

Settle Only If Every Leg Ties: A Runnable Evaluation of Language Models on an Options-Component ETF Redemption

Eight frontier models held the broken redemption and settled the clean one; the arithmetic one layer down did not always follow

Dmitry Krutous, PMP · Independent · Greater Boston(<https://evals.finance>) · welt.management.solutions@gmail.com · ORCID [0009-0003-7594-4671](https://orcid.org/0009-0003-7594-4671)

Version 1 (October 2026). This is a short technical report on one evaluation, the third in a series. The TieOutBench report (SSRN, DOI [10.2139/ssrn.7243025](https://ssrn.com/abstract=7243025)) describes the suite, the grading method and the first five evaluations; the checker report (SSRN, DOI [10.2139/ssrn.7611299](https://ssrn.com/abstract=7611299)) measures a maker-checker protocol on the sixth. Everything reported here is reproducible from the accompanying repository, github.com/DimaMerc/TieOutBench, where every run's answer, raw completion, scored report and run record is committed under `outputs/eval8-live/`, together with the six attacker notes of the grader review and the note that checks the fixture's mechanics against public documents. SSRN: papers.ssrn.com/abstract=7655899, DOI [10.2139/ssrn.7655899](https://ssrn.com/abstract=7655899) (October 2026).

The option contract specifications are real: FLEX options on the iShares Russell 2000 ETF carry the OCC root 4IWM, a contract multiplier of 100, European exercise and an expiry on the last business day of the outcome period. The fund, the Authorized Participant, the clearing members, the custodian and the ETF agent are fictitious, the prices are representative, and every mechanical assertion in the case files is labelled confirmed or constructed against the public documents that were checked. Nothing here is advice regarding any security.

Abstract

When an exchange-traded fund whose portfolio is made of listed options is redeemed in kind, the option positions have to move at the clearing house from the fund's account to the redeeming dealer's account, and in 2026 that step still runs outside the clearing house's own process. In its rule filing of January 2026, NSCC describes it as "fragmented and heavily dependent on manual intervention, which increases the potential for errors." This report builds a runnable evaluation of that step and measures eight frontier models on it. The task is a settlement tie-out: given the redemption order, three portfolio composition files of which one is the basket of record, the fund's final net asset value, the clearing house's record of four position moves, the custodian's positions the next morning and the agent's own status report, tie every option leg to the basket, recompute the cash component, and decide whether to settle. The gold case has one leg moved at 12,000 contracts against 120 due, the order's underlying-share equivalent keyed as the contract count, while the agent's report says all four legs moved; the right answer is to hold and have 11,880 contracts returned first. A clean twin settles. Before any model ran, six independent attackers tried to make the grader score a wrong answer well or a right answer badly, about 1,500 graded mutations; every gate was then moved onto an evidence hook so that nothing fires on a missing field or a notation, and 267 of the attackers' checks were pinned as standing regressions. Every mechanical assertion of the fixture was checked against public primary documents and labelled. Then eight models from three vendors ran both cases once. Every model held the broken order and settled the clean one; nine of sixteen runs met every criterion; three models were perfect on both cases. The losses were arithmetic

one layer down: a small model valued the option legs per share instead of per contract and fired the gate built for that confusion on both cases; a mid-tier model sized the excess at ten times its value; another moved the written legs' positions the wrong way; a small model converted a per-share impact to a percentage fifty times too small. No model assumed what the dealer had paid for its shares, though one stated its reason for refusing in the answer schema's own words. Running real models also found six places where the grader rejected legitimate answers; all runs were re-graded with the original scores kept, and eleven scores moved without any decision or gate changing. These are recorded results from one break pattern, one run per model per case, not rates.

1. Introduction

The earlier reports in this series graded models on written analyses and on a document-store episode, and then on a maker-checker protocol over that episode [15, 16]. This report moves to a step of ETF primary-market processing that a market infrastructure describes, in a regulatory filing, as still manual.

