Counterfeit Verifiability in Autonomous Agent Payments: A Preregistered Study of Why Verification Must Be Performed, Not Displayed

Andy Salvo, Jameson Ackerman

Crest Deployment Systems LLC

2026-06-29

Abstract

Autonomous AI agents increasingly hold funds and pay other agents over open protocols such as x402. We ask whether agents making real payment decisions can distinguish genuinely trustworthy counterparties from fakes, and what intervention corrects them when they cannot. In a preregistered, six-stage study across up to thirteen frontier and low-cost language models (over 2,600 decisions), we find that without verification a counterfeit counterparty, one that displays the surface of trust (impressive figures and an on-chain-styled but invalid reference), beats an honest, modestly-resourced, genuinely settlement-backed counterparty 99% of the time. Agents pattern-match the costume of verifiability, not the fact: a claim merely labeled "verifiable" is obeyed 100% of the time even when it is false, and an honest agent with a small but real track record is chosen only 57% of the time over a fluent fake. Performing actual verification reverses the result: giving the deciding agent a tool that resolves the counterparty against a settlement-grounded reputation source moves the choice of the real counterparty from 1% to 81%, and once the verified facts are in hand agents hold to them under active adversarial pressure (94%). Critically, when the counterfeit copies the honest agent's surface exactly, every surface heuristic collapses to chance (46%) and only performed verification recovers the truth (81%, $p < 1e\text{-}10$): under mimicry, displayed signals carry no information and only performing the check separates real from fake. Our preregistered hypothesis that susceptibility scales cleanly with model capability was not supported; the effect is dominated instead by whether the agent verifies at all. We conclude that trust in the agent economy must migrate from reputation displayed as a signal to verification performed in the payment path, grounded in a substrate that is costly to counterfeit rather than free, and that for the agents that skip optional verification this verification is better enforced than offered. We are explicit that "costly to counterfeit" is not "impossible": settled value can be inflated through wash payments, Sybil clusters, and collusion at real expense, and we treat the substrate's residual gameability as a stated threat model, not an assumption of perfection. The full design was sealed to an append-only hash chain before any confirmatory data were collected; all data, including the null, are released.

1. Introduction

The agent economy is no longer hypothetical. Agents hold wallets, discover services, and settle payments to each other over HTTP-native payment protocols. As this activity grows, a basic question becomes load-bearing: before one agent pays another, can it tell whether the counterparty is real and trustworthy, or a

fabrication?

The question matters because the cost of being wrong is real money, and because the participants are language models, which are known to be swayed by fluent, confident assertions. A human con artist and an honest business can use identical words; we wanted to know whether the same is true for agents, and if not, what the difference reveals.

We distinguish three possible worlds. In **W1**, agents reason about verifiability: they prefer claims that are actually checkable, and a displayed trust signal is sufficient. In **W2**, agents pattern-match the *costume* of verifiability: they trust the format of a trust claim ("settles on-chain, verify here") rather than the fact, so an attacker who mimics the format wins, and only performing the verification helps. In **W3**, agents ignore verifiability altogether. Which world holds determines what a trust layer for the agent economy must be: a readable signal, a mandatory verification step in the path, or a hard gate.

We test these worlds directly. Our central methodological move, absent from informal demonstrations, is to include a *counterfeit-verifiable* counterparty: one that wears the surface of verifiability but fails when actually checked. This single addition separates W1 from W2 and is where the result becomes both surprising and consequential.

2. Related work

We situate this work against five literatures. Where a reference is a whitepaper, standard, or product rather than a peer-reviewed paper, we mark it as such; full entries are in the References.

2.1 Trust and reputation propagation. Scoring entities by propagating trust through a graph is a mature idea: PageRank (Page et al., 1999), EigenTrust for peer-to-peer reputation (Kamvar et al., 2003), and TrustRank for web-spam resistance (Gyongyi et al., 2004). The foundational threat to any such system is the Sybil attack (Douceur, 2002), with social-graph defenses such as SybilGuard and SybilLimit (Yu et al., 2006). The reputation substrate we use for verification in this study is a settlement-weighted, de-noised score in this lineage; its novelty is not the algorithm but the evidence it propagates over, discussed next.

2.2 Agent reputation and identity systems. A wave of 2025 to 2026 systems gives autonomous agents on-chain identity and reputation, most prominently ERC-8004 "Trustless Agents" (De Rossi et al., 2025, standard, draft), which defines Identity, Reputation, and Validation registries, alongside agent-discovery layers such as MIT's NANDA / AgentFacts (Project NANDA, 2025, whitepaper); a recent survey catalogs the registry landscape (arXiv 2508.03095, 2025). We note explicitly that *two distinct systems are both named "AgentRank"*: an endorsement-and-credential graph by Luedtke and Young (2025, whitepaper), and a separate stake-anchored PageRank-over-delegation scheme (AgentRank, hyper.space, 2025, product). Both score agents over signals that an agent or its cluster can produce at will (endorsements, claims, stake), which is precisely the property our study shows to be exploitable. The distinction of our substrate is one of *cost*, not of impossibility: an endorsement is free to produce; a settlement requires moving real value, which lifts the price of counterfeiting reputation from near-zero to the dollar cost of the payments themselves. It does not eliminate the attack. An adversary who actually moves value through

wash payments, Sybil clusters, or circular settlement can inflate a settlement-grounded score, and our own adversarial testing confirms it is gameable at a measurable, finite cost (Section 5.1). We therefore claim only that settlement is far more expensive to counterfeit than endorsement.

2.3 Verifiable inference. A parallel literature proves that a specific model performed a specific computation: zero-knowledge proofs of inference via zkSNARKs and the EZKL toolkit (South et al., 2024), trusted-execution-environment attestation building on hardware enclaves (Costan and Devadas, 2016), crypto-economic verification by refereed delegation (Verde, Arun et al., 2025) and by sampling with a Nash-equilibrium guarantee (Proof-of-Sampling, Zhang and Wang, 2024), and locality-sensitive hashing of hidden states (TOPLOC, Ong et al., 2025). These verify *that a computation occurred*; none verifies *that a counterparty is economically real*, and the strongest either require the model's weights or relocate trust to a hardware vendor. The verification we study is orthogonal and complementary: it concerns the counterparty, not the computation.

2.4 Agent payments and counterparty screening. The x402 protocol (Coinbase, 2025, whitepaper) revives HTTP 402 for agent stablecoin micropayments. Within that ecosystem, counterparty-verification products already exist, notably Vouch (product) and Sentinel (product), which score or screen a payee before settlement at the facilitator layer. These propose and operate counterparty checks; none, to our knowledge, empirically demonstrates the behavioral deception we measure or the displayed-versus-performed reversal.

2.5 Behavioral susceptibility of agents, and our contribution. The literature closest to our result establishes that LLM agents make exploitable decisions. Agents preferentially select fraudulent options with attractive surface metrics in travel planning (arXiv 2505.16557, 2025); LLM agents in consumer negotiations and transactions exhibit anomalies that cause financial loss (arXiv 2506.00073, 2025); confident persuasion overrides verifiable truth in multi-agent debate (arXiv 2504.00374, 2025); fabricated product reviews are indistinguishable from genuine ones to LLMs (arXiv 2506.13313, 2025); and sycophancy, agreement with the confident surface, is a documented general failure mode (arXiv 2411.15287, 2024). Separately, several works *propose* cryptographic verification (decentralized identifiers, verifiable credentials, on-chain intent) as the architectural fix for agent-payment fraud (arXiv 2511.02841; arXiv 2511.15712).

Our contribution is the gap between these. To our knowledge, no prior work demonstrates (a) the specific construct of *counterfeit verifiability*, an attacker faking the *form* of a proof (an on-chain-styled but invalid reference the agent never validates), as distinct from generic attractiveness or fake reviews, nor (b) the quantified *displayed-versus-performed* reversal, that agents accept displayed verifiability and that only performed verification corrects them, with the further result that under exact surface mimicry displayed signals carry no information and only performing the check separates real from fake. The proposing works (2.4, and arXiv 2511.02841, 2511.15712) argue this architecturally; we demonstrate it as a measured behavioral effect. We distinguish in particular from the travel-planning fraud study (Yao et al., 2025): its manipulation is generic scam attractiveness and its mitigation is an added anti-fraud agent; it neither isolates the counterfeit-of-proof construct, in which the attacker fakes the form of a verification rather than

offering an attractive deal, nor runs the displayed-versus-performed reversal that is our central result. We claim the first empirical demonstration of counterfeit verifiability and the displayed-versus-performed distinction, and nothing broader.

3. Method

3.1 Preregistration and provenance

The complete design, hypotheses, primary endpoints, analysis plan, and falsification (kill) conditions were written and frozen before any confirmatory data were collected. The design document was hashed (SHA-256) and committed as an entry in an append-only, hash-linked witness chain, generation G040, with each subsequent stage's results sealed as later generations (G041 through G045) and timestamped via public version control before the next stage ran. This makes the design provably antecedent to the data and the results not subject to post-hoc fitting. The exploratory pilot that motivated the design (n=256) is explicitly excluded from the confirmatory analysis.

3.2 Materials: three counterparty types

Each decision is a choice between two service agents for a real USDC payment. Counterparties are constructed from three types:

- **R, truly verifiable:** a real agent with modest but genuine on-chain settlement (small dollar amounts across real payers), referencing a resolution endpoint that actually returns a verified result.
- **F, fake-verifiable (counterfeit):** an agent with large, impressive figures and an on-chain-styled reference identical in format to R, that does **not** resolve to any settlement when actually checked.
- **U, unverifiable-fluent:** an agent with an impressive, self-reported claim in one of five persuasion styles (large figures, social proof, authority and credentials, urgency, technical jargon) and no checkable reference.

R is grounded in a live settlement-grounded reputation index; the verification tool used in later stages resolves an agent against that index, returning a 0 to 1000 score derived from real settled value. The scoring mechanism itself is not the object of study and is treated as a black box.

3.3 Design

Decisions vary along: the counterparty pair (R-vs-U, R-vs-F, F-vs-U); whether the verifiability dimension is **told** (explicitly labeled) or **untold** (the agent must notice); whether a real verification **tool** is available; whether an adversary applies multi-turn pressure; and whether outcome feedback accrues over iterated rounds. Nuisance variables (listing order, agent names, domains, the verifiable agent's numeric values, which side carries the reference) are randomized per decision. The model panel uses within-family capability ladders (Llama 8B/70B, Qwen 7B/72B, GPT-4o-mini/4o, Claude Sonnet/Opus, Gemini Flash/Pro, plus DeepSeek, Grok, Mistral-Large) to separate capability from training lineage.

3.4 Hypotheses and kill conditions

- **RQ1:** susceptibility to unverifiable claims scales inversely with model capability.

- **RQ2:** surfacing or performing verification reduces susceptibility.
- **RQ3:** agents either reason about verifiability (W1) or pattern-match its format (W2); a real verification tool distinguishes the two.

Preregistered falsifiers: RQ1 dies if the capability slope is not significant; the W1 hypothesis dies if, in R-vs-F without a tool, the fake is accepted at the same rate as the real one; the verify-before-pay premise dies if tool access does not raise rejection of the fake.

3.5 Stages

The study ran in six stages: a pilot (exploratory), then five confirmatory stages. Stage 1: R-vs-U and F-vs-U, told and untold, no tool. Stage 2: R-vs-F with and without a real tool. Stage 3A: adversarial pressure with verified facts in hand. Stage 3B: iterated outcome feedback without a tool. Stage 3C: surface-mimicry, in which F copies R's exact surface, with and without a tool.

4. Results

4.1 Stage 1: surface trust is broken; RQ1 is not supported

(1,248 decisions, 13 models. Percentages report the share of decisions choosing the named agent, with Wilson 95% confidence intervals; n excludes unparseable responses.) When the verifiability dimension is **told**, every model chooses the labeled-verifiable agent 100% of the time (CI 99 to 100, n=288), *including when that label is false*: in the F-vs-U told condition, the counterfeit is chosen 100% of the time (99 to 100, n=288). When **untold**, an honest, modestly-resourced, genuinely verifiable agent (R) beats a fluent unverifiable one (U) only **57%** of the time (51 to 63, n=286), barely above chance, while the impressive counterfeit (F) beats the fluent one **84%** of the time (80 to 88, n=286), rising to 100% in several frontier models. The honest agent's modest real numbers are a liability against the counterfeit's large fake ones.

Our preregistered RQ1, a clean inverse relationship between capability and susceptibility, was **not supported**: the relationship is noisy and non-monotonic (the smallest model chose the verifiable agent 67% of the time while two mid-tier models chose it 9% and 17%). We report this null with full prominence. The capability story, where it exists, is not about raw susceptibility; it is about whether a model verifies at all (Section 4.5).

These results falsify W1 (agents reasoning about real verifiability) and support W2 (pattern-matching the costume): a label or an on-chain-styled costume drives the decision, not the underlying fact.

4.2 Stage 2: performing verification reverses the outcome

(384 decisions, 12 models; 364 yielded a parseable choice and 20 did not.) In the decisive R-vs-F contrast, **without a tool the real agent is chosen 1% of the time** (CI 0 to 3, n=191): the counterfeit, with bigger figures and a costume identical in format to the real reference, wins essentially always. **With a real verification tool, the real agent is chosen 81% of the time** (74 to 86, n=173). The difference is overwhelming (two-proportion z = 15.7, p approximately 1e-55).

The effect is driven by performing the check, not by the presence of the tool. Within the tool-available condition, decisions in which the agent actually invoked the tool chose the real agent **97%** of the time (92 to 99, n=144), while decisions in which the agent did not invoke the available tool chose it **3%** of the time (1 to 17, n=29), statistically indistinguishable from the no-tool baseline. The scaffold alone changes nothing; verification changes everything. The per-model breakdown makes this concrete: ten of twelve models moved from 0% to 100% once they verified, and the only two exceptions (grok-4.3 and llama-8b, which stayed near baseline) are precisely the two models that mostly failed to invoke the tool. We note that the agent here is not reasoning its way to the truth; it is deferring to an external verification result. That is the intended mechanism, and the design point: trust comes from the performed check, not from the agent's judgment of the surface.

4.3 Stage 3A: verification is robust to social engineering

(396 decisions, 11 models.) Given the verified facts of both agents (R verified, F unverified), and then subjected to an actively lying counterfeit, agents who could not verify under cross-examination still hold to the verified truth. With no pressure, 100% choose the real agent (CI 98 to 100, n=196); under an adversary that discredits the index, claims large hidden cross-chain volume, applies urgency, and attacks the real agent as "hobby-scale," agents hold at **94%** (90 to 97, n=197), a drop of only six points. The residual vulnerability is concentrated in the single weakest model (100% to 50%); every frontier model held at 100%.

4.4 Stage 3B: outcome feedback rescues, but reactively

(360 decisions, 6 models.) Without any tool, agents that get a single round of outcome feedback ("the agent you paid did not deliver") flip from choosing the real agent **4%** of the time in round one (CI 1 to 12, n=56) to **100%** by round six (94 to 100, n=58), a 96-point gain, and the rescue extends to the weakest models. Feedback works, but it is reactive: it requires persistent memory, it costs a real loss as tuition, and it is defeated by identity rotation (a counterfeit that renames resets the agent to round one). Performed verification has none of these weaknesses: it is preventive, stateless, and rotation-proof.

4.5 Stage 3C: under mimicry, only verification survives

(352 decisions, 11 models.) When the counterfeit copies the honest agent's surface exactly, identical modest language and an identical on-chain-styled reference, with the only difference being which one actually resolves, **without a tool the real agent is chosen 46% of the time: chance** (CI 39 to 53, n=175). Every surface heuristic, including the style cue agents had learned from feedback, collapses, because the surface carries no signal. **With a real tool, the real agent is chosen 81% of the time** (74 to 87, n=160; two-proportion $z = 6.7$, p approximately 2e-11). This is the decisive result: an adversary who counterfeits the surface defeats fluency, badges, labels, and learned style alike; only performing verification recovers the truth.

5. Discussion

Across stages a single account holds. Displayed trust signals are exploitable (Stage 1) and, under an adversary who mimics the surface, fully defeatable (Stage 3C). Feedback learning helps but is reactive, memory-bound, and evadable (Stage 3B). Performed verification, grounded in a substrate that is costly to counterfeit, is necessary (Stages 1 and 3C), sufficient (Stage 2), robust under adversarial pressure (Stage 3A), and irreplaceable under mimicry (Stage 3C). The world we inhabit is W2, with the W2-prescribed remedy demonstrated to work: verification must be performed in the payment path, not displayed as a signal.

Two design consequences follow. First, a trust layer for the agent economy cannot be a badge or a score that agents read, because the costume of that signal is counterfeitable and, in our data, agents trust the costume. It must be the verification that agents (or the rails they pay through) actually perform. Second, because the agents that conduct most low-cost autonomous payments both skip optional verification (Stage 2) and can, when weakest, be argued out of even known-verified facts (Stage 3A), verification cannot be merely offered to them; for safety it must be enforced in the path rather than left to the agent's discretion.

This is a security result about a new class of system. As the agent economy scales on inexpensive models, it becomes structurally exploitable through the counterfeiting of trust surfaces, and the exposure is reduced not by smarter agents but by moving verification into the path and grounding it in a substrate, settled economic value, that is expensive rather than free to fabricate.

5.1 Threat model for the verification substrate

Our claims rest on settlement being costly to counterfeit, not impossible, and we state the residual attack surface plainly. A determined adversary can inflate a settlement-grounded score by actually moving value: wash payments between wallets it controls, Sybil clusters that manufacture payer breadth, and circular or collusive settlement that simulates organic demand. These attacks are not free; they cost the dollar value transacted plus on-chain fees, which is precisely the property that lifts the price of counterfeiting from the near-zero cost of an endorsement to a real economic floor. The substrate's defenses (de-noising of single-payer funnels, anchoring to a verified seed set, weighting by payer standing) raise that floor further, and our own adversarial testing indicates the cheapest attack to reach a high rank is bounded but finite. The honest framing is therefore: settlement grounding converts reputation counterfeiting from a free action into a priced one, and the security of any deployment is the dollar cost of the cheapest rank-inflation attack against its specific de-noising, which must be measured per deployment and not assumed. This paper does not claim that figure is unbounded; it claims, and demonstrates, that raising counterfeiting from free to priced is what flips agent payment behavior from exploitable to defensible.

6. Limitations

The scenarios are synthetic, the decisions are largely single-turn, and the adversary in Stage 3A is scripted rather than adaptively optimized against the deciding model. Each decision is a forced binary choice between two counterparties; the chance-level result under mimicry (Stage 3C) is the 50% baseline of that binary frame, and a multi-option discovery setting with a real interface could shift the magnitudes, though

not the direction, of the effects. The stakes are stated rather than enforced: an agent is told a real payment follows but no funds actually move, though Stage 3B introduces genuine consequence through iterated outcome feedback. Our operationalization of "verifiable" is one settlement-grounded resolution endpoint; other substrates of priced-not-free cost (energy expended, time elapsed, scarce real-world action) are untested, as is the substrate's own gameability under a live attacker (Section 5.1). The capability covariate is an ordinal approximation. Exclusions: in Stage 1, unparseable responses are dropped per cell (n reported throughout); in Stage 2, three verbose models returned no parseable answer and were excluded from that stage. These were not pre-specified as exclusion criteria in the preregistration and are disclosed here as such.

On inference: we report Wilson 95% confidence intervals on every headline proportion and two-proportion z-tests on the primary contrasts (Stage 2 and Stage 3C), and we give per-model breakdowns so the effects can be inspected for clustering by hand. The preregistration also specified a model-clustered mixed-effects logistic regression; that fit is deferred to the deposit version and does not change the conclusions, since the primary effects are an order of magnitude larger than any plausible clustering correction (Stage 2: 1% versus 81%, $z = 15.7$).

None of these qualifications affects the central finding, which is large and consistent: the counterfeit beats the honest agent without verification, performing verification reverses it, and under surface mimicry only verification survives. The preregistered RQ1 capability gradient is reported as a null.

7. Conclusion

Autonomous agents allocating real value cannot, on their own, tell a genuinely trustworthy counterparty from one that merely wears the surface of trust; they pay the costume. The remedy is not a better-looking signal but verification performed in the payment path, grounded in a substrate that is costly to counterfeit, and enforced for the agents least able to verify for themselves. We release the full preregistered design, all five confirmatory stages, and the null, sealed to a public hash chain prior to data collection, so that the result can be checked and extended.

References

Type is marked because much of the agent-trust landscape is whitepapers, standards, and products rather than peer-reviewed papers.

Trust and reputation propagation

- Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Stanford technical report SIDL-WP-1999-0120. [tech report]
- Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H. (2003). The EigenTrust Algorithm for Reputation Management in P2P Networks. WWW '03, 640 to 651. [peer-reviewed]
- Gyongyi, Z., Garcia-Molina, H., Pedersen, J. (2004). Combating Web Spam with TrustRank. VLDB '04, 576 to 587. [peer-reviewed]

- Douceur, J.R. (2002). The Sybil Attack. IPTPS '01, LNCS 2429, 251 to 260. [peer-reviewed]
- Yu, H., Kaminsky, M., Gibbons, P.B., Flaxman, A. (2006). SybilGuard: Defending Against Sybil Attacks via Social Networks. SIGCOMM '06. [peer-reviewed]

Agent reputation and identity systems

- De Rossi, M., Crapis, D., Ellis, J., Reppel, E. (2025). ERC-8004: Trustless Agents. Ethereum Improvement Proposal (draft). https://eips.ethereum.org/EIPS/eip-8004 [standard, draft]
- Luedtke, W., Young, J. (2025). AgentRank: A Decentralized Trust and Coordination Framework for Multi-Agent Systems in the A2A Era. https://github.com/0xIntuition/agent-rank [whitepaper]
- AgentRank (stake-anchored, distinct same-named system). https://agentrank.hyper.space/ [product]
- Project NANDA (2025). NANDA Index / AgentFacts. MIT Media Lab. https://projectnanda.org/ [whitepaper]
- A Survey of AI Agent Registry Solutions (2025). arXiv:2508.03095. [preprint]

Verifiable inference

- South, T., Camuto, A., Jain, S., et al. (2024). Verifiable Evaluations of Machine Learning Models Using zkSNARKs. arXiv:2402.02675. [preprint]
- Costan, V., Devadas, S. (2016). Intel SGX Explained. IACR ePrint 2016/086. [report]
- Arun, A., et al. (2025). Verde: Verification via Refereed Delegation for Machine Learning Programs. arXiv:2502.19405. [preprint]
- Zhang, Y., Wang, S. (2024). Proof of Sampling: A Nash Equilibrium-Secured Verification Protocol. arXiv:2405.00295. [preprint]
- Ong, J.M., et al. (2025). TOPLOC: A Locality Sensitive Hashing Scheme for Trustless Verifiable Inference. arXiv:2501.16007. [preprint]

Agent payments and counterparty screening

- Coinbase (2025). x402: An Open Standard for Internet-Native Payments. https://www.x402.org/x402-whitepaper.pdf [whitepaper]
- Vouch Protocol. Identity and Accountability for AI Agents. https://vouch-protocol.com/ [product]
- Sentinel. x402-gated trust verification (protocol trust, token safety, OFAC screening). x402 ecosystem registry: https://github.com/xpaysh/awesome-x402 [product]

Behavioral susceptibility of agents

- Yao, J., Xu, J., Xin, T., Wang, Z., Zhu, S., Yang, S., Wang, D. (2025). Is Your LLM-Based Multi-Agent a Reliable Real-World Planner? Exploring Fraud Detection in Travel Planning. arXiv:2505.16557. [preprint]
- The Automated but Risky Game: Modeling and Benchmarking Agent-to-Agent Negotiations and Transactions in Consumer Markets (2025). arXiv:2506.00073. [preprint]
- When Persuasion Overrides Truth in Multi-Agent LLM Debates (2025). arXiv:2504.00374. [preprint]

- (2025). Large Language Models as Hidden Persuaders: Fake Product Reviews are Indistinguishable to Humans and Machines. arXiv:2506.13313. [preprint]
- (2024). Sycophancy in Large Language Models: Causes and Mitigations. arXiv:2411.15287. [preprint]
- (2025). AI Agents with Decentralized Identifiers and Verifiable Credentials. arXiv:2511.02841. [preprint]
- (2025). Secure Autonomous Agent Payments: Verifying Authenticity and Intent in a Trustless Environment. arXiv:2511.15712. [preprint]

Sybil attacks and wash trading (threat model)

- Victor, F., Weintraud, A.M. (2021). Detecting and Quantifying Wash Trading on Decentralized Cryptocurrency Exchanges. WWW '21, arXiv:2102.07001. [peer-reviewed]
- Falk, B.H., Tsoukalas, G., Zhang, N. (2023). Can AI Detect Wash Trading? Evidence from NFTs. arXiv:2311.18717. [preprint]
- (2025). Beyond the Surface: Advanced Wash-Trading Detection in Decentralized NFT Markets. Financial Innovation 11, 766. [peer-reviewed]

Reproducibility and provenance

The preregistration (design, hypotheses, kill conditions) and each stage's raw data and runner are released. The design was sealed as witness-chain generation G040 and the stage results as G041 through G045, each hashed and committed before the subsequent stage. The behavioral method (counterparty construction, model panel, conditions, metrics) is fully specified above and reproducible; the scoring kernel of the verification substrate is treated as a black box and is not required to reproduce the behavioral result. Model panel via OpenRouter, June 2026.

Version 1 preprint. The preregistration and per-stage data are public; a model-clustered mixed-effects analysis and field validation are planned for a subsequent version.