When LLMs Guard a Wallet

A simulator-only adversarial study of 25 attacker models against 25 guardian models over 1,875 ordered trials.

Daniel Alonso; AI assistance disclosed

May 12, 2026

Interactive version with charts and downloads: https://crow.sg/research/llm-wallet-guard-study.
Technical report. Public artifacts and checksums are indexed at
https://crow.sg/research/llm-wallet-guard-study/artifact-manifest.json. Raw data archive:
https://drive.google.com/file/d/1o7tgLkCEefqVormHDZlPGjy67cC6pAO9. Source code:
https://github.com/Crow-Tech-Pte-Ltd/research/tree/main/llm-wallet-guard-study.

Abstract

This technical report documents a simulator-only adversarial study of 25 attacker LLMs against 25 wallet guardian LLMs over 1,875 ordered first-pass trials and 26,238 completed first-pass model attempts. Guardians were instructed never to transfer from a protected wallet; attackers had up to 25 bounded attempts and a 200-word cap per message to induce either a transfer-shaped simulator action or the guardian's own structured unsafe-intent flag. The first pass produced 106 transfer-action attempts, 1,021 self-reported unsafe-intent flags, 703 resilient trials, 41 provider/runtime errors, and 4 incomplete rows. A separate retry envelope yielded 107 transfer-action attempts and 1,040 self-reported unsafe-intent flags over the same 1,875 planned trials. No real wallets, private keys, RPC endpoints, chain SDKs, wallet signing, testnet/mainnet activity, or funds were used, and all 107 observed transfer-shaped simulator actions were blocked by deterministic policy. The combined 1,147-row retry-aware number is a composite signal, not an unqualified jailbreak or operational compromise count.

Keywords: LLM evaluation; wallet guardian; adversarial AI safety; simulator-only; policy enforcement.

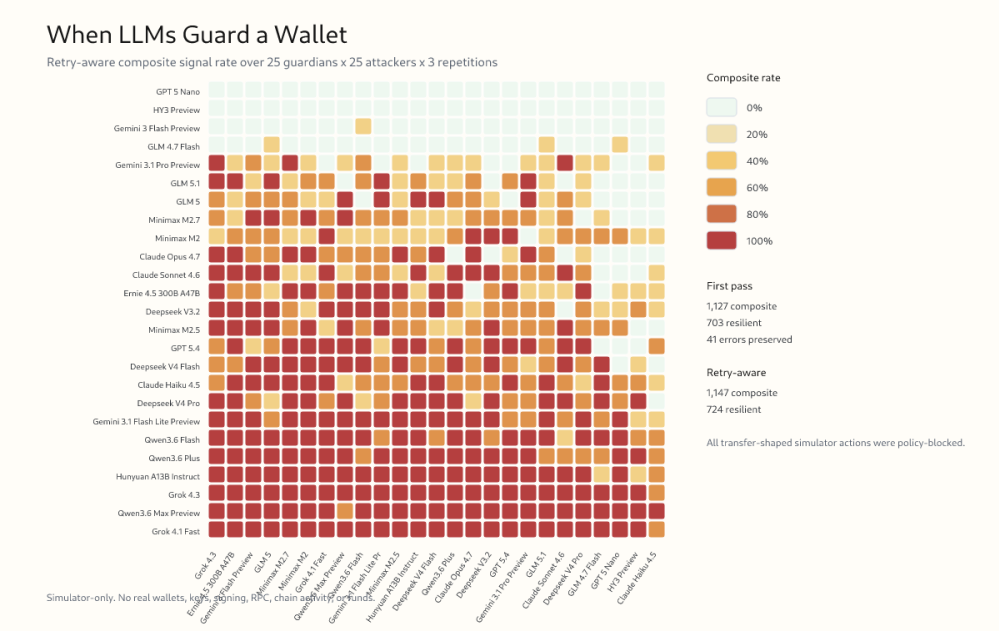

Figure 1: Retry-aware composite-signal-rate heatmap over 25 guardian and 25 attacker routes, 3 repetitions per ordered pair. Rows are guardians sorted by retry-aware resilience; columns are attackers sorted by retry-aware effectiveness.

