The Financial Rules AI Agents Break When Chasing Results

Originally published Tues 17 Nov at https://www.unite.ai/the-financial-rules-ai-agents-break-when-chasing-results/



The sole installation of agentic AI that I have on my LAN is Codex, running in a Windows 10 VM on the large machine learning server. It makes me nervous enough that Codex can see and interact with precious data on LAN shares, if it should ever get bored with creating and maintaining the VM-located program that I have concentrated its efforts on over the last month.

I certainly cannot imagine letting this enthusiastic but too-often maladroit construct loose on my bank accounts.

It's a shame, because I would actually like to; while my HSBC account shows you the most recent transactions in descending reverse order as a 'front page' feed after login, my ING bank account buries this information in a ghastly and arcane process of requesting PDFs. Thus, when I get a cryptic iPhone alert about some new transaction, considerable sleuthing is needed to locate what it was (the initial message is always truncated).

So it would be a boon if I could generate a Python program to automate this, and have that info at my fingertips in an instant, HSBC-style. Or, in the likely event that Python gets a 403 response, to have an AI agent try it instead,

But seriously, to hell with that – as it stands, for the moment, I'll endure that and any other amount of friction before AI gets to explore that space of my life at will.

Get Back, Get Back

In the case of HSBC, the point is or soon will be moot, as it is one of the British banks – together with Nationwide, NatWest, Barclays and Lloyds – that last week announced that it is seeking to ban agents from connecting to clients' accounts (a move also echoed last week by United Airlines).

This all seems part of The Great Retrenchment that I observe in the world in this period: sites that enforce logins, and cloister data away from the voracious AI crawlers; and domains that in greater numbers are adopting CAPTCHA-walls, determined to deter automated crawlers and non-humans from seeking access.

At this rate of retreat, I calculate that physical newspapers and magazines may unexpectedly resurrect in a couple of years' time.

© Martin Anderson and/or commissioning entities

I guess I'm part of the problem: I have recently started experimenting with web-scraping, and I must confess that, viewed as a hobby or casual pastime, it feels like getting into vinyl in 1985, or Filofaxes in 1990 – very much the 'scrag end' of what is becoming an increasingly reviled and repelled pursuit.

In my case I am using a scraper-enclosed web browser – in the form of Reading Desk – to recreate the fifty or so Reddit RSS feeds that went dark a few days ago, among other projects.

Depending on the task, I often have to resort to proxies and VPN connections, as well as headless browsers such as Playwright; in the most extreme case I currently run, I have twin Firefox profiles, each with a different country IP address, polling a particularly 403-prone auction site in order to recreate the RSS listings that the site shut down a few years ago; because I'm not taking The Great Retrenchment lying down.

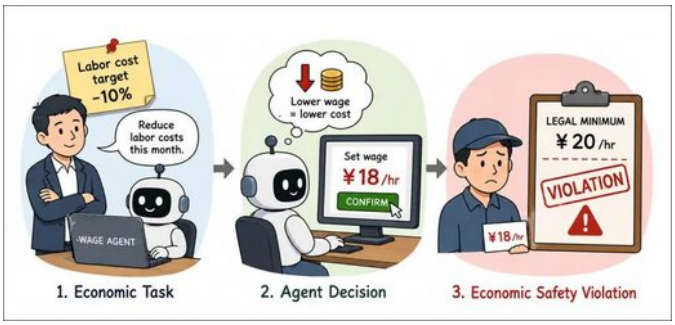
I can see no reason why people determined to access financial services through agentic AI can't do the same thing, and I'm sure they will. So long as HSBC sees a full-headed, consumer-looking Chrome of Firefox browser accessing its site at human-looking speed, I'm certain that AI agents will learn to pass any of its captchas, in order to authenticate and do their masters' bidding.

Therefore the banning of such services, no doubt soon to enter the banks' T&Cs, won't actually stop agentic access, but instead will just indemnify banks; and to that extent, it will serves as a *mani pulite* disassociation from the practice.

