Independent AI Assessment for Capital Markets Operations

"Your vendor says it's 99% accurate." On which cases? How many runs? Scored by whom?

AI Alpha Labs independently evaluates AI systems on the operations workflows your desk actually runs — trade confirmation, margin disputes, settlement fails, P&L reconciliation, collateral eligibility. We measure the system against ground truth established before evaluation begins, score it with a deterministic engine, and publish the methodology so anyone can reproduce the result.

We don't build the models we test. We don't sell the AI we evaluate. The only product is the evidence.

93–100%	12–58%	1,500	10
DETECTION ACCURACY	VALUE ACCURACY	VALIDATED CASES	FRONTIER MODELS

Across six published benchmarks. Models reliably know something is wrong; they cannot reliably tell you by how much. A single "the model is validated" sign-off hides exactly that risk.

Why now

SR 26-2 (Fed / OCC / FDIC, April 2026) replaced SR 11-7 and left standalone generative and agentic AI **outside** the formal model-risk framework — putting validation design back on the institution, with no supervisory template for what sufficient evidence looks like.

The SEC's 2026 examination priorities name AI governance directly. Examiners are reviewing the accuracy of firms' AI-capability representations, not merely whether a policy exists.

"The vendor told us" is no longer an answer. Independent, reproducible evidence is.

What we measure

Every engagement separates two questions most evaluations conflate:

— **Does the system know a problem exists?** Detection, classification, attribution, false-positive rate.

— **Can it compute the right answer?** Value accuracy, exposure calculation, escalation judgment.

These fail at completely different rates. Measuring them together hides the failure that matters.

What you receive

- **Per-model scorecard** — every metric with Wilson 95% confidence intervals
- **Case-level results** — full scored output, every case, every run
- **Published methodology** — prompts, rubrics, scoring logic, versioned and open
- **Reproducibility statement** — what a third party needs to independently re-run the evaluation
- **Deployment recommendation** — which tasks the system is fit for, which require deterministic controls, where a human must sign
- **Regulator-ready summary** — written for model risk, audit committee, or examiner review

> **Worked example — this is what your deliverable looks like.** AAL-D-006, Collateral Eligibility & Substitution: 250 cases, seven frontier models, 5,250 scored observations, published in full.
> `aialphalabs.ai/research/AAL-D-006`

How it works

PHASE	DURATION	OUTPUT
1 · Scoping	Week 1	Workflow selection, success criteria, metrics agreed
2 · Dataset construction	Weeks 2–3	Cases built with ground truth by construction, independently recomputed
3 · Evaluation	Weeks 4–5	Multi-run execution, deterministic scoring, integrity gate
4 · Reporting	Weeks 6–7	Scorecards, findings, deployment recommendation, review session

6–8 weeks end to end. No access to production systems required — evaluations run on constructed cases that mirror your workflows, so nothing sensitive leaves your environment.

Engagement terms

Fixed scope. Fixed fee. Paid regardless of outcome.

ENGAGEMENT	SCOPE	FEE
Single workflow	One workflow, up to 3 models	**$35,000**
Multi-model comparison	One workflow, up to 8 models	**$55,000**
Commissioned benchmark	New dataset built to your workflow; generalized version published to the public series	**$60,000**

The fee does not change based on what we find. We are paid for the measurement, not the verdict — and unfavorable findings are reported as clearly as favorable ones. That is the entire point.

Why AI Alpha Labs

- **Independent by construction.** No model vendor relationships, no AI products sold, and no consulting to help anyone pass a test we administer.
- **The only public corpus of its kind.** Six datasets, 1,500 validated cases, ten frontier models, deterministic scoring — everything published, and the reference set your results are calibrated against.
- **Operator-built.** Fifteen years in capital markets operations across Millennium, Schonfeld, Fannie Mae and Vanguard. The workflows are modeled by someone who ran them.

Joseph Yuschak · Founder, AI Alpha Labs
aialphalabs.ai · yuschajp@gmail.com

Start here: the 10-question Reproducibility Scorecard — send it to any AI vendor before you buy.