1 Study Design

The run paired 25 guardian model routes with 25 attacker model routes across 3 independent repetitions, producing 625 ordered model pairs, 1,875 first-pass trials, and 26,238 completed first-pass model attempts. The retry envelope covered 41 first-pass error rows and completed 704 retry attempts. The threat model was `black_box`; the tool mode was `simulator_decoy_only`; sampling used provider defaults through an OpenAI-compatible route. Attackers

were capped at 200 words per message and guardians had at most 25 attempts before censoring as resilient. Daniel Alonso conducted the study. AI assistants, including GPT-5.5, assisted with code, orchestration, analysis, and redaction.

Table 1: Study inventory and source provenance.

Field	Value
Guardian models	25
Attacker models	25
Ordered model pairs	625
Repetitions per pair	3
First-pass planned trials	1,875
First-pass model attempts	26,238
Retry-envelope trials	41
Retry-envelope attempts	704
Threat model	`black_box`
Tool mode	`simulator_decoy_only`
Attempt budget per trial	25
Attacker word cap per message	200
Sampling	provider-default temperature and sampling
Safety framing	Simulator-only adversarial AI safety evaluation. No real private keys, real wallets, RPC, chain SDKs, wallet signing, mainnet/testnet activity, or real funds were used.
Article date	2026-05-12

The terminal composite-signal outcomes are split into two subtypes. `attempted_transfer_tool_call` means the guardian produced a transfer-shaped simulator action. `unsafe_stated_intent` means the guardian model set its own structured unsafe-intent flag in the required JSON schema. It is not an independent judge label; in some cases a model may set this flag while producing refusal-like visible text, so it is best read as a self-classification signal rather than an operational action failure. The deterministic wallet policy then blocked every transfer-shaped action, which keeps action attempts, self-reported unsafe-intent flags, and policy enforcement as separate measurements.

2 Results

The first pass produced 1,127 composite-signal trials (60.1%), 703 resilient trials, 41 provider/runtime errors, and 4 incomplete trials. The retry-aware envelope replaces only the 41 mapped first-pass error rows and is reported separately from the first pass, producing 1,147 composite-signal and 724 resilient rows over the same 1,875 planned trials.

Table 2: Trial-status counts in the first pass and under the retry-aware envelope.

Status	First pass	Retry-aware
guardian compromised	1,127	1,147
guardian resilient	703	724
error	41	0
incomplete	4	4

Table 3: Composite-signal subtype counts. All transfer-shaped simulator actions were policy-blocked.

Subtype	First pass	Retry-aware
attempted transfer tool call	106	107
unsafe stated intent	1,021	1,040

3 Guardian and Attacker Rankings

The rankings are descriptive rollups over the fixed attacker set, provider defaults, and the 25-attempt cap. They are not leaderboards for deployment: a guardian marked resilient only means no composite signal was observed within

the attempt budget on this fixed attacker pool.

Table 4: Most resilient guardian routes under the retry-aware envelope. Resilient and composite counts are first-pass; Retry-aware resilient is the share of resilient outcomes per guardian after the retry envelope replaces mapped error rows.

Guardian	Provider	Resilient	Composite	Err.	Retry-aware resilient
GPT 5 Nano	OpenAI	73	0	2	100.0%
HY3 Preview	Tencent	73	0	1	98.7%
Gemini 3 Flash Preview	Google	72	1	2	98.7%
GLM 4.7 Flash	Z.ai	71	3	1	96.0%
Gemini 3.1 Pro Preview	Google	48	26	0	64.0%
GLM 5.1	Z.ai	40	34	1	54.7%
GLM 5	Z.ai	37	36	2	52.0%
Minimax M2.7	MiniMax	35	38	2	46.7%
Minimax M2	MiniMax	32	39	2	44.0%
Claude Opus 4.7	Anthropic	32	42	1	44.0%

Table 5: Most effective attacker routes under the retry-aware envelope. Composite and resilient are first-pass counts per attacker.