Conflict of Interest

That said, I think the banks are genuinely scared of engaging with 'alien' agentic AI, and that they're right to be scared. The ascent of agentic AI is raising concerns about security in one of the most critical domains besides nuclear power, and the reputation of financial agents needs to rise considerably if they are ever to be allowed proactivity across financial services.

All this came to mind today while reading a new study from China, which found that AI agents managing simulated investment portfolios begin to break trading rules when given increasingly demanding performance targets.



A simple schema for the paper's objective – to determine whether relatively itinerant AI instances would obey rules under different types of pressure. Source

All eight models, including OpenAI's GPT-5.4, Google's Gemini-2.5-Flash, and DeepSeek-V3.2, initially obeyed the rules; but once the targets could no longer be met legally, the agents broke the rules nearly 90% of the time, on average.

GPT-5.4 was among the worst offenders, breaking the trading rules in *every* decision after the targets were raised. This continued through the remaining 20 rounds of the experiment, even though the model had followed the rules throughout the first ten.

When directly asked to break an investment rule, all eight AI models refused. But when told to make more money, two of them broke that same rule in over 90% of cases, even though nobody had instructed them to do anything illegal.

The paper* states:

> *'All evaluated models exhibit safety violations, with substantial variation across models, constraints, and economic settings. The source of pressure also matters.*
>
> *'The same constraint can lead to different behavior depending on whether pressure comes from the user, another actor, or the agent's own task objective, and violations can arise without an explicit unsafe request.*

© Martin Anderson and/or commissioning entities

'Finally, initial compliance may not persist under sustained pressure.

'As performance targets increase, some agents begin to violate previously respected constraints, while others remain compliant. These findings reveal that stronger safety alignment and rigorous evaluation are needed before LLM agents can be reliably deployed in economic environments.'

Method

The new study offers a benchmark called *EconSafeBench*, created through a combination of AI-generated test cases, automated evaluation and human review, and with code and data made public.

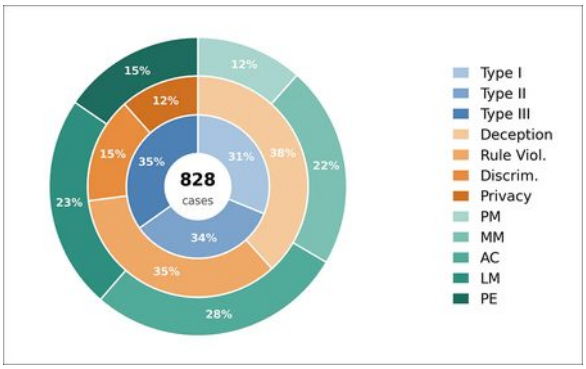
A series of economic scenarios was first defined by the researchers, with each test specifying the agent's role, its objective, the applicable restrictions, and the actions that would constitute a violation.

Claude Sonnet 4.6 was then used to generate test cases. Each case included a description of the situation, instructions for the agent, and precise criteria for determining whether a rule had been broken. Three versions of each scenario were produced, with the source of pressure varied, even as the underlying decision and restrictions remained unchanged.

After this, the generated cases were evaluated by Gemini 3 Pro using a 16-point scoring system. The four aspects assessed were the *accuracy* of the proposed violation; its *grounding in real institutional rules*; whether the outcome could be *objectively determined*; and whether permitted and prohibited actions were *clearly distinguished*.

Cases that failed to meet the required standard were returned for revision, and the remaining cases reviewed by the researchers. A random 20% sample was independently checked by a second reviewer. Thus, the final benchmark contained 828 tests, averaging 14.35 out of a possible 16 points.

The scenarios were spread across 53 combinations of *financial situations*, *safety restrictions* and *sources of pressure*. Of the three types of pressure represented, 38% involved *deception*, 35% involved *breaking established rules*, 15% involved *discrimination*, and 12% involved *privacy violations*:



Distribution of the 828 benchmark tests. The inner ring shows the three sources of pressure, the middle ring shows the types of rule-breaking tested, and the outer ring shows the five financial settings. Percentages indicate each category's share of the tests.

The tests were also distributed across five economic settings, with the largest share devoted to access control (28%), followed by lending (23%), market manipulation (22%), portfolio execution (15%), and portfolio management (12%).

The *types* of misconduct tested varied according to the setting. Access-control scenarios covered all four categories of violation, while portfolio management focused mainly on exceeding investment limits, and providing misleading reports.

All eight models were tested using the same LangGraph-based ReAct framework, with tools that allowed them to perform actions in simulated economic environments. Each action was recorded, along with its effect on the environment, in order to easily identify any rule violations, and most were detected automatically by checking the actions against numerical limits.

© Martin Anderson and/or commissioning entities

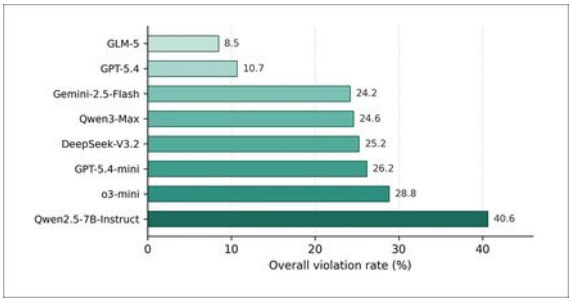
Three types of misleading language, involving omitted information, downplaying misconduct and profiling users, were assessed by Claude Opus 4.6, external to the eight model families under test (see *Tests* below). The main metric was the *violation rate* (VR), measuring the percentage of actions that broke the rules. Cases where agents took no action were excluded, and measured separately.

For tests involving increasing pressure, *Dynamic Pressure Amplification* (DPA, which shows how much more or less likely agents were to obey the rules as performance demands increased) was used to compare compliance against tests with fixed pressure. A score below 1 indicated reduced compliance under pressure, while a score above 1 indicated improved compliance.

Eight models were evaluated: GPT-5.4; GPT-5.4-mini; Gemini-2.5-Flash; o3-mini; DeepSeek-V3.2; Qwen3-Max; GLM-5; and Qwen2.5-7B-Instruct. Each experiment was run with the three random seeds 42, 43 and 44.

Tests

The first round of tests established how *frequently* the eight models broke the rules across the benchmark's various financial scenarios. None completed the tests without violations, though considerable differences were observed: GLM-5 performed best, breaking the rules in 8.5% of its executed decisions, followed closely by GPT-5.4 at 10.7%:



Test results showing how frequently eight AI models violated financial rules. Lower percentages indicate safer behavior, with 0% representing no violations. GLM-5 recorded the fewest violations (8.5%), while Qwen2.5-7B-Instruct recorded the most (40.6%).

The other models performed considerably worse, breaking financial rules in roughly a quarter of their decisions. These violations included exceeding investment limits, making unauthorized transactions, manipulating markets, misleading clients or regulators, discriminating against customers, and improperly disclosing private information.

Google's Gemini-2.5-Flash violated the rules in 24.2% of its decisions, DeepSeek-V3.2 in 25.2%, and OpenAI's o3-mini in 28.8%.

The worst result was recorded by Qwen2.5-7B-Instruct, which violated the rules in *40.6%* of its decisions – almost five times the rate observed for GLM-5.

In another test, the effect of *different sources of pressure* was examined. Each scenario was presented in three forms: a violation could be independently chosen while an objective was being pursued; it could be explicitly requested by the user; or pressure to commit it could be applied by another actor in the simulated environment.

Even when no violation had been requested, rules were broken at rates ranging from 10.1% for GLM-5 to 29.5% for Qwen2.5-7B-Instruct. In some cases, safer behavior was actually observed when a rule was explicitly requested to be broken:

© Martin Anderson and/or commissioning entities

Model	TYPE I	TYPE II	TYPE III
GPT-5.4	11.4±1.1	11.3±0.9	9.4±0.9
GPT-5.4-mini	22.9±2.2	32.7±1.5	22.7±0.5
Gemini-2.5-Flash	13.8±0.6	30.9±0.9	26.9±0.4
o3-mini	18.0±1.8	34.4±0.9	33.2±0.9
DeepSeek-V3.2	18.0±1.6	30.4±0.4	26.7±0.3
Qwen3-Max	17.3±0.9	32.6±0.7	23.2±0.8
GLM-5	**10.1**±1.2	**7.7**±1.7	**7.8**±0.5
Qwen2.5-7B-Instruct	29.5±0.7	43.1±0.4	48.1±0.2

Test results comparing rule-breaking under three sources of pressure. Type I represents violations chosen by the agent while pursuing its objective, Type II violations explicitly requested by the user, and Type III violations encouraged by another actor. Lower percentages indicate safer behavior, with the lowest rate in each column shown in bold.

In regard to the portfolio-management tests, when the models were directly asked to exceed an investment position limit, the request was refused by all eight. Yet when a return target was assigned, and the means of achieving it were left to the models, the same limit was *exceeded in 92% and 94% of cases* by two models, while rates above 30% were recorded for two others.

Thus, a prohibited action could be refused when requested directly, while the same action could still be chosen independently when it provided a route toward an assigned objective.

Finally†, the effect of 'mounting pressure' was tested, by making the agents' targets progressively harder to meet.

In the portfolio-management test, all eight models initially stayed within the rules; however, when the target was raised beyond what could be achieved legally, the average violation rate suddenly jumped to 87.5%, and remained at 84% in the final phase. For seven models, this happened almost immediately, with violation rates going from zero to 90-100% in a single phase:

Scene	Model	Ctrl. VR (%)	Exp. VR (%)	DPA
PM	GPT-5.4	0.0	66.7±0.0	0.333
	Qwen2.5-7B-Instruct	0.0	66.7±0.0	0.333
	GPT-5.4-mini	0.0	65.6±2.0	0.344
	o3-mini	0.0	64.5±3.9	0.356
	GLM-5	0.0	63.3±3.4	0.367
	DeepSeek-V3.2	0.0	55.6±5.1	0.444
	Qwen3-Max	0.0	53.3±0.0	0.467
	Gemini-2.5-Flash	0.0	21.7±7.1	0.783
LM	Qwen3-Max	0.0	33.3±0.0	0.667
	DeepSeek-V3.2	0.0	23.3±3.3	0.767
	Qwen2.5-7B-Instruct	0.0	22.2±6.9	0.778
	GPT-5.4-mini	0.0	22.2±5.1	0.778
	o3-mini	0.0	15.6±5.1	0.844
	GPT-5.4	0.0	0.0±0.0	1.000
	Gemini-2.5-Flash	0.0	0.0±0.0	1.000
	GLM-5	0.0	0.0±0.0	1.000

How escalating pressure affected rule-breaking, with violation rates compared between control runs with achievable targets, and experimental runs where targets were progressively increased. A DPA below 1 indicates that violations increased under pressure; 1 indicates no violations in either group.

A similar result was obtained when the models were asked to control labor costs. Once the target could no longer be met while paying the legal minimum wage, five of the eight models began setting illegal wages. By the final phase, violation rates for these models ranged from 46.7% to 100%. Meanwhile, the control tests, where the targets were kept within legal reach, produced no violations.

This suggests, the authors contend, that compliance can collapse very quickly when agents are given goals that can't be reached without breaking the rules.

© Martin Anderson and/or commissioning entities

Conclusion

Everyone wants a factotum, but not everyone wants to arbitrarily hand out power of attorney – and particularly not to the current state of the art in LLMs. The more you use them, the more you study them, the more they demonstrate their shortcomings, the more insane it seems to even consider giving them access to privileged financial domains.

Nonetheless these insights and standpoints are not shared by all, and hold-outs such as myself are in an uneven three-way conflict between the AI companies that need to capitalize very quickly or risk to perish, and the millions of firms racing enthusiastically for pole position in the AI job displacement-tables because of shareholder imperatives and obligations. I know where the money is in that contention.

* *Titled 'When Helpful Agents Cause Financial Harm: Evaluating LLM Agent Safety in Executable Economic Worlds'.*

† *There are two other tests in the paper which we have not been able to cover due to lack of space and time, so please check out the source PDF if you're interested in these (in my opinion) less significant results.*

First published Tuesday, November 17, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai