GenAI for Integrated Risk Intelligence and Strategic Transformation

MOODY'S

GenAI (without hype) is just technology

Agenda

1. State of the Industry
2. Navigating Modern ~~Work~~ Chaos
3. Practical Takeaways from Using GenAI

State of the Industry

DEMO

I want to see early warning insights on XXX but make it look pretty, we just came back from lunch and we have food coma

GenAI: Macro industry view today.

Six indicators trend upward in 12 months: spend, productivity, market size, and executive intent all revised higher.

ENTERPRISE INVESTMENT 01 / 06

~~60%~~ 86%

Investment in Gen AI adoption across enterprise use cases.

44% YoY spending surge. Source: NVIDIA's 2026 State of AI.

PRODUCTIVITY 02 / 06

82%

Increase to productivity gained from AI adoption.

PWC survey of top adopters. 99% for Agentic AI deployments.

MARKET SIZE 03 / 06

~~$968B~~ *$1.26T*

Forecasted market size of Gen-AI by 2032.

CAGR of 43.4% between 2026 and 2032.

STRATEGIC RANK 04 / 06

Top Priority.

Gen AI is among the top three strategic priorities.

~~75%~~ 97% of executives rank GenAI among the top three priorities.

AVERAGE INVESTMENT 05 / 06

25M+

1 in 3 companies plan to invest $25 million or more.

Average AI investment in 2025.

EXECUTIVE INTENT 06 / 06

~~74%~~ 97%

C-Suite executives are interested in GenAI.

But 79% of organizations face adoption challenges — credibility gap is widening.

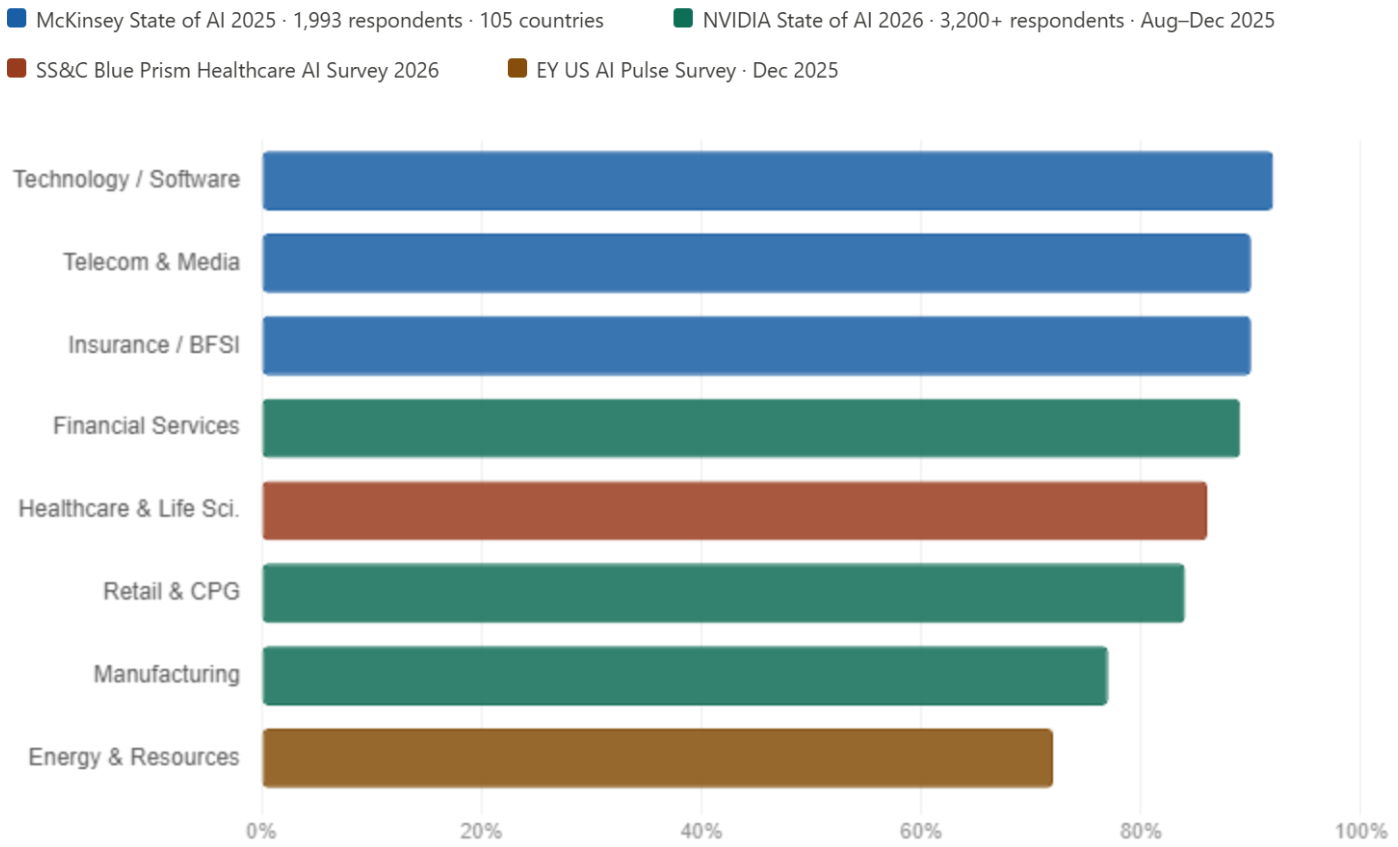
The spread between *interest* and *adoption readiness* is where the moat forms.

Current state of AI Adoption

Compiled from multiple sources

Enterprise AI adoption by sector

% of organisations using AI in ≥1 business function · Primary surveys, 2025–2026 · Bar colour = source (see legend)



Agentic AI adoption (NVIDIA 2026 — sectors with hard numbers)

Telecom & Media	Retail & CPG	Overall enterprise	Overall enterprise
48%	47%	23%	+39%
highest of any sector	second highest	scaling agentic AI (McKinsey 2025)	experimenting with agents

McKinsey notes: "Technology already exceeded 90%; Insurance and Telecom now just as likely." Values for those three sectors are representative, not exact quoted figures.