Defined-outcome funds, the “buffer” ETFs, hold four FLEX option legs on a reference ETF for an outcome period: a deep-in-the-money purchased call that carries the exposure, a purchased put at the reference price, a written put at the buffer's lower bound and a written call at the cap. When an Authorized Participant redeems such a fund in kind, the option positions have to move at OCC, the options clearing house, from the fund's account to the dealer's. NSCC's rule filing SR-NSCC-2026-001, published in the Federal Register on January 29, 2026 and approved on March 12, 2026, states that “the entire redemption process is generally managed ex-clearing, requiring multiple manual steps to ensure completion, including the tracking, pricing, validation and ultimate execution of options positions transfers at OCC by the APs, ETF Agents and primer brokers [sic]” and that “This process, as it stands, is fragmented and heavily dependent on manual intervention, which increases the potential for errors and operational” risk [17]. The filing adds a link from NSCC to OCC so that the moves can be instructed from the portfolio composition file; DTCC's release page lists the link as in production on May 15, 2026 [19]. The filing quantifies nothing: no count of affected funds, assets or orders.

That description is, as far as the author could find, the only regulator-grade statement in 2025 and 2026 that a specific creation or redemption step is manual and error-prone. It is also a step nobody appears to have built an evaluation for. The public finance benchmarks grade analyst work [2, 3, 4, 6]; the one general professional-work suite that includes fund operations, Surge AI's DAYJOB, includes net asset value work but not primary-market processing [20]. So the question of this report is narrow: given the documents a settlement desk has when it decides whether a redemption with option legs may settle, do frontier language models make the decision a desk would make, and when they do not, where exactly do they fail?

The contributions are:

1. **A runnable, gated evaluation of the step itself.** Eight checkpoints from pinning the order to the decision, 42 point-weighted criteria, six auto-fail gates, two gold cases (a break and a clean twin), a live path for any OpenAI-compatible endpoint, and a selftest that proves the live contract against the grader. Every number in the gold answers is arithmetic on the packet.
2. **A fixture whose mechanics are checked, not assumed.** The settlement documents of such a redemption are not public, so the cases are constructed. Every mechanical assertion was checked against

public primary documents - the NSCC filing, OCC's by-laws and its 2025 rule filing, an issuer's statement of additional information, DTCC's processing pages, the SEC's Form N-CEN data - and labelled confirmed, constructed or representative in the case files. Section 3.3 reports what each label covers.

3. **A grader attacked before any model ran.** Six independent attackers, one per grading surface, tried to beat the grader in both directions. Section 4 reports what they found and what changed, including a design rule the review forced: every gate fires on evidence of the error and never on a missing field or a notation.
4. **Recorded results for eight models on both cases,** with every answer, raw completion and run record published, the grader's own calibrations disclosed with before-and-after scores, and the one hard gate that fired traced to the model's own arithmetic.

2. Related work

The grading method follows the suite's earlier reports: point-weighted criteria in the style of HealthBench [1], collapsed by auto-fail gates for the errors that make a financial work product unusable regardless of its other merits, with a calibrated-refusal probe in the style of FailSafeQA [5]. The finance benchmarks the author could find grade the analyst's work: FinanceBench's evidence-cited question answering over filings [2], FinQA's numerical reasoning [3], the Vals Finance Agent benchmark and its second version [4], and GDPval's expert-graded tasks [6]. DAYJOB [20] grades agentic professional work with binary expert rubrics and includes fund operations among its finance tasks. None of them covers ETF creation or redemption processing, portfolio composition files or options position transfers, and the author is not aware of any benchmark that does.

The operational literature here is regulatory and infrastructural rather than academic: the NSCC filing [17], OCC's rule filing of July 2025 that codified the permission to adjust options positions in support of ETF creations and redemptions [18], DTCC's ETF processing documentation [19], and the issuers' statements of additional information that define how such a fund redeems [21, 22]. Section 3.3 quotes them where the fixture rests on them.

3. The task and the fixture

3.1 The process

An in-kind redemption with option components settles bilaterally. The dealer delivers the creation units' shares through DTC to the fund's transfer agent; the fund's clearing member moves each option leg to the dealer's clearing member at OCC by a position adjustment, an off-floor transfer between clearing member accounts [11, 12]; the cash component, less the transaction fee, is wired. None of it passes through NSCC's netting system, so there is no central-counterparty guaranty on the option legs, a point the NSCC filing makes itself: the process runs "without the benefit of NSCC's CCP guaranty, which introduces operational and counterparty credit risks among market participants" [17]. The tie-out before settlement is the control: each move's series, side and quantity against the basket times the number of units; the custodian's post-transfer positions against the pre-transfer positions less the due quantity; the cash component recomputed from the final NAV and the marks against the agent's figure; the shares received. Settle only if every leg ties.

3.2 The fixture

The fund is a fictitious Rule 6c-11 buffer ETF of the suite's fictitious issuer [9], holding four FLEX legs on IWM [10] for the outcome period October 1, 2025 to September 30, 2026: a purchased call struck at 2.40, a purchased put at 240.00, a written put at 204.00 (a 15 percent buffer) and a written call at 282.00 (a 17.5 percent cap), all expiring September 30, 2026. The fund holds 24,000 contracts of each leg against 20,000,000 shares; a creation unit is 50,000 shares, so the basket carries 60 contracts per leg per unit. On June 23, 2026 the dealer redeems two units at a final NAV of 30.5000: 100,000 shares, a redemption value of 3,050,000.00, 120 contracts due per leg. At the day's marks (254.90, 3.95, 0.80 and 4.65) the four legs are worth 1,520,400.00 per unit against a unit value of 1,525,000.00, so the cash component is 4,600.00 per unit, 9,200.00 for the order, 8,700.00 after a 500.00 transaction fee.

The packet the model receives has seven documents: the order ticket; the dealer's file drop of three portfolio composition files, of which one is the basket of record, one is the next day's basket published on the trade date, and one is the issuer's November fund; the final NAV statement with the marks; OCC's record of the four position adjustments; the custodian's position report the next morning, with the share receipt and the cash component's payment status; and the ETF agent's status report, which says all four legs moved and the order is ready to settle. The break is in one line of OCC's record: the purchased put at 240.00 was moved at 12,000 contracts against 120 due, the order's underlying-share equivalent (120 contracts times 100 shares) keyed as the contract count. The fund held 24,000, so the clearing house processed it, and the receiving account took in 11,880 free long puts worth 4,692,600.00 at the mark. The custodian's report shows the fund's put position at 12,000 against 23,880 due. If the position is not recovered, the remaining shareholders carry 0.2358 per share, 0.7731 percent of NAV. The cash component in the agent's report is right; the delivery is not. The clean twin is the same order with every leg at 120 and the custodian at 23,880 per leg.

3.3 What is confirmed and what is constructed

A fixture that asserts mechanics it cannot support would mislead the practitioners it is written for, so every assertion was checked against public primary documents and labelled in the case headers. The check confirmed the skeleton. NSCC's filing confirms that these redemptions run outside NSCC with the option legs moved at OCC [17]. OCC's by-laws name the move an adjustment, and its 2025 rule filing added the purpose "for any other purpose permissible under Exchange rules" to the permitted adjustments; the filing's own text says it "would not make any substantive changes to OCC policies or practices" [18]. The by-laws test only position sufficiency: "A Clearing Member shall not effect a closing sale transaction in an account unless, at the time of such transaction, such Clearing Member has a long position in such account for at least the number of cleared contracts involved" [11], so a move of 12,000 out of a position of 24,000 passes the one check the clearing house applies. The issuer's statement of additional information defines the cash component as the difference between the unit's NAV and the value of the delivered instruments "less the applicable Redemption Transaction Fee", sets the cut-off at the close of the New York Stock Exchange's regular session, says redemption shares are delivered "through DTC to the transfer agent" and gives two business days as the outer limit for the cash, within which NSCC's regular-way T+1 for index receipts sits [21, 13]; the dealer agreements the check opened route shares through DTC and cash through Fedwire [8]; both issuers checked state that the fund holds its FLEX positions "through accounts at clearing members" [21, 22]. DTCC's pages confirm "ETF Agent" as NSCC's term, the portfolio composition file deadline of 10:00 PM Eastern on the day before the trade date, and the noon supplemental deadline [19].

The check also found what no public document supports, and those items are labelled constructed: a receiving clearing member's acceptance step and the intraday timings; the per-transfer confirmation documents (OCC publishes position information; the fixture's rows render it); the agent's status report; the option identifier and share-equivalent columns of the portfolio composition file (NSCC says such files carry "the shares and their associated quantities" for option components); the dealer clearing for itself; and the characterisation of the share leg as free of payment. Two items are labelled representative rather than typical: the SEC's Form N-CEN data show the leading issuer of such funds with units of mostly 25,000 shares and a 300 dollar fee, against the fixture's 50,000 and 500 [23].

The check also corrected the fixture's dating. DTCC lists NSCC's link to OCC for these moves, Options ETF Clearing Connect, as in production on May 15, 2026, before the fixture's trade date [19]. The case files now state that the fixture's dealer and agent have not onboarded to the new service and settle under their pre-existing bilateral arrangement, which the NSCC filing describes as participant-elected, and they note that the planted keying error is a risk of the manual flow: in the messaged flow, the filing says, "instructions would be created by using the daily portfolio composition files" [17].

3.4 Checkpoints, gates and the refusal probe

The task is decomposed into eight checkpoints. Planning pins the order: identity, the trade and settlement dates, the basket of record among the three files, the fund's own four series, the order terms and the settlement route. Two extraction checkpoints read the basket and the agent's report as claimed, then the independent evidence: OCC's record, the custodian's positions, the NAV statement, the share receipt. Three calculation checkpoints do the work: the per-leg tie with the custodian cross-check; the cash component recomputed from the legs (contracts times the 100 multiplier times the mark, long positive and written negative) and tested against the agent's figure; the break quantified in contracts, value, direction and impact per remaining share. The decision checkpoint grades the settle-or-hold call, its localisation to the leg, the remedy (a corrective transfer of 11,880 contracts back to the fund before any cash is released or any share cancelled) and the parties escalated to. The refusal checkpoint asks for the dealer's profit or loss on the redemption, which needs its acquisition cost for the shares, a figure no document in the packet carries; the answerable twin is the net cash due to the dealer.

Six gates collapse the score. The signature gate fires when a model settles a redemption whose legs do not all tie, in any wording, and raises a headline flag. Two hard gates fire when the model ties the wrong basket: the next day's file (a different cash estimate) or the sibling fund's (different strikes and expiry). A third hard gate fires when the model's own arithmetic confuses contracts with shares, a figure sitting on the hundredfold axis of the right value. A scoped gate fires when the direction of the break is inverted, the over-delivery reported as a shortfall owed to the dealer. The fabrication gate voids an invented document row or an invented cost, price or profit. A penalty sized to zero the decision checkpoint on its own scores the over-cautious mirror: holding the clean twin.

3.5 Scoring

Scoring is the suite's: 42 criteria with point weights, aggregated by checkpoint and collapsed by the gates, a reported gated score and an "AllPass" flag when every criterion is met, and the refusal checkpoint scored by the F-beta of refusal quality and twin accuracy. Every criterion of this evaluation is deterministic; no language-

model judge is used [7]. The live contract is a JSON schema whose key paths mirror the gold, and the selftest proves that a schema-perfect answer grades 1.000 with every criterion met on both cases.

4. Testing the test before the models

The grader is a program, and a program that scores language can be beaten by language. Before any model ran, six independent attackers, each a language model working from the same brief, took one surface each: the settle-or-hold classifier, the refusal grader, the numbers and the multiplier gate, extraction and the tie ledger, the planning pins and their gates, and a sweep of legitimate answer formats a careful analyst might produce. Each was told that a finding without a reproduced run was not a finding. They produced about 1,500 graded mutations and six notes, all committed.

They found the grader wrong in both directions. The hard planning gates fired on the wrong things: a perfect answer that wrote a date as “23-Jun-2026”, left the basket date blank or listed the four option series in another notation lost two thirds of its score to a gate meant for using the wrong basket, while the basket of record itself was never verified, so a model could name the November fund’s file, or an invented one, and keep a perfect score. The decision classifier read words, not clauses: “No exceptions. Settle.” counted as a hold because a negation stood three words before the verb; “Don’t settle” with a curly apostrophe counted as a settle; a correct hold that spelled out “do not release the cash, do not cancel the shares, do not mark the order settled” counted as a contradiction; a release instruction placed next to the word “escalate” passed; and the two yes-or-no switches for releasing the cash and cancelling the shares were read by nothing. The remedy was a bag of words: a remedy that moved the 11,880 contracts the wrong way, to the dealer, satisfied it. Ledger statuses were sorted by the first family word found, so “quantity matches” read as a quantity break. Row labels and the fabrication test were whitelists, so a ledger row labelled “share delivery” fired the fabrication gate while a second, invented confirmation for the broken leg did not. The multiplier gate could be dodged by withholding the figures it read and fired on a single mark slip that landed near one hundredth of the right value. The refusal grader keyed on its own designed phrasings: “the acquisition cost of the 100,000 shares is not in the packet” was gated as a fabrication because the share count read as a price, while “since the cost is not in the packet, I assume the dealer bought at the prior NAV” passed because a negation stood earlier in the sentence.

What changed is one design rule and its consequences. Every gate now fires on evidence of the error and never on a missing field or a notation. The date, series and multiplier gates moved from the positive pins to penalty hooks: a missing or oddly written pin costs its points and nothing more, while a parseable date that is a different day, a file identifier that resolves to another basket, a basket actually read from a distractor file, or two figures on the hundredfold axis fire. The decision classifier works clause by clause, with negation and conditions bound inside their clause, a settle conditioned on the corrective action read as a hold and one conditioned on a time read as a settle, and the switches read. The remedy is graded on its own text for the return of the excess, its size, its direction back to the fund, the release sequenced after it, and the parties. Statuses separate the verdict from the kind of mismatch. Rows key by leg number, by the option series in any notation, or by a side-type-strike description. Fabrication is tested against the packet. Numbers parse units, locales and accounting signs, and a hedge is no value. The refusal grader scans every string of the answer for an assumed cost, a stated price or a profit, binds negation to the same clause in both directions, and treats share counts and the packet’s own derived totals as known.

Where the author’s judgment differed from an attacker’s expectation, the judgment is pinned and recorded: an unrecognised decision word earns nothing and fires no gate; a settle conditioned on a time is a settle; three of four series pinned is incomplete, not wrong; a hedge is an unmet pin, not a gate. After the repair the selftest passed for all eight suites of the repository, the oracle scored 1.000 on both cases, and the 22 prior cases of the other evaluations graded byte for byte as before. No live model had been run when the grader changed, so no published number moved. 267 of the attackers’ checks run with every selftest.

5. Results

Eight models from three vendors ran both cases once each on October 21, 2026, through the live path, with the same packet, the same schema and the same grader. Table 1 gives the gated scores; “AP” marks a run that met every criterion.

Model	Tier	Break case	Clean case
Claude Opus 4.8	flagship	1.000 AP	1.000 AP
Claude Sonnet 4.6	mid	0.868	1.000 AP
Claude Haiku 4.5	small	0.543 · multiplier gate	0.549 · multiplier gate
GPT-5.6-sol	flagship	1.000 AP	1.000 AP
GPT-5.5	flagship (previous)	1.000 AP	1.000 AP
GPT-5.4	mid	0.988	1.000 AP
GPT-5.4-mini	small	0.968	1.000 AP
Gemini 3.6 Flash	small, fast	0.956	0.989

Table 1. Gated scores, one run per model per case, after the grader calibrations of Section 5.5. The two Gemini runs are the re-runs at a 16,000-token budget; the first attempts, cut off at 8,000 tokens, are kept in the repository.

5.1 Every model held the broken order and settled the clean one

The headline is uniform. All eight models returned a hold on the break case, named the purchased put as the leg, the 12,000 contracts against 120 due and the 11,880 excess, described the corrective transfer back to the fund’s account before any release, and escalated to both clearing members; all eight settled the clean twin. Nine of sixteen runs met every criterion, and three models, Claude Opus 4.8, GPT-5.6-sol and GPT-5.5, were perfect on both cases. The signature gate never fired: no model settled a redemption whose legs did not tie, as no model settled the broken basket, affirmed the mismatched confirmation or released the wrong NAV in the suite’s earlier evaluations [15]. No model held the clean order. On the refusal probe, no model assumed what the dealer had paid for its shares or invented a closing price; every model labelled the question not answerable from the packet and produced the net cash due to the dealer as the answerable twin.

5.2 The hundredfold slip, and the gate built for it

The one hard gate that fired is the one this evaluation was built around, and it fired in a model’s own arithmetic rather than in a label. Claude Haiku 4.5 valued the four option legs per share of the underlying instead of per

contract: 15,294.00, 237.00, minus 48.00 and minus 279.00 per unit, each one hundredth of the right figure, and an option value per unit of 15,204.00 against 1,520,400.00. On the break case it reported the creation unit's value as 30.50, the NAV per share, rather than 1,525,000.00. It then copied the agent's cash component of 4,600.00 per unit as its own recomputation, so its cash figures tied while its legs could not have produced them. The multiplier gate fired on both cases (0.543 and 0.549), and the gate's evidence is visible in the answer: four leg values and the option value on the hundredfold axis. The model still held the broken order and settled the clean one, and it sized the excess at 47,052,000.00.

5.3 Magnitude one layer down

The remaining losses are errors of size with the decisions right. Claude Sonnet 4.6 valued the excess at 46,926,000.00, ten times the right 4,692,600.00, and carried the error into the impact per remaining share (2.36 against 0.24) and into the hold record, while its per-unit leg values were right; the right call with a tenfold break (0.868). GPT-5.4 moved the fund's positions the wrong way on the two written legs, expecting minus 24,120 where the custodian showed minus 23,880, and flagged a 240-contract shortfall that does not exist on two legs that tie; everything else was right (0.988). GPT-5.4-mini sized the excess and the impact per remaining share correctly, 0.2358, and then reported it as 0.0154 percent of NAV instead of 0.7731 (0.968). None of these fired a gate: a tenfold error is a mis-scaling of the kind the suite's scale gates fire on in other evaluations, but this evaluation's multiplier gate is deliberately built for the hundredfold axis of contracts and shares, and the review had shown that a wider gate catches coincidences.

5.4 The refusal, and an echo

Gemini 3.6 Flash got every number and both decisions right and lost points twice. On the break case its stated reason for refusing the profit question was the answer schema's own instruction text, returned word for word: "for the AP's profit or loss: state that the AP's acquisition cost of the shares it tenders is not in the packet and name that missing input; do NOT assume a cost (not 'bought at NAV', not 'at the close') and do NOT invent a price or a profit". The refusal grader reads that as a refusal without a stated reason and gives it half credit (0.956); the same model did the same thing on the suite's seventh evaluation. On the clean case it left the quantities of the share row of the tie ledger blank (0.989). Its first attempts at both cases were cut off by the live path's default 8,000-token budget, most of which the endpoint reported as hidden reasoning; the visible answers ended in the extraction section and scored 0.264 and 0.361 as parsed. They are kept in the repository under each run's `prior/` folder, and the scores in Table 1 come from the re-runs at 16,000 tokens.

5.5 Corrections to the grading, not to the models

Running real models found six places where the grader rejected a legitimate answer, none of which the attackers had reached. On the clean twin the models left the position fields for the offending leg empty, because there is no offending leg; the grader demanded numbers. The schema field for the option value the dealer received was never defined, and three models read it three ways: the total received, the excess only, the whole offending leg; it is no longer graded, and the schema now defines it. Two models stated the settlement actions in the structured switches and left the remedy text empty on the clean case; the switches now count. One model's refusal opened with the probe's own words and was treated as an echo of the instructions; the echo guard now looks for the instruction's imperatives only. Two models filed an "exception" on the cash whose text says the figures agree; such a row is now informational when the model's own cash ties. One model cited the agent's totals line, a line of the packet the gold alternates did not list; it is listed now. All fourteen runs that

existed at the time were re-graded offline with the original scores kept in each run record; the two Gemini runs came later and were graded by the calibrated grader from the start. Eleven scores moved, none of them a decision or a gate (Table 2).

Model	Break before → after	Clean before → after
Claude Opus 4.8	0.935 → 1.000	0.882 → 1.000
Claude Sonnet 4.6	0.868 → 0.868	0.903 → 1.000
Claude Haiku 4.5	0.528 → 0.543	0.534 → 0.549
GPT-5.6-sol	1.000 → 1.000	0.935 → 1.000
GPT-5.5	0.956 → 1.000	0.855 → 1.000
GPT-5.4	0.923 → 0.988	0.903 → 1.000
GPT-5.4-mini	0.903 → 0.968	0.882 → 1.000

Table 2. Gated scores before and after the six grader calibrations found by running real models. Both scores are kept in each run’s `run.json`.

6. What the results say

Three observations, stated as narrowly as the evidence allows.

The control transfers. Given the documents a desk has, every one of eight models tied the four legs to the basket, found the leg moved at one hundred times its due quantity, refused to settle, named the return of the excess as the first step and escalated to the right parties, and every one settled the clean twin. In the suite’s four reconciliation-type evaluations - creation and redemption, confirmation matching, NAV oversight and now this one - the signature gate (settle, affirm, release, settle) has never fired in a live run. Where the frontier models differ from a desk is not in the decision.

The arithmetic one layer down does not follow the tier cleanly. The hundredfold confusion of contracts and shares appeared in one small model and nowhere else, and the gate built for it caught it in the model’s own figures. But a tenfold error appeared in a mid-tier model that had the per-unit values right, the sign of a written leg’s position movement defeated another mid-tier model, and a percent conversion defeated a small one. Each of these would have been carried into a corrective instruction, a loss estimate or an escalation memo. The evaluation’s value is that it says which figure, in which checkpoint.

A fixture’s mechanics are claims, and claims can be checked. The author could not validate the operational details of this process from experience, and said so. Checking them against public documents confirmed the skeleton, corrected the dating, and separated what the documents support from what the fixture constructs. The labels in the case files are the result. The evaluation’s scores do not depend on those details - the gold answer is arithmetic on the packet - but a practitioner’s trust in the fixture does, and that trust is the point of building it.

7. Limitations

One break pattern, one run per model per case, eight models. Sixteen runs do not give rates, and a model that lost points on one arithmetic step might not lose them on the next run. The break is a quantity error on one leg; a leg left pending at the cut-off, a stale mark in the agent's cash component, the sibling fund's series on one leg, and the messaged flow with a rejected instruction are planned cases, not built ones. The fixture is constructed: the documents that a settlement desk actually sees were not available, so the packet renders what public documents say such documents contain, and several of its details are labelled as unsupported by any public source. The grader is deterministic and was attacked by six language models and repaired by the author; the six calibrations found by running real models show that the attackers did not reach every legitimate format, and a wider set of live runs would likely find more. The attackers, the document check and much of the harness were produced with a language-model assistant whose vendor makes three of the eight models under test; the structural safeguards are stated in the declaration below.

8. Reproducibility

Everything is in the repository [14]. The workflow decomposition is `workflow/options-redemption-analysis.md`; the rubric is `rubric/criteria-options-redemption.yaml` (version 1.1.0, 42 criteria, six gates); the two cases are `cases/grbo-redeem-2026.case.yaml` and `cases/grbo-redeem-2026-clean.case.yaml`, with every figure's arithmetic in their verification blocks and the mechanics labels in their headers; the grader is `harness/suites/options_redemption.py`, the live path `harness/live_options_redemption.py`, the pinned regressions `harness/gaming_review_eval8.py`. `python -m harness selftest` runs every invariant of every evaluation in the suite. The sixteen runs are under `outputs/eval8-live/<model>/<case>/` as `answer.json`, `raw.txt`, `report.txt` and `run.json`, the last carrying the endpoint's usage and finish reason, the hashes of the grader, suite, live module, case and rubric files, the git revision, and the re-grade record with both scores. The six attacker notes and their scripts are under `outputs/eval8-live/gaming-review/`, the mechanics check is `outputs/eval8-live/MECHANICS-NOTE.md`, and the taxonomy with the grader-hardening record and the live findings is `outputs/eval8-live/TAXONOMY.md`.

Author note

The author led the delivery of an institutional ETF servicing platform - creation, redemption, corporate actions - and worked on derivatives processing across options, futures, forwards and swaps, and has built AI-assisted finance tools independently since 2022. He did not run a settlement desk, and this report does not claim that experience; where the fixture's mechanics mattered, they were checked against public documents rather than asserted from memory, and the result of the check is labelled in the case files.

Declaration on the use of generative AI

A generative AI assistant (Anthropic's Claude) was used throughout this work: to draft and refine the harness code, the case files and this text; to act as the six attackers of the grader review, each a separate session working from a written brief; and to act as the research assistant that checked the fixture's mechanics against public documents and wrote the mechanics note. All factual claims were verified against committed artifacts (the

public documents cited, the case files, the sixteen runs' answers and run records, the scored reports), and the author reviewed and edited all content and takes full responsibility for it. Because generative AI is also this report's subject, and because three of the eight models under test are made by the vendor of the assistant used, the safeguard against circularity is structural: gold answers are closed-form arithmetic on the packet committed in the repository, grading is deterministic, every answer and run record is published verbatim, and every reported number is reproducible by any reader from the published artifacts.

References

1. Arora, R. K., et al. (OpenAI). *HealthBench: Evaluating Large Language Models Towards Improved Human Health*. arXiv:2505.08775, 2025.
2. Islam, P., et al. *FinanceBench: A New Benchmark for Financial Question Answering*. arXiv:2311.11944, 2023.
3. Chen, Z., et al. *FinQA: A Dataset of Numerical Reasoning over Financial Data*. arXiv:2109.00122, 2021.
4. Vals AI. *Finance Agent Benchmark*. arXiv:2508.00828, 2025; *Finance Agent v2*, vals.ai/benchmarks/fabv2, 2026.
5. Kamble, K., Russak, M., et al. *Expect the Unexpected: FailSafe Long Context QA for Finance*. arXiv:2502.06329, 2025.
6. Patwardhan, T., et al. (OpenAI). *GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks*. arXiv:2510.04374, 2025.
7. Zheng, L., et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. arXiv:2306.05685, 2023.
8. Vanguard. *Form of Authorized Participant Agreement for Vanguard ETF Shares*. EDGAR accession 0001683863-23-005607.
9. Securities and Exchange Commission. *Rule 6c-11 under the Investment Company Act of 1940* (17 CFR 270.6c-11).
10. Cboe Exchange, Inc. *Equity FLEX Options: Contract Specifications*. cboe.com.
11. Options Clearing Corporation. *By-Laws and Rules*, as last modified August 19, 2025 (Article VI, Section 1, Interpretation .01, and Section 15).
12. Cboe Exchange, Inc. *Rule 6.7 Off-Floor Position Transfer Form*, cdn.cboe.com; SEC Release 34-87340 (October 17, 2019), 84 FR 56877, on Cboe Rule 6.9 transfers in connection with ETF creations and redemptions.
13. Securities and Exchange Commission. *Release 34-100048* (May 2, 2024), NSCC: T+1 settlement for index receipts.
14. Krutous, D. *TieOutBench* repository, github.com/DimaMerc/TieOutBench; the runs of this report under `outputs/eval8-live/`, their taxonomy in `outputs/eval8-live/TAXONOMY.md`.
15. Krutous, D. *TieOutBench: A Runnable, Rubric-Gated Evaluation Suite for Financial Analysis and Post-Trade Operations*. SSRN, 2026. DOI 10.2139/ssrn.7243025. Data and code snapshot: Mendeley Data, DOI 10.17632/m5wfrnhdsp.
16. Krutous, D. *A Second AI as Checker: Measuring a Maker-Checker Protocol for LLM Agents on Corporate-Actions Processing*. SSRN, 2026. DOI 10.2139/ssrn.7611299.

17. National Securities Clearing Corporation. *Proposed Rule Change to Support ETF Creations and Redemptions with Options Components* (SR-NSCC-2026-001). SEC Release 34-104687 (January 26, 2026), 91 FR 3938 (January 29, 2026); approval order Release 34-104981 (March 12, 2026).
18. Options Clearing Corporation. *Proposed Rule Change* SR-OCC-2025-010. SEC Release 34-103521 (July 22, 2025), with Exhibit 5.
19. DTCC Learning Center. *ETF Processing: Timeline and Submission Processing; ETF Release* (page dated September 21, 2026, listing Options ETF Clearing Connect in production May 15, 2026). dtcclearing.com.
20. Surge AI. *DAYJOB*. arXiv:2610.01306, October 1, 2026.
21. Innovator ETFs Trust. *Statement of Additional Information*, Form 485BPOS, accession 0001213900-24-079100 (sections on the placement of creation or redemption orders, the delivery of redemption proceeds, and clearing member default risk); Form N-CEN, accession 0000894189-26-000385.
22. First Trust Exchange-Traded Fund VIII. *Statement of Additional Information*, Form 485BPOS, accession 0001445546-26-007606; Form N-CEN, accession 0000940400-26-028992.
23. Securities and Exchange Commission. *Form N-CEN Data Sets*, 2025 Q4 to 2026 Q3 (creation-unit sizes, cash versus in-kind redemption percentages, transaction fees). sec.gov/data-research.