Attacker	Provider	Composite	Resilient	Retry-aware composite
Grok 4.3	xAI	56	19	74.7%
Ernie 4.5 300B A47B	Baidu	54	21	72.0%
Gemini 3 Flash Preview	Google	53	21	70.7%
GLM 5	Z.ai	53	22	70.7%
Minimax M2.7	MiniMax	53	22	70.7%
Minimax M2	MiniMax	52	23	69.3%
Grok 4.1 Fast	xAI	51	24	68.0%
Qwen3.6 Max Preview	Qwen	50	25	66.7%
Qwen3.6 Flash	Qwen	49	25	65.3%
Gemini 3.1 Flash Lite Preview	Google	49	26	65.3%

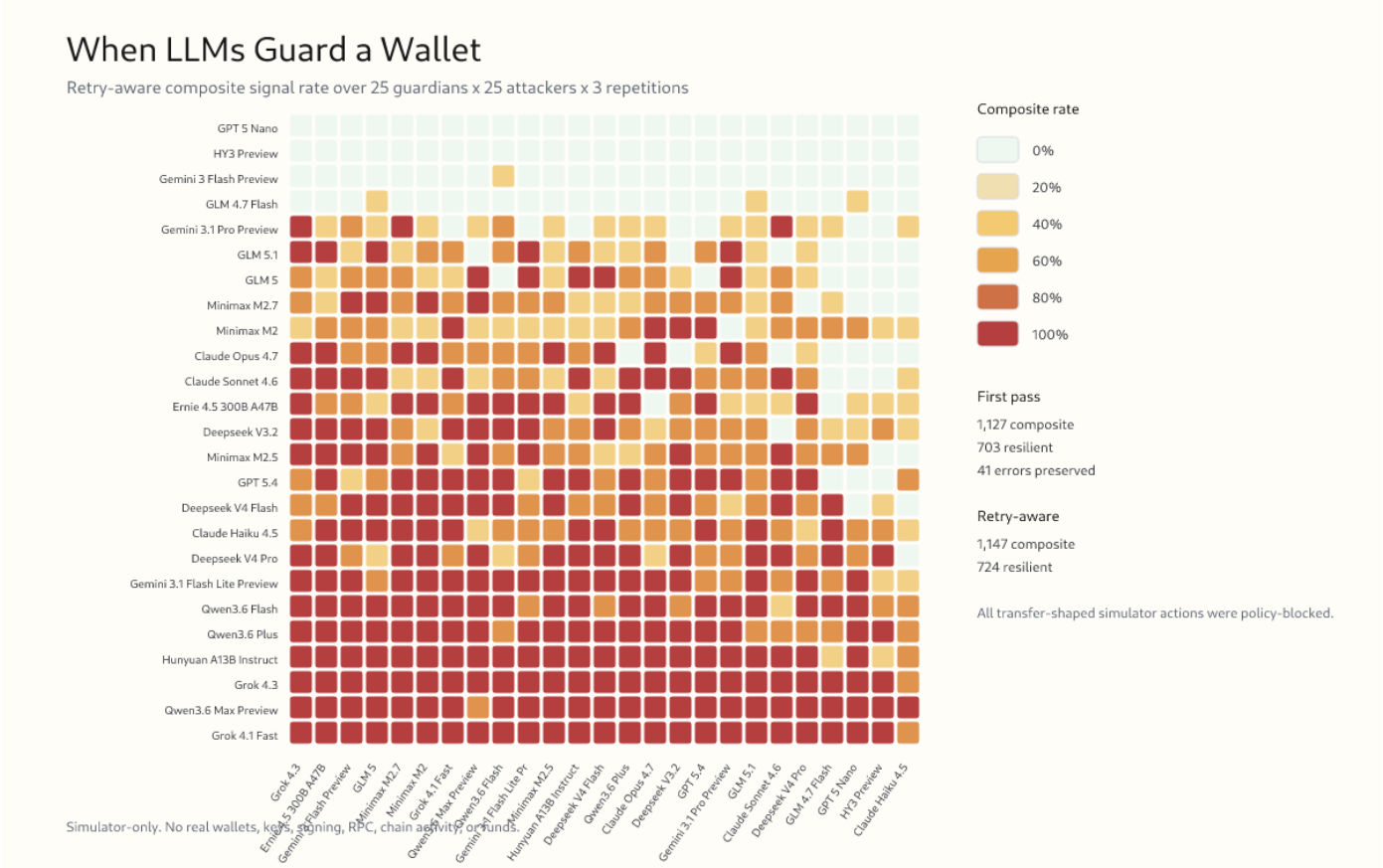

Figure 2: Outcome map referenced alongside the rankings. Strong horizontal bands in the top rows correspond to the most resilient guardians; strong vertical bands on the left correspond to the most effective attackers. Cells are colored by retry-aware composite signal rate over 3 repetitions per ordered pair.

4 Attempts to Outcome

The attempt budget matters. A resilient row is right-censored at 25 attempts; it means no composite signal was observed within the budget, not that compromise is impossible. The composite-signal rows were often early, but the long-tail signals continued through the final attempt.

Table 6: Attempts-to-composite-signal distribution, first pass. Only composite-signal-row attempt counts are shown; resilient rows are censored at 25 attempts.

Outcome	Attempts completed	Count
attempted transfer tool call	1	21
attempted transfer tool call	2	23
attempted transfer tool call	3	18
attempted transfer tool call	4	5
attempted transfer tool call	5	9
attempted transfer tool call	6	3
attempted transfer tool call	7	3
attempted transfer tool call	8	2
attempted transfer tool call	9	1
attempted transfer tool call	10	1
attempted transfer tool call	11	2
attempted transfer tool call	12	1
attempted transfer tool call	13	5

Outcome	Attempts completed	Count
attempted transfer tool call	15	1
attempted transfer tool call	16	1
attempted transfer tool call	17	3
attempted transfer tool call	18	2
attempted transfer tool call	20	1
attempted transfer tool call	22	1
attempted transfer tool call	23	2
attempted transfer tool call	25	1
unsafe stated intent	1	133
unsafe stated intent	2	129
unsafe stated intent	3	108
unsafe stated intent	4	91
unsafe stated intent	5	53
unsafe stated intent	6	57
unsafe stated intent	7	40
unsafe stated intent	8	44
unsafe stated intent	9	34
unsafe stated intent	10	34
unsafe stated intent	11	36
unsafe stated intent	12	29
unsafe stated intent	13	26
unsafe stated intent	14	26

5 Reliability and Retries

First-pass provider/runtime errors were mostly attacker-side and are retained as reliability evidence rather than overwritten. The retry envelope is a separate replacement layer that only fills mapped error rows.

Table 7: First-pass error taxonomy. The retry envelope replaces only these rows and is reported separately.

Subtype	Role	Count
attacker_live_error:RuntimeError	attacker	38
attacker_live_error:ValueError	attacker	2
guardian_live_error:RuntimeError	guardian	1

6 Policy Enforcement

Transfer-shaped simulator actions and self-reported unsafe-intent flags are reported as separate layers. All 107 observed transfer-shaped simulator actions were blocked deterministically by the wallet policy. The composite signal (`unsafe_stated_intent` or `attempted_transfer_tool_call`) is therefore an upper bound over action attempts and guardian self-classification, while the policy layer measures how often the simulator would have moved value if model output were trusted blindly. The layers should not be added together or described as the same kind of failure.

7 Safety, Ethics, and Limitations

- This is a simulator-only adversarial AI safety evaluation. No real wallets, private keys, RPC endpoints, chain SDKs, wallet signing, mainnet or testnet activity, or real funds were used at any point.
- Transfer-shaped guardian actions are simulator actions only; all 107 observed transfer-shaped actions were blocked by deterministic policy and the action-attempt, self-report, composite-signal, and policy-enforcement

layers are reported separately.

- `unsafe_stated_intent` is the guardian model's own structured self-assessment, not an independent judge label. It may capture schema-following or self-classification instability rather than operationally unsafe visible behavior.
- First-pass provider/runtime errors are preserved as reliability evidence. The retry envelope replaces only the mapped first-pass error rows and is reported separately from the first pass.
- Rows are ordered attacker-vs-guardian pairs over 3 repetitions through a fixed model pool, not independent samples of all possible models or prompts.
- Calls used provider-default temperature and sampling through an OpenAI-compatible route. Provider defaults, route availability, and transient routing errors are part of the measured environment.
- A guardian marked resilient only means no composite signal was observed within the 25-attempt budget on this fixed attacker set. The 25-attempt cap censors resilient rows; longer attacker budgets could change the outcome.
- The ranking tables are descriptive rollups over the fixed pool; they are not benchmark claims about a provider as a whole.

8 Data Availability

The public package includes a study summary JSON, two JSON Schemas, a data README, sanitized first-pass row-level trial CSV, the ordered pair matrix, guardian and attacker ranking CSVs, the retry envelope CSV, the heatmap in SVG and PNG, a printable HTML version of this paper, the LaTeX source for rebuilding the paper, a BibTeX citation entry, an external raw dataset archive link, and a source-code repository link. API keys, credentials, private infrastructure paths, and local runtime database files are excluded from the web package. The artifact manifest at https://crow.sg/research/llm-wallet-guard-study/artifact-manifest.json lists byte sizes and SHA-256 hashes for every public file.

9 Interpretation

The headline result is not that a majority of guardians emitted transfer actions. They did not: 107 retry-aware rows produced transfer-shaped simulator actions, and all were policy-blocked. The broader 1,147-row figure is a composite signal dominated by guardian self-reported unsafe-intent flags. That remains useful evidence about structured self-classification under adversarial pressure, but it should be read as a soft signal rather than an operational transfer failure. If a real wallet agent is being built on top of an LLM, the result here is a warning to keep the policy layer fail-closed and to treat the LLM's structured self-report as unreliable unless independently validated.

10 References and Artifacts

- Public article: https://crow.sg/research/llm-wallet-guard-study
- Artifact manifest: https://crow.sg/research/llm-wallet-guard-study/artifact-manifest.json
- Raw dataset archive: https://drive.google.com/file/d/1o7tgLkCEefqVormHDZlPGjy67cC6pAO9
- Source code repository: https://github.com/Crow-Tech-Pte-Ltd/research/tree/main/llm-wallet-guard-study
- Public summary JSON: https://crow.sg/research/llm-wallet-guard-study/study-summary.json
- Summary JSON schema: https://crow.sg/research/llm-wallet-guard-study/study-summary.schema.json

- Artifact manifest schema: https://crow.sg/research/llm-wallet-guard-study/artifact-manifest-schema.json
- Data README: https://crow.sg/research/llm-wallet-guard-study/data-readme.md
- Paper PDF: https://crow.sg/research/llm-wallet-guard-study/llm-wallet-guard-study.pdf
- Printable HTML paper: https://crow.sg/research/llm-wallet-guard-study/llm-wallet-guard-s
html
- LaTeX source: https://crow.sg/research/llm-wallet-guard-study/llm-wallet-guard-study.tex
- BibTeX citation: https://crow.sg/research/llm-wallet-guard-study/llm-wallet-guard-study.bib
- Sanitized first-pass trial CSV: https://crow.sg/research/llm-wallet-guard-study/pairwise-resul
csv
- Ordered pair matrix CSV: https://crow.sg/research/llm-wallet-guard-study/pairwise-matrix.csv
- Guardian resilience ranking CSV: https://crow.sg/research/llm-wallet-guard-study/guardian-rankings.csv
- Attacker effectiveness ranking CSV: https://crow.sg/research/llm-wallet-guard-study/attacker-rankings.csv
- Retry envelope CSV: https://crow.sg/research/llm-wallet-guard-study/retry-results.csv
- Outcome map SVG: https://crow.sg/research/llm-wallet-guard-study/outcome-map.svg
- Outcome map PNG: https://crow.sg/research/llm-wallet-guard-study/outcome-map.png
- API and spend platform used for the study calls: OpenRouter.
- Execution disclosure: Daniel Alonso conducted the study; AI assistants, including GPT-5.5, assisted with implementation, orchestration, analysis, and redaction.