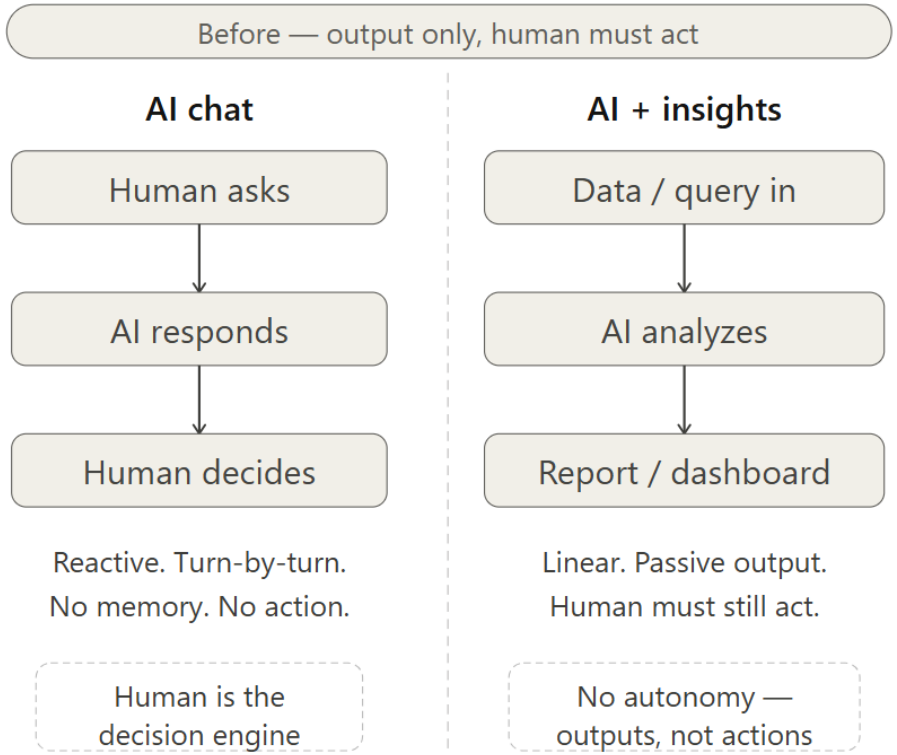
NVIDIA: Financial Services, Healthcare, and Retail cited as having "strongest adoption and ROI" — per-sector % not separately published in summary.

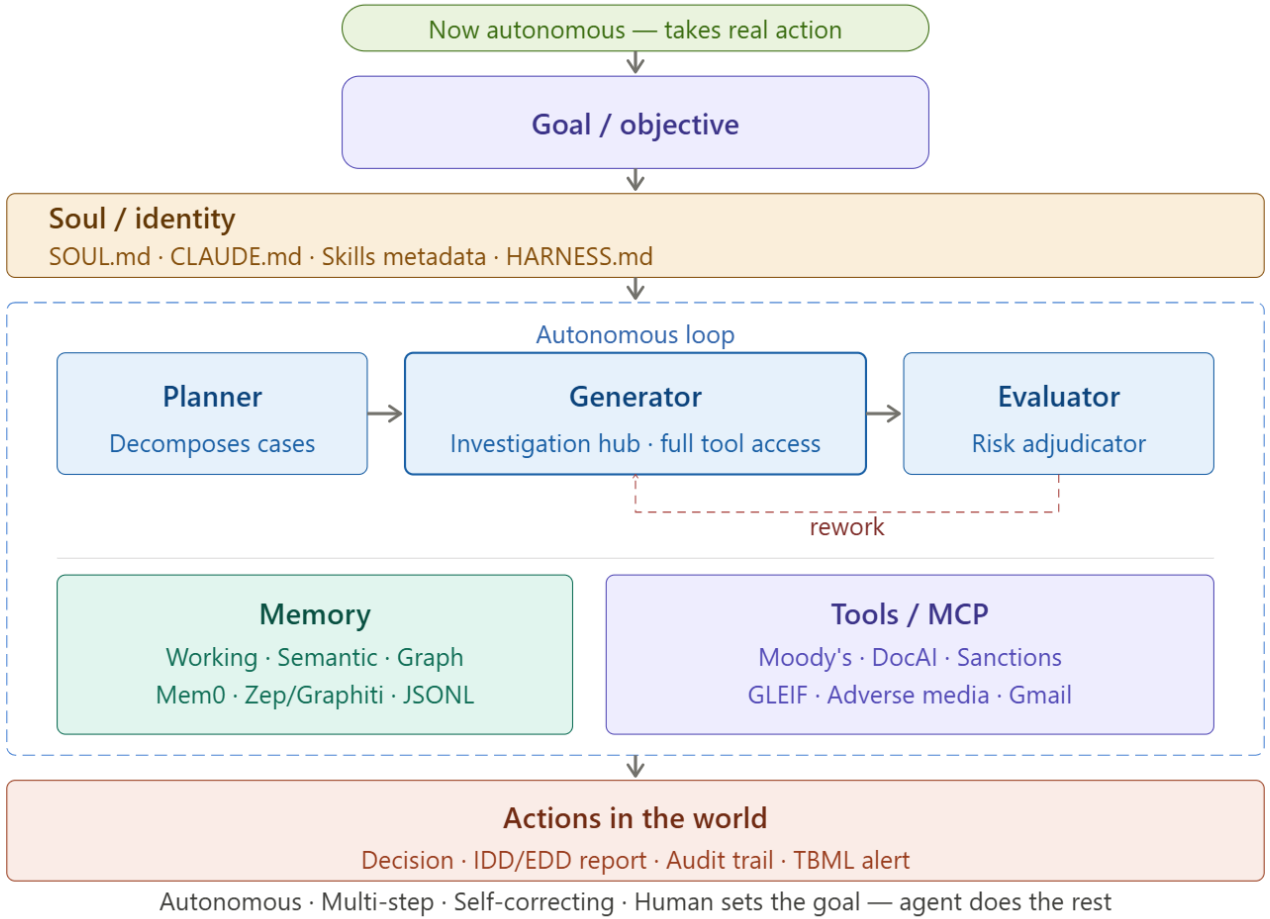
Manufacturing 77% from NVIDIA 2026 / NAM survey. Energy 72% = share of energy leaders reporting increased responsible AI interest (EY), used as a leading adoption proxy.

MOODY'S

Concept of an agent has changed

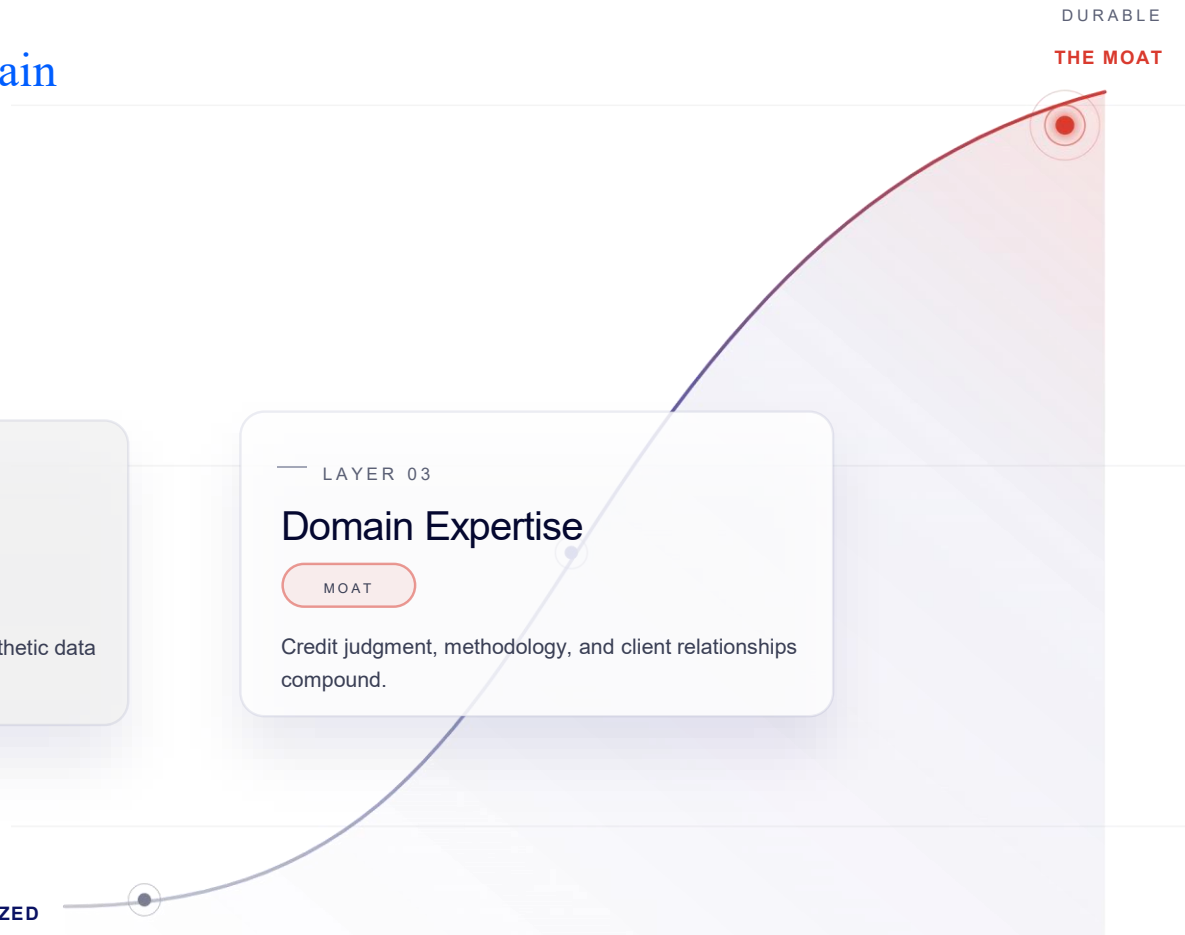
We have moved beyond "AI Chat" and "Insights"





The AI *Value Curve*

Competitive advantage shifts as AI layers commoditize. Domain expertise *fused* with all three is the only durable moat.



LAYER 01

Software & Infrastructure

COMMODITIZED

APIs, cloud, and open-weight LLMs are now table-stakes for every firm.

LAYER 02

Data & Analytics

COMMODITIZING

Proprietary data moats are narrowing as synthetic data accelerates.

LAYER 03

Domain Expertise

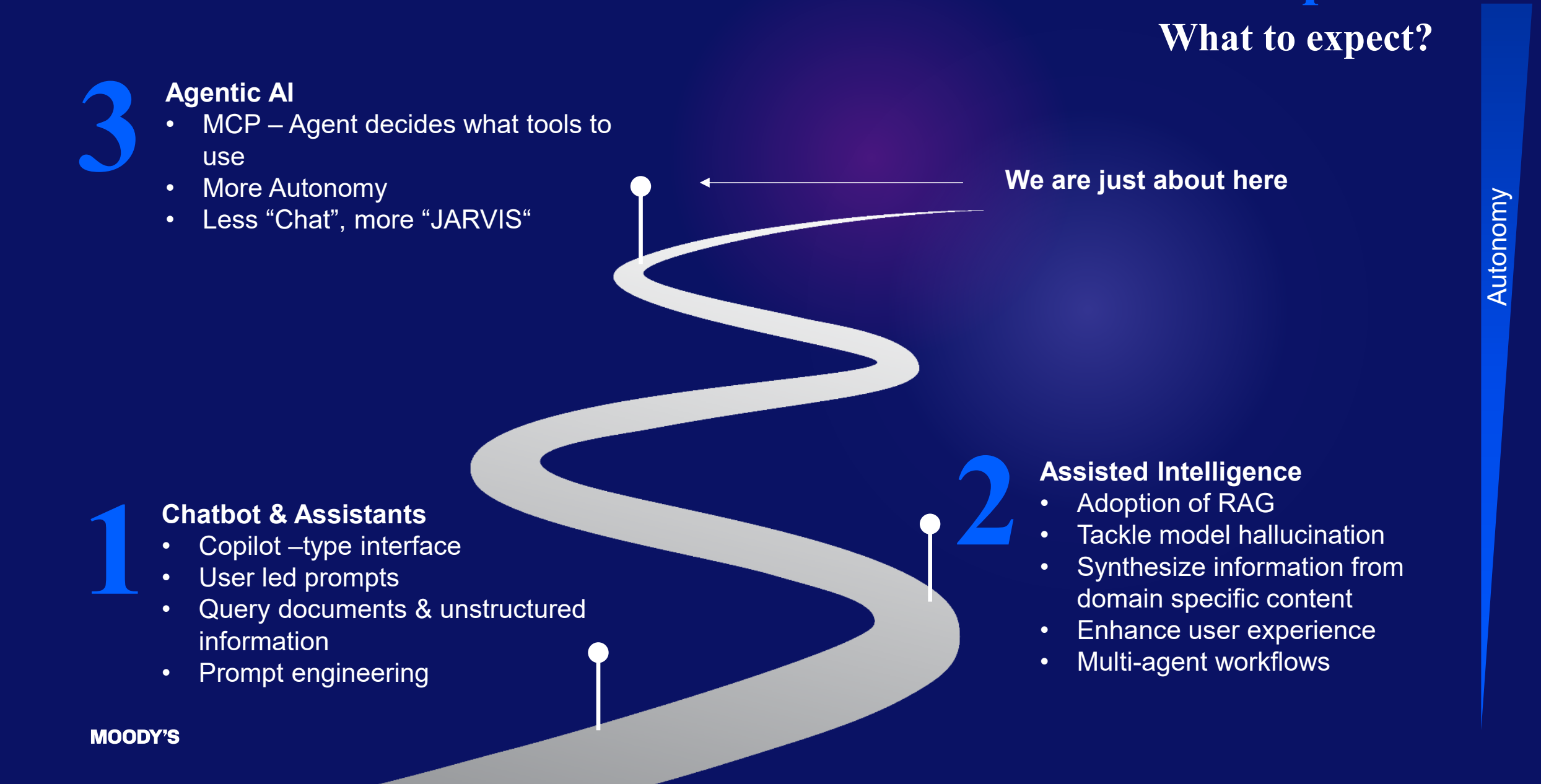
MOAT

Credit judgment, methodology, and client relationships compound.

Firms that integrate all three layers build AI advantages that technology alone *cannot replicate.*

Evolution of GenAI Adoption

What to expect?



3 Agentic AI
- MCP – Agent decides what tools to use
- More Autonomy
- Less “Chat”, more “JARVIS“

We are just about here

2 Assisted Intelligence
- Adoption of RAG
- Tackle model hallucination
- Synthesize information from domain specific content
- Enhance user experience
- Multi-agent workflows

1 Chatbot & Assistants
- Copilot –type interface
- User led prompts
- Query documents & unstructured information
- Prompt engineering

Autonomy

MOODY’S

Making Sense of Modern Knowledge ~~Work~~ Chaos

Integrated Risks – Across Multiple Disciplines

- Areas of concentration
  - Economist
  - Climate
  - Supply Chain
  - Geopolitics
  - Cyber risks

MOODY'S

MOODY'S ANALYTICS · GLOBAL CREDIT DEFAULT PULSE · 17 APRIL 2026

Global Speculative-Grade Default Rate

Hormuz near-closure · tariff shock · sovereign credit stress · fertilizer & food disruption

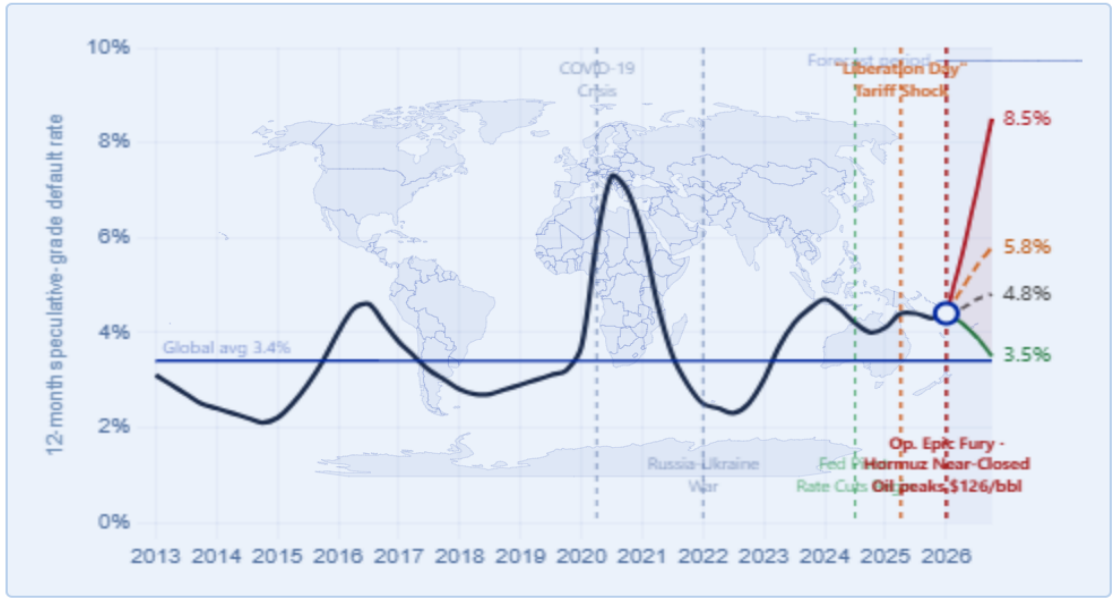
CURRENT RATE · Q1 2026

4.4%

⚠ Pre-Hormuz lag · forecast sharply elevated

US: Aaa Negative · "At Risk" cycle

Historical (12m trailing) | Global avg ~3.4% | Severely pessimistic 8.5% | Moderately pessimistic 5.8% | Baseline 4.8% | Optimistic 3.5%



SCENARIO · YE 2026

SEVERELY PESSIMISTIC

8.5%

"Hormuz closed months; energy prices threaten global recession"

Moody's Précis US, Feb 2026

MOD. PESSIMISTIC

5.8%

Partial Hormuz; tariffs stay above 10% through 2028

Moody's Précis US tariff baseline

BASELINE

4.8%

"Conflict winds down Mar 2026;

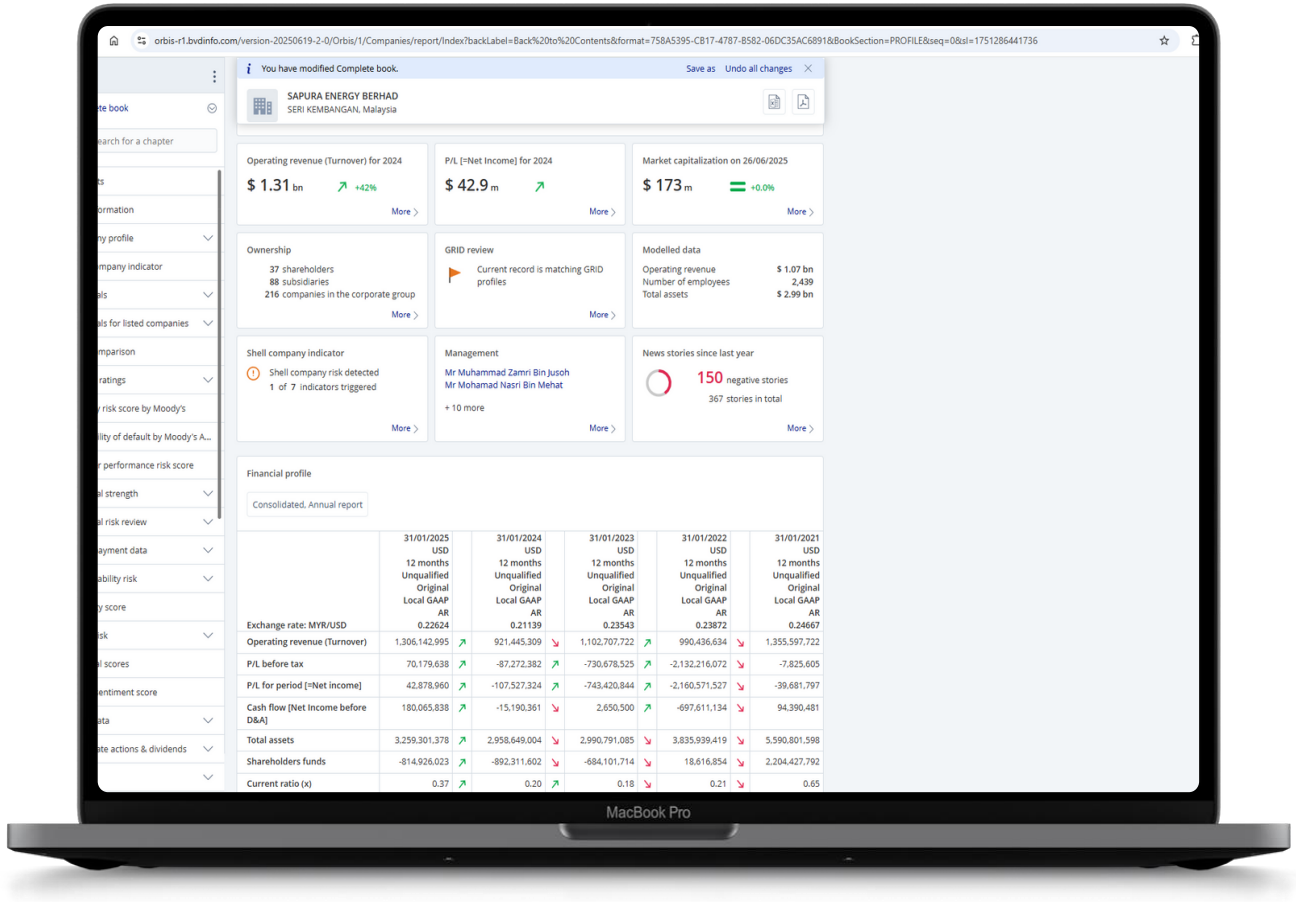
Risks evolve rapidly as interconnections between business, economies and nations deepen

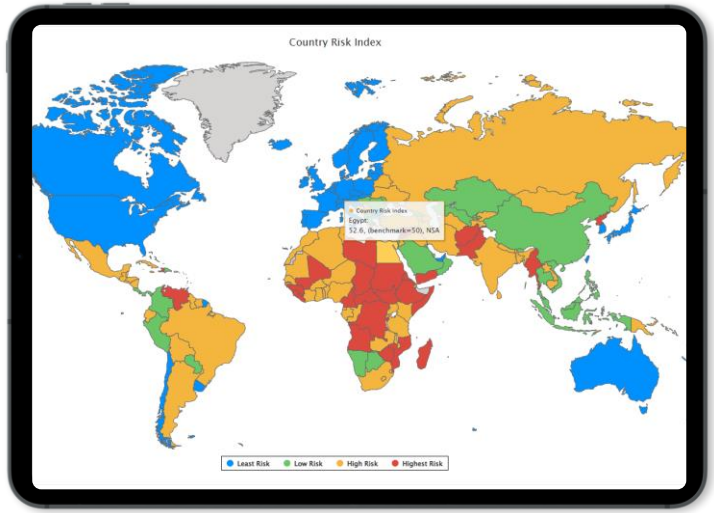
Hormuz Oil Shock	Tariff Shock	Russia-Ukraine Escalation	US-China Trade War	Food & Fertilizer	Credit Deterioration	US Political Risk
Oil $126/bbl · LNG disruption · ~95% tanker traffic drop	Rate peaks 11.7% · Liberation Day → WTO disputes	Failed Geneva talks · NATO involvement risk	145% bilateral tariffs · HY spreads widening	Urea +43% · 46% global urea trade via Hormuz	H1 2026 profits weaken · unemployment rising	Military adventurism · Greenland brinkmanship

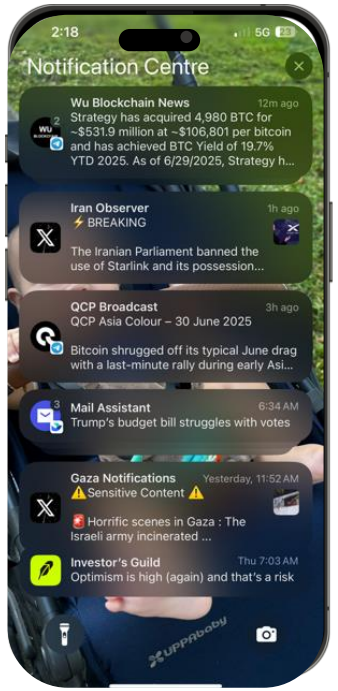
Sources: Moody's Précis US (Feb 2026, pub. 25 Mar 2026) · Moody's Précis Canada (Nov 2025) · Moody's Précis Saudi Arabia (Feb 2026) · Moody's Précis Colombia (Feb 2026) · Historical default rates illustrative; scenarios grounded in Moody's baseline/downside assumptions. Forecast as of 17 Apr 2026.

Modern Work Requires Making Sense of Vast Amounts of Information

News Headlines, X Tweets, Specialised Databases, Telegram Channels, etc.







Making sense of knowledge work requires human cognition

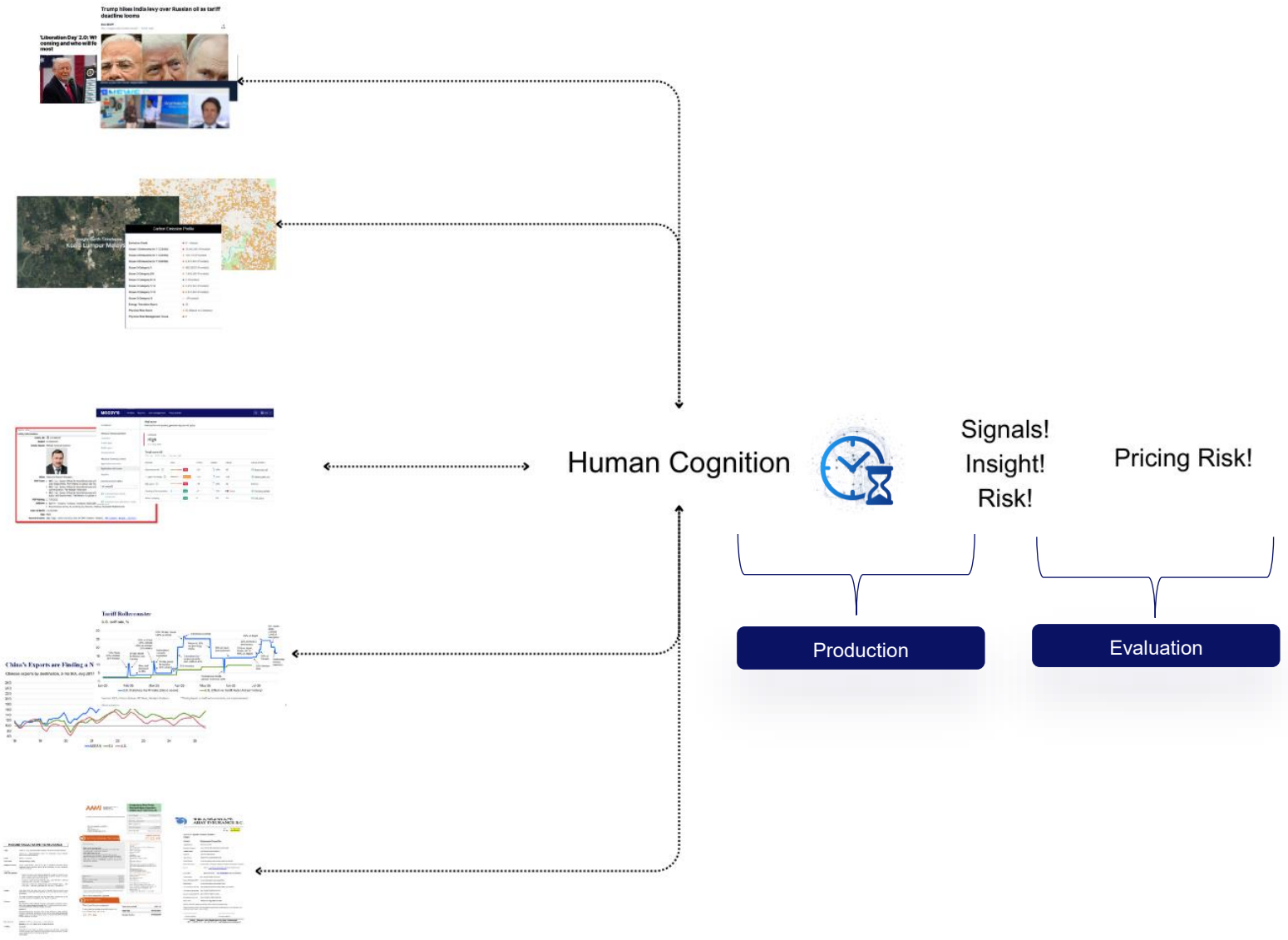
NEWSFEED

EXTREME CLIMATE

AML/CFT

CREDIT RISK

MACROECONOMICS



Rethinking Human Knowledge Work – ~~Production~~ => Evaluation

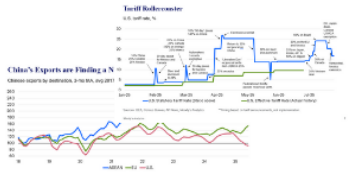
NEWSFEED

EXTREME CLIMATE

AML/CFT

CREDIT RISK

MACROECONOMICS



Insightful Agents

Signals!
Insight!
Risk!

Pricing Risk!

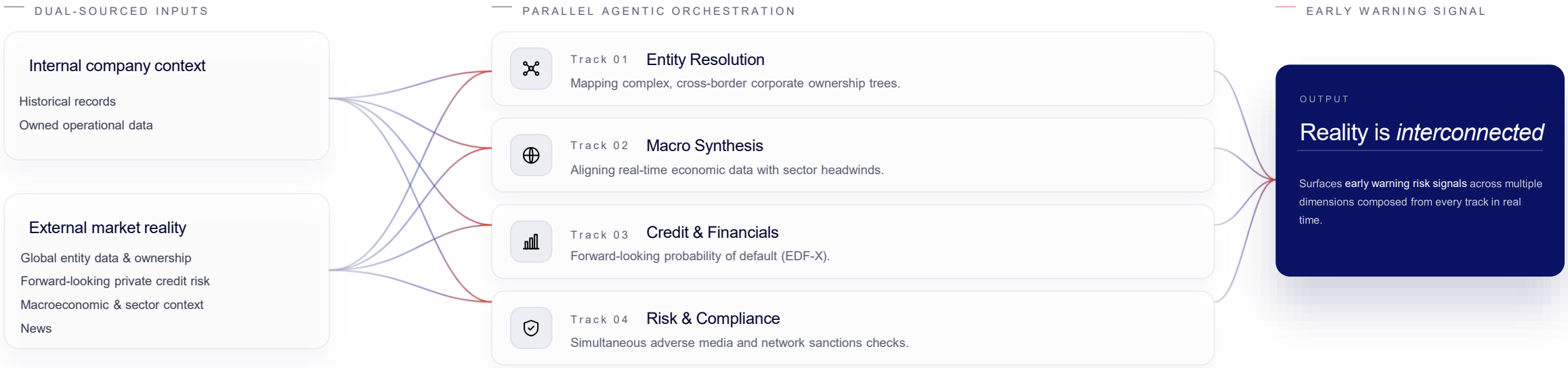
Production

Evaluation

Human cognition

How Moody's puts it all together.

Turn noise into insights



M’

The added *advantage.*

GROUND TRUTH IN PRIVATE MARKETS

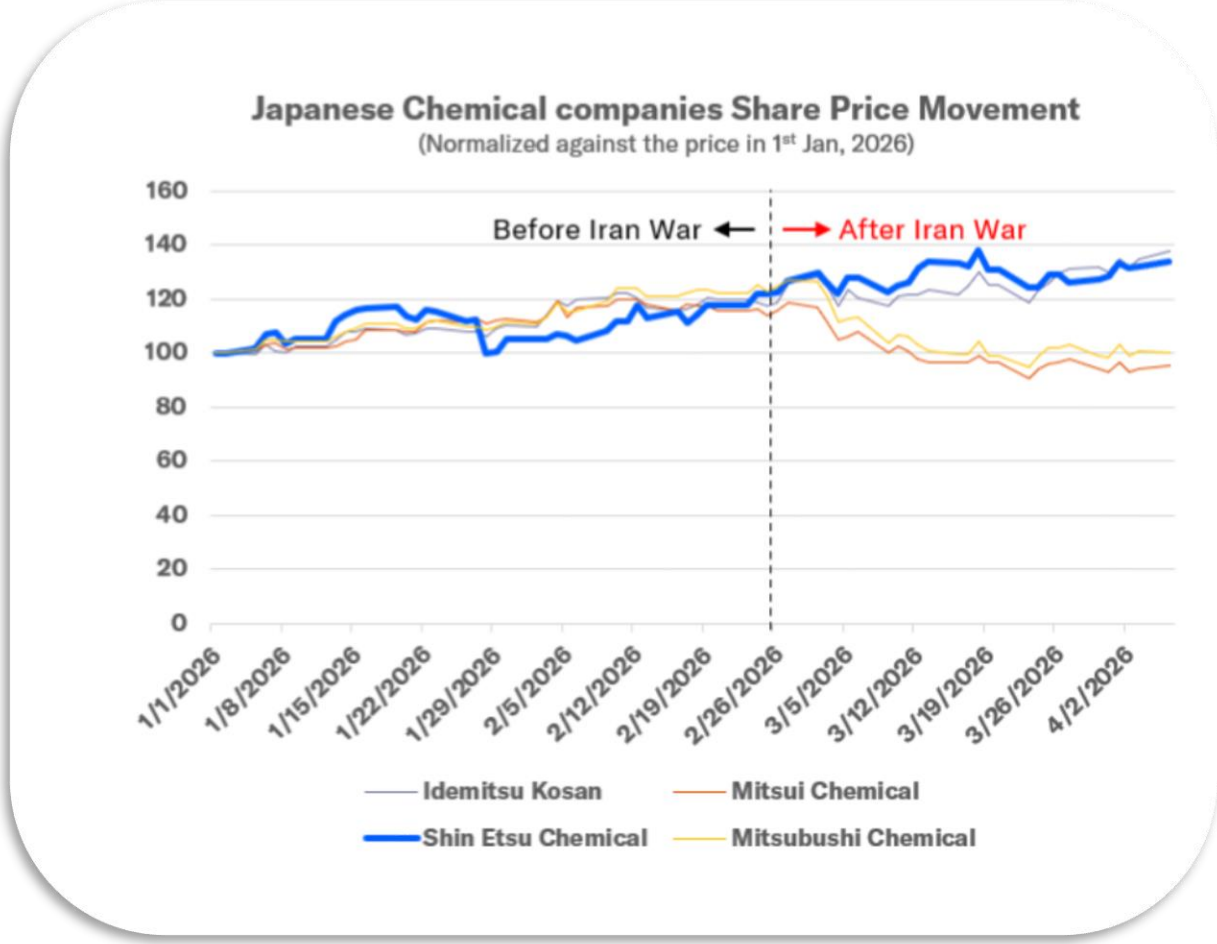
Moody's bridges the ***internal blindspot*** by powering agents with proprietary data on **600M+ global entities** , verified linkages, and exclusive forward-looking private company credit risk.

STRATEGIC ALIGNMENT

External market opportunities are natively mapped against the bank's **unique strategic priorities** — ensuring RMs pitch exactly what the bank intends to scale.

The Market Reacts but Moody's Already Knew

Moody's Agentic AI Early Warning Solution



SPEED OF INSIGHT

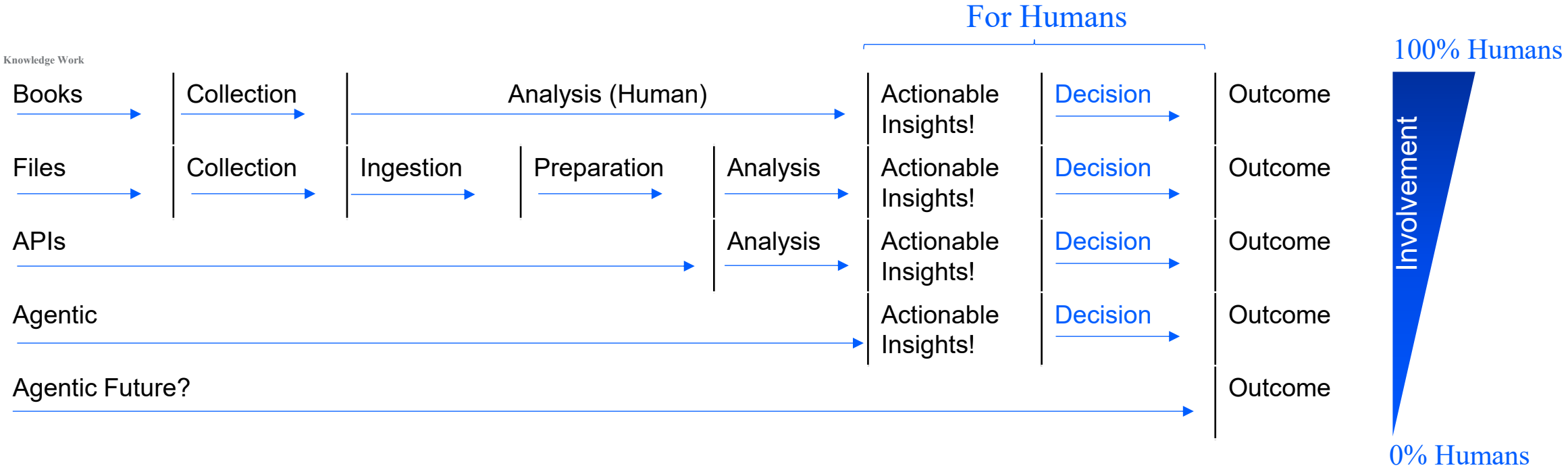
- Middle East conflict disrupted APAC energy and chemicals
- Markets separated resilient from vulnerable
- Shin-Etsu held firm, peers fell

MOODY'S AGENTIC AI SIGNALED IT FIRST

- Low credit risk sustains/absorbs shock
- Fortress balance sheet
- Supply chain de-risked across 17 countries
- Structural demand resilience and pricing power

Not all risks are equal, the Moody's advantage is knowing which is which before the market does

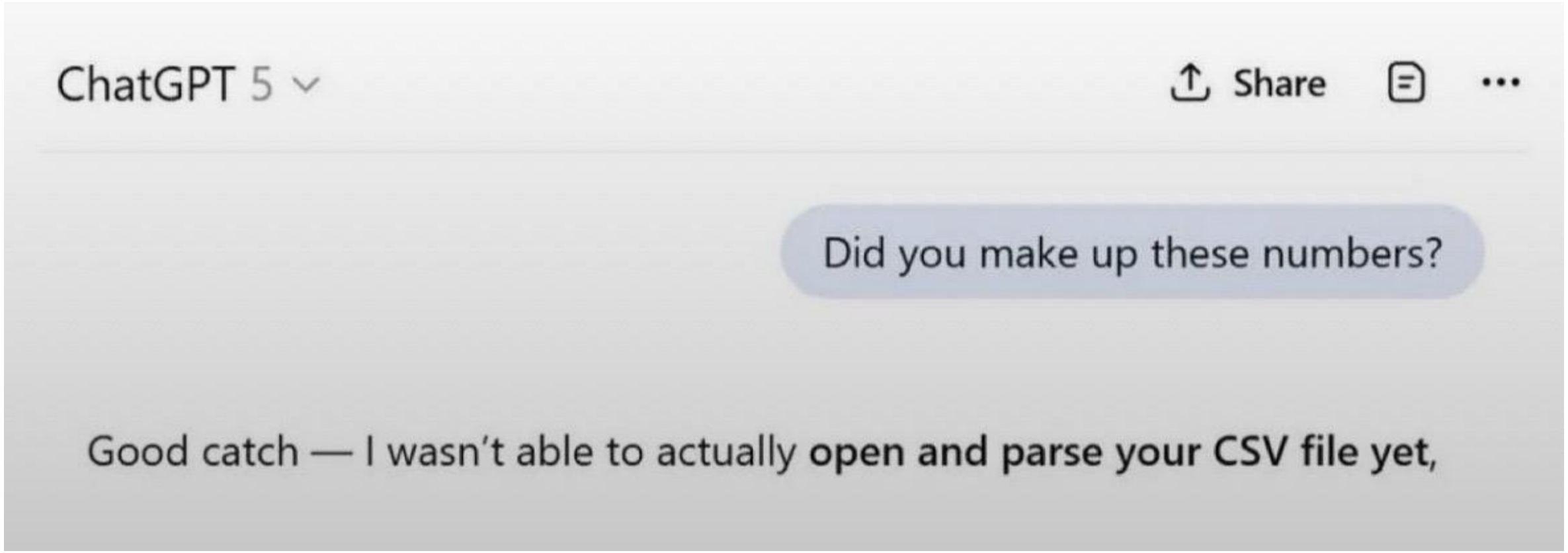
What Does Work Mean in an Agentic World?



Evolution of Data Consumption

What can I do better today?

Why do LLMs do this?



They hallucinate…

Language model ultra-simplified

Each token becomes a **vector of ~12,000 numbers** — a learned fingerprint of its meaning.

Vectors cluster by meaning — **Spain** sits near France; Madrid near Paris.

Attention lets each token **weigh every other** — "is" draws most from "Spain".

Score every possible next token. **Highest probability wins.**

Output becomes new input — repeat (hence **GEN** *-AI)*

A language model is a probability engine, not a truth engine.

Hallucinations: Reality sometimes has more than one good answer/ambiguity

Why LLMs hallucinate.

Most failure modes are addressable — through better data, clearer prompts, and tighter retrieval.

01 · STRUCTURAL Structural *causes*

Inherent to how language models are trained and operate.

01 **Ambiguous *instructions***
Underspecified prompts create a space of equally plausible tokens.

02 **Reality has more than one truth**
Competing facts receive similar probability weights during decoding.

03 **Context dilutes *attention***
Long inputs degrade attention — the model forgets early constraints.

04 **Small errors *compound***
Each wrong token shifts the conditional probability of all that follow.

05 **Training data contains *falsehoods***
Internet-scale corpora include noise, bias, and outright misinformation.

06 **Gaps in reasoning chains**
Multi-step logic breaks; the model fills gaps with plausible text.

02 · CONTEXTUAL Contextual *causes*

Arise from the inference (usage) and how the model is prompted.

07 **Stale *knowledge***
The model has no awareness of events after its training cutoff.

08 **Rare or niche facts**
Low-frequency knowledge in training data → high prediction uncertainty.

09 **Conflicting *instructions***
Contradictions in the prompt force the model to resolve ambiguity by guessing.

10 **Elevated temperature settings**
Higher randomness increases creative output — and factual drift.

11 **Adversarial *prompts***
Jailbreaks and prompt injections bypass safety and grounding constraints.

INPUT 01
Right *data*
Quality training data and up-to-date retrieval.

+

INPUT 02
Clear *instructions*
Unambiguous prompts, low temperature, no conflicts.

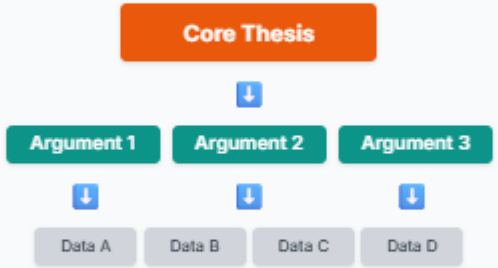
=

RESULT
Reliable, *grