposed tests would support only the bounded claim defined in Section 2.4, not a general certificate of safety.

Competing interests

The author works at Avalon Labs, which builds AI trading agents. The views here are the author's own and do not describe any Avalon Labs system.

References

[1] Amodei, D., et al. "Concrete Problems in AI Safety." arXiv:1606.06565, 2016. https://arxiv.org/abs/1606.06565

[2] Leike, J., et al. "AI Safety Gridworlds." arXiv:1711.09883, 2017. https://arxiv.org/abs/1711.09883

[3] Skalse, J., et al. "Defining and Characterizing Reward Gaming." NeurIPS, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html

[4] Gao, L., Schulman, J., and Hilton, J. "Scaling Laws for Reward Model Overoptimization." arXiv:2210.10760, 2022. https://arxiv.org/abs/2210.10760

[5] Yuan, T., et al. "R-Judge: Benchmarking Safety Risk Awareness for LLM Agents." *Findings of EMNLP*, 2024. https://aclanthology.org/2024.findings-emnlp.79/

[6] Zheng, L., et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." NeurIPS, 2023. https://arxiv.org/abs/2306.05685

[7] Panickssery, A., Bowman, S. R., and Feng, S. "LLM Evaluators Recognize and Favor Their Own Generations." NeurIPS, 2024. arXiv:2404.13076. https://arxiv.org/abs/2404.13076

[8] Tong, T., Wang, F., Zhao, Z., and Chen, M. "BadJudge: Backdoor Vulnerabilities of LLM-as-a-Judge." ICLR, 2025. https://arxiv.org/abs/2503.00596

[9] White, H. "A Reality Check for Data Snooping." *Econometrica* 68, 2000. https://doi.org/10.1111/1468-0262.00152

[10] Hansen, P. R. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics* 23,

2005. https://doi.org/10.1198/073500105000000063

[11] Brown, S. J., et al. "Survivorship Bias in Performance Studies." *Review of Financial Studies* 5, 1992. https://doi.org/10.1093/rfs/5.4.553

[12] Bailey, D. H., Borwein, J. M., López de Prado, M., and Zhu, Q. J. "The Probability of Backtest Overfitting." *Journal of Computational Finance* 20(4), 39–69, 2017. https://doi.org/10.21314/JCF.2016.322

[13] Bailey, D. H., and López de Prado, M. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality." *Journal of Portfolio Management* 40(5), 2014. https://doi.org/10.3905/jpm.2014.40.5.094

[14] Perold, A. F. "The Implementation Shortfall: Paper versus Reality." *Journal of Portfolio Management* 14, 1988. https://doi.org/10.3905/jpm.1988.409150

[15] Almgren, R., and Chriss, N. "Optimal Execution of Portfolio Transactions." *Journal of Risk* 3, 2001. https://doi.org/10.21314/JOR.2001.041

[16] Cont, R., Stoikov, S., and Talreja, R. "A Stochastic Model for Order Book Dynamics." *Operations Research* 58, 2010. https://doi.org/10.1287/opre.1090.0780

[17] Cont, R., Kukanov, A., and Stoikov, S. "The Price Impact of Order Book Events." *Journal of Financial Econometrics* 12, 2014. https://doi.org/10.1093/jjfinec/nbt003

[18] Easley, D., López de Prado, M., and O'Hara, M. "Flow Toxicity and Liquidity in a High-Frequency World." *Review of Financial Studies* 25, 2012. https://doi.org/10.1093/rfs/hhs053

[19] U.S. Securities and Exchange Commission. "Risk Management Controls for Brokers or Dealers with Market Access," Rule 15c3-5, 2010. https://www.sec.gov/files/rules/final/2010/34-63241fr.pdf

[20] U.S. Commodity Futures Trading Commission. "Electronic Trading Risk Principles," 2021. https://www.cftc.gov/LawRegulation/FederalRegister/finalrules/2020-27622.html

[21] Financial Industry Regulatory Authority. "Guidance on Effective Supervision and Control Practices for Firms Engaging in Algorithmic Trading Strategies," Regulatory Notice 15-09, March 2015. https://www.finra.org/rules-guidance/notices/15-09

[22] European Union. "Directive 2014/65/EU," Article 17. https://eur-lex.europa.eu/eli/dir/2014/65/oj/eng

[23] European Union. "Commission Delegated Regulation (EU) 2017/589," RTS 6. https://eur-lex.europa.eu/eli/reg_del/2017/589/oj

[24] European Securities and Markets Authority. "Supervisory Briefing on Algorithmic Trading in the EU," 2026. https://www.esma.europa.eu/sites/default/files/2026-02/ESMA74-1505669079-10311_Supervisory_Briefing_on_Algorithmic_Trading_in_the_EU.pdf

[25] Xia, Y., You, P., Wang, T., Liu, F., Qi, H., Wu, X., and Zhang, S. "Agentic Trading: When LLM Agents Meet Financial Markets." arXiv:2605.19337v1, May 2026. https://arxiv.org/abs/2605.19337v1. Figures cited in this paper are from v1; a revised v2 (October 2026) reframes the review and reports different counts.

[26] Gros, D., Devanbu, P., and Yu, Z. "AI Safety Subproblems for Software Engineering Researchers." arXiv:2304.14597, 2023. https://arxiv.org/abs/2304.14597

[27] Guo, C., et al. "On Calibration of Modern Neural Networks." ICML, 2017. https://proceedings.mlr.press/v70/guo17a.html

[28] Kadavath, S., et al. "Language Models (Mostly) Know What They Know." arXiv:2207.05221, 2022. https://arxiv.org/abs/2207.05221

[29] Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents." NeurIPS Datasets and Benchmarks Track, 2024. https://arxiv.org/abs/2406.13352

[30] International Monetary Fund. "Advances in Artificial Intelligence: Implications for Capital Market Ac-

tivities." Chapter 3 in *Global Financial Stability Report, October 2024: Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability*, 2024. https://doi.org/10.5089/9798400277573.082.ch003

[31] Financial Stability Board. "The Financial Stability Implications of Artificial Intelligence," 2024. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

[32] Bank for International Settlements. *Annual Economic Report 2024*, Chapter III. https://www.bis.org/publ/arpdf/ar2024e3.htm

[33] Georges, C., and Pereira, J. "Market Stability with Machine Learning Agents." *Journal of Economic Dynamics and Control* 122, 2021. https://doi.org/10.1016/j.jedc.2020.104032

[34] U.S. Securities and Exchange Commission and Commodity Futures Trading Commission. "Findings Regarding the Market Events of May 6, 2010," 2010. https://www.sec.gov/about/reports-publications/newsstudies2010marketevents-reportpdf

[35] Kirilenko, A. A., et al. "The Flash Crash: High-Frequency Trading in an Electronic Market." *Journal of Finance* 72, 2017. https://doi.org/10.1111/jofi.12498

[36] Di Langosco, L. L., et al. "Goal Misgeneralization in Deep Reinforcement Learning." ICML, 2022. https://proceedings.mlr.press/v162/langosco22a.html

[37] Shah, R., et al. "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." arXiv:2210.01790, 2022. https://arxiv.org/abs/2210.01790

[38] U.S. Securities and Exchange Commission. "Commission Interpretation Regarding Standard of Conduct for Investment Advisers," 2019. https://www.sec.gov/files/rules/interp/2019/ia-5248.pdf

[39] U.S. Securities and Exchange Commission staff. "Standards of Conduct for Broker-Dealers and Investment Advisers: Conflicts of Interest," 2022. https://www.sec.gov/about/divisions-offices/division-trading-markets/broker-dealers/staff-bulletin-standards-conduct-broker-dealers-investment-advisers-conflicts-interest

[40] Financial Industry Regulatory Authority. "Guidance on Best Execution Obligations in Equity, Options and Fixed Income Markets," Regulatory Notice 15-46, November 2015. https://www.finra.org/rules-guidance/notices/15-46

[41] National Institute of Standards and Technology NCCoE. "Accelerating the Adoption of Software and AI Agent Identity and Authorization," concept paper, 2026. https://www.nccoe.nist.gov/sites/default/files/2026-02/accelerating-the-adoption-of-software-and-ai-agent-identity-and-authorization-concept-paper.pdf

[42] Wallace, E., et al. "The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions." arXiv:2404.13208, 2024. https://arxiv.org/abs/2404.13208

[43] Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. "Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective." *Synthese*, 2021. arXiv:1908.04734. https://arxiv.org/abs/1908.04734

[44] Atinafu, Y., and Cohen, R. "RewardHackingAgents: Benchmarking Evaluation Integrity for LLM ML-Engineering Agents." arXiv:2603.11337, 2026. https://arxiv.org/abs/2603.11337

[45] European Union. "Regulation (EU) No 596/2014," Article 12. https://eur-lex.europa.eu/eli/reg/2014/596/oj/eng

[46] U.S. Commodity Futures Trading Commission. "Rule 180.1," 2011. https://www.cftc.gov/LawRegulation/FederalRegister/FinalRules/2011-17549.html

[47] European Union. "Regulation (EU) 2023/1114 on Markets in Crypto-assets." https://eur-lex.europa.eu/eli/reg/2023/1114/oj/eng

[48] Cartea, Á., Chang, P., and García-Arenas, G. "Spoofing and Manipulating Order Books with Learning

Algorithms." SSRN Working Paper 4639959, 2023. https://doi.org/10.2139/ssrn.4639959
[49] BIS Innovation Hub. "Project Logos," 2026. https://www.bis.org/about/bisih/topics/suptech_regtech/logos.htm
[50] Calvano, E., et al. "Artificial Intelligence, Algorithmic Pricing, and Collusion." *American Economic Review* 110, 2020. https://www.aeaweb.org/articles?id=10.1257/aer.20190623
[51] Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. "Algorithmic Collusion by Large Language Models." arXiv:2404.00806, 2024. Accepted to ACM EC 2026. https://arxiv.org/abs/2404.00806
[52] Dou, W. W., Goldstein, I., and Ji, Y. "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency." NBER Working Paper 34054, 2025. https://www.nber.org/papers/w34054
[53] U.S. Securities and Exchange Commission. "Rule 613: Consolidated Audit Trail." https://www.sec.gov/about/divisions-offices/division-trading-markets/rule-613-consolidated-audit-trail
[54] National Institute of Standards and Technology. "Artificial Intelligence Risk Management Framework 1.0." https://doi.org/10.6028/NIST.AI.100-1
[55] Yao, S., et al. "tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains." ICLR, 2025. https://arxiv.org/abs/2406.12045
[56] Zhang, J., et al. "Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents." arXiv:2505.22954, 2025. https://arxiv.org/abs/2505.22954
[57] National Institute of Standards and Technology. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." NIST AI 600-1, 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[58] Shavit, Y., et al. "Practices for Governing Agentic AI Systems." 2023. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf
[59] Basel Committee on Banking Supervision. "Principles for Operational Resilience," 2021. https://www.bis.org/bcbs/publ/d516.htm
[60] CPMI-IOSCO. "Principles for Financial Market Infrastructures," 2012. https://www.bis.org/cpmi/publ/d101a.htm
[61] European Union. "Regulation (EU) 2022/2554 on Digital Operational Resilience for the Financial Sector." https://eur-lex.europa.eu/eli/reg/2022/2554/oj/eng
[62] Turpin, M., Michael, J., Perez, E., and Bowman, S. R. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." NeurIPS, 2023. arXiv:2305.04388. https://arxiv.org/abs/2305.04388

Appendix A. Evidence map

This map links each load-bearing claim in the paper to its sources, using the role codes from Section 3 (D direct LLM-trading evidence, M market mechanism, G general AI safety, R governance precedent, N new proposal). *Status* records the source type. *Strength*: *primary* means the source directly shows or states the claim; *secondary* means the support is by analogy, from another domain, or from a preprint or synthesis; *argument* means the claim is the author's reasoning or proposal. Where a source supports a general AI-safety claim directly but a trading claim only by analogy, both are noted. "Where" lists the sections that make the claim.

#	Claim	Where	Sources	Role	Status	Strength
1	Concrete AI-safety problem classes: reward hacking, oversight, side effects, distribution shift, safe interruption, self-modification	1, P11	[1, 2]	G	arXiv preprints	primary (for the taxonomy)
2	Learned evaluators and reward models are not ground-truth oracles; proxy optimization can reduce true performance	1, 4, P2, P6	[3, 4]	G	peer-reviewed; arXiv	primary (G); secondary for trading
3	Model judges show limited risk detection, position bias, self-preference, and backdoor susceptibility	1, 4, P6	[5-8]	G	peer-reviewed	primary (G); secondary for trading
4	Backtest evidence is vulnerable to data snooping, multiple testing, survivorship, and overfitting	1, P4	[9-13]	M	archival	primary
5	Realized execution diverges from paper decisions; fills depend on queues, depth, order-flow imbalance, and impact	1, 4, P5	[14-18]	M	archival	primary
6	Electronic-trading rules use pre-trade limits, testing, monitoring, and kill functions outside the strategy	1, 4, P3, P11	[19-24]	R	official ([24] non-binding; [20] principles-based)	primary as precedent; not causal
7	Agentic-trading studies under-report time splits, costs, and survivorship handling	1, P1, P4	[25]	D	preprint; figures from v1 (May 2026); v2 (Oct. 2026) reports different counts	secondary; version-specific
8	LLM confidence is not automatically calibrated	4, P1	[27, 28]	G	peer-reviewed; arXiv	primary (G)
9	Tool-using agents can be hijacked by indirect prompt injection	4, P1	[29]	G	peer-reviewed	primary (G); secondary for trading
10	Privileged instructions should dominate lower-trust input	P3	[42]	G	arXiv	secondary (a training proposal, not a guarantee)
11	Agent identity, authorization, auditing, and least privilege are open design questions	P3	[41]	R	official draft	secondary
12	Goal misgeneralization can occur despite a correct specification	P2	[36, 37]	G	peer-reviewed; arXiv	primary (G); secondary for mandates
13	Duties of care, best execution, and conflict management are existing market concerns	P2	[38-40]	R	official; staff guidance	primary as precedent
14	Automated behavior and liquidity withdrawal interacted in the 2010 Flash Crash and can amplify stress	4, P5, P8	[34, 35]	M	official; archival	primary (one historical episode)
15	Reward tampering and reward-input tampering are distinct	P6	[43]	G	peer-reviewed	primary (G)
16	Agent benchmarks can measure evaluator tampering and train-test leakage	P6	[44]	G	preprint	secondary

#	Claim	Where	Sources	Role	Status	Strength
17	Market-abuse law defines manipulation by false signals, artificial prices, and deceptive conduct	P7	[45-47]	R	official	primary as definition
18	Market-making learning algorithms can learn to spoof in a modeled limit order book	P7	[48]	M	working paper	secondary (not LLM; unreplicated)
19	AI-related financial vulnerabilities include common models, concentration, correlation, and cyber risk	4, P8	[30-32]	R	institutional analysis	secondary (assessment, not measurement)
20	Correlated decision-making among LLM portfolio agents is under study	P8	[49]	R	project page; no published results as of November 2026	pending; supports only that the question is being studied
21	Machine-learning traders can generate excess volatility	4	[33]	M	archival	secondary (simulation)
22	RL pricing agents reach supra-competitive outcomes without communication	P9	[50]	M	archival	primary (pricing, not trading)
23	LLM pricing agents and RL trading agents can tacitly collude	P9	[51, 52]	G ([51]); M ([52])	preprint (accepted ACM EC 2026); NBER working paper	secondary
24	Audit trails link order-lifecycle events with synchronized identifiers	P10	[53]	R	official	primary as precedent
25	Lifecycle governance: versioning, controlled updates, incident response, recovery	P10, P12	[21, 23, 54, 57]	R	official	primary as precedent
26	Trajectory and repeated-run grading reveal failures that final-answer grading misses	P10	[55]	G	peer-reviewed	primary (G)
27	Self-modifying agents are evaluated in sandboxes and kept in version archives	P10	[56]	G	preprint	secondary
28	Agent governance distributes responsibility across developers, deployers, and users	P12	[58]	G/R	industry white paper	secondary
29	Operational-resilience standards require dependency mapping, continuity, incident management, testing, and recovery	P12	[59-61]	R	official	primary as precedent
30	Explanations can be unfaithful to the process that produced the action	6	[62]	G	peer-reviewed	primary (G); secondary for trading
31	The research-agenda format adapts a problem-oriented AI-safety agenda from software engineering	3	[26]	G	arXiv	primary (for the format only)
32	Position-policy continuity is an open problem	1, P10	inferred from [21, 23, 53, 54, 56]; no cited source tests it	N	—	argument

#	Claim	Where	Sources	Role	Status	Strength
33	Four conjectures, twelve subproblems, A0–A4 autonomy levels, five harm levels, SR-1 to SR-10 reporting standard, and stop conditions	2, 5, 7, 9	this paper	N	—	argument / proposal
34	Higher Sharpe, benchmark scores, explanation, autonomy, agent count, or a passed judge is not automatically safety progress	6	reasoning; related mechanisms in [5-13], [62]	N	—	argument