KITEEDGE WHITE PAPER

Auditable Judgement

Why buy-side AI programmes stall, and how to fix the missing layer

November 2026

kiteedge.co.uk

SUMMARY

The layer AI did not make cheap

Institutions get results from AI only when three capabilities work together: **Organisation**, **Judgement** and **Action**. The relationship is multiplicative, so a zero anywhere means a zero overall.

AI has made Organisation and Action cheap. Search, summarisation, drafting and workflow automation are now commodities.

Judgement has not got cheaper. Deciding which evidence counts, resolving conflicting sources, knowing what was not covered, and owning the conclusion are still where errors cost the most. In most AI programmes, judgement is also where accountability is least visible.

This paper argues that the next gains in buy-side AI come from making judgement **auditable**. That means bounded by licensed evidence, explicit about confidence and gaps, and signed off by a named person.

It also includes a 12-point self-check, so you can see which layer limits your own results.

THE PROBLEM

AI got cheap in the wrong places

Most buy-side AI programmes have followed the same arc. First came a chat assistant. Then came access to more data, then automation of a few workflows: summarising research, drafting notes, screening documents. Each step made something faster, but few made investment decisions measurably better.

The reason is structural. AI is very good at two of the three things an institution needs.

- **Finding and arranging information.** Retrieval, search and summarisation have become cheap and widely available.
- **Producing output.** Drafts, briefs, tables and alerts now take seconds.

AI is much weaker at the third: **deciding what the evidence supports, and being able to show why.** A model will answer every question with the same fluency, whether the evidence is strong, thin or missing. It rarely tells you what it did not look at.

The gap is visible in independent testing. A June 2026 benchmark by a financial data provider asked 13 models 150 analyst-style questions. With web search alone, the models captured on average 13% of the key facts an expert answer required. Connected to professional data, they captured 32%.[1] Better inputs more than doubled performance. Even so, most of what an analyst needed was still missing, and the answers rarely made that obvious.

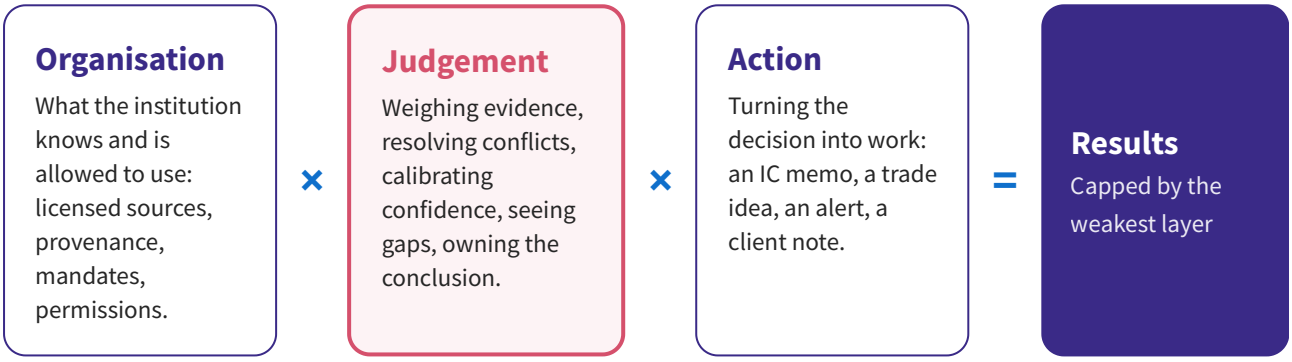
> Fluent output with invisible gaps is the real risk. A PM can't calibrate conviction on an answer that looks complete but isn't.

Equally, a compliance team can't defend a conclusion whose sources, permissions and sign-off weren't recorded. The result is a familiar pattern. Pilots impress in demos, usage plateaus, and the most important decisions still route around the AI.

THE FRAMEWORK

Results = Organisation × Judgement × Action

We use one simple model to diagnose where AI programmes stall.



The multiplication is deliberate. Strength in one layer does not make up for weakness in another; it amplifies it.

Layer that is weak	What the institution experiences
Organisation	Confident conclusions built on inputs nobody can vouch for. Licensing exposure.
Judgement	Fast output that nobody can defend. In automated workflows, mistakes made at scale.
Action	Good analysis that never reaches a decision. Duplicated work across desks.

The model is also a loop, not a pipeline. What was decided, on what evidence, and how it turned out should flow back into Organisation, so that the next decision starts from a better base.

WORKING HYPOTHESIS

Most buy-side firms have invested heavily in Organisation (data platforms, research subscriptions) and increasingly in Action (copilots, internal assistants). Judgement is the layer left ungoverned. The self-check in this paper is designed to test that hypothesis, firm by firm.

IN PRACTICE

What auditable judgement looks like

Auditable judgement does not mean replacing the analyst or the PM. It means the reasoning AI contributes is bounded, inspectable and owned. In practice that comes down to five properties. Each one is something you can test today, in your own tools.

Property	What it means	Test it with this question
1. Bounded evidence	The AI reasons only over sources you are licensed and permitted to use, and that boundary is enforced before an answer is generated, not checked afterwards.	"If I ask about a provider we don't subscribe to, what happens?"
2. Traceable claims	Every claim links to the passage that supports it, with its provenance: who published it, when, and under what licence. A citation to a document is not enough.	"Show me the paragraph behind this sentence, and who we licensed it from."
3. Visible coverage	The answer shows what was searched and what was not, so a thin evidence base looks thin. Where evidence is insufficient, the system says so instead of filling the gap.	"What did this analysis not cover?"
4. Surfaced conflict	When sources disagree, the disagreement is shown, not averaged away, and confidence reflects it.	"Where do our sources disagree on this, and by how much?"
5. Named accountability	A person commits to the conclusion before it is acted on, and the record of evidence, scope and sign-off can be replayed later.	"Who signed off on this, on what evidence, and can we re-run it?"

The first two properties belong to Organisation: they make inputs trustworthy. The last three are Judgement itself. In our experience, most AI tools address the first two at best. Very few make coverage, conflict and accountability visible at the point of decision. That is where the conviction of a portfolio manager, and the confidence of a compliance team, are actually won or lost.

> In an AI-assisted process, the PM is no longer a manager of information volume. The PM becomes a manager of calibrated certainty.

That means deciding how much weight a conclusion can bear, given what is known and what is not. Auditable judgement is the infrastructure for that role.

BUILD OR BUY

The internal-build question

Many asset managers now have an internal AI team building assistants on top of foundation models. That is a sensible investment, and it changes what a firm should buy, not whether it should.

Internal builds tend to be strongest in **Action**. They plug models into existing workflows, systems and user interfaces, and the people building them understand the firm's processes better than any vendor. They tend to be weakest in two places:

- **Licensed Organisation.** Third-party research comes with entitlements, redistribution limits and audit requirements. Building and maintaining a clean, licensed corpus, with permissions enforced at the passage level, is specialist infrastructure that sits outside most internal teams' remit.
- **Governed Judgement.** Coverage, conflict detection, confidence and sign-off records are hard to retrofit onto a general-purpose assistant. They are usually the first items cut from an internal roadmap, because the demo works without them.

The practical answer is to treat these as layers, not alternatives. The internal team owns Action and the user experience. A governed layer supplies licensed evidence and auditable judgement, through APIs and model-agnostic connections the internal build can call. The firm keeps control of its workflows and gains the parts that are hardest to build and most expensive to get wrong.

A TEST FOR ANY BUILD-OR-BUY DISCUSSION

Take the five properties in the previous section. For each one, ask whether your current plan delivers it, who maintains it, and how you will evidence it to your compliance team.

SELF-CHECK

Which layer limits your results?

Score each statement for one workflow you care about, for example producing an IC memo, an earnings review or a research brief. Be strict: score 3 only if you could show the evidence to a sceptical colleague.

Scale: 0 = No · 1 = Partly · 2 = Mostly · 3 = Yes, and we could show you

#	Layer	Statement	Score
O1	Organisation	We can trace any figure in last week's IC pack back to its original source	
O2	Organisation	We know which research we are licensed to use in AI tools, and the tools enforce it	
O3	Organisation	Our research, notes and models sit in one governed place, not across inboxes and drives	
O4	Organisation	Fund mandates and constraints are written down where our AI tools can apply them	
J1	Judgement	When two sources disagree, we have a set way to resolve it, and we record the decision	
J2	Judgement	Our PMs can see what the research did not cover, not just what it found	
J3	Judgement	AI-generated analysis shows its evidence and confidence before anyone acts on it	
J4	Judgement	A named person signs off before AI output reaches an investment decision or a client	
A1	Action	A research question becomes a usable brief in hours, not days	
A2	Action	Analysts rarely redo work another desk has already done	
A3	Action	AI output arrives in the formats we actually use: IC memos, alerts, client notes	
A4	Action	We can re-run a past analysis quickly when new information arrives	

Reading your result

1. Add up each layer: O1 to O4, J1 to J4 and A1 to A4. Each layer scores out of 12.
2. Your **weakest layer caps your results**, however strong the other two are. Start there.
3. A gap of 4 or more points between your strongest and weakest layer usually means investment has gone into the wrong place. Typically that is more output, or more data, on top of ungoverned judgement.

Weakest layer	What it usually means	Where to start
Organisation	Your AI reasons over inputs you can't fully vouch for	Map sources, licences and mandates for one workflow
Judgement	Your AI moves fast, but nobody can show why a conclusion was trusted	Design the evidence, coverage and sign-off gate for one decision type
Action	Your inputs and judgement are sound, but the insight isn't reaching decisions	Turn one recurring question into a repeatable brief

NEXT STEPS

Where to start

Start with one workflow, measure it, and fix the weakest layer first, rather than starting with a platform decision. We recommend three steps.

1. **OJA Diagnostic, 1 to 2 weeks.** Take one workflow and score Organisation, Judgement and Action with the people who run it. Then run your own questions through a governed environment, so the gaps show up on your real work, not on a vendor demo. You receive a gap map and a prioritised plan, whether or not that plan involves our platform.
2. **Scoped proof of concept.** Aim it at the weakest layer, with success measures agreed before it starts. For example: the share of claims traceable to a licensed source, the time from question to brief, or the number of decisions with a complete sign-off record.
3. **Rollout, workflow by workflow.** Extend to the next workflow only when the first one meets its measures. Build the sign-off gate and decision records in from the start, so the audit trail grows with usage.

Book an OJA Diagnostic
One workflow, one to two weeks, a gap map you can act on.
kiteedge.co.uk/contact-us

ABOUT KITEEDGE

KiteEdge builds governed AI decision infrastructure for institutions where decisions need to be defensible, licensed and audit-ready. The company has been building since 2017, and is based in London.

Its first vertical, **Contours**, serves buy-side capital markets. It is a provenance-first research platform that reasons over a licensed corpus of institutional research, links every claim to its source, and supports human sign-off before outputs are acted on. KiteEdge also runs a consulting practice for financial institutions, grounded in building and operating these platforms for real users.

kiteedge.co.uk · contours.ai

Sources

1. Aiera, “The Aiera Lift”, 24 June 2026, aiera.com/newsroom/the-aiera-lift. A vendor-run benchmark of 13 models on 150 analyst-style questions, scored by key-facts entailment; figures are averages across models.

© 2026 KiteEdge Ltd. Registered in England and Wales, Company No. 10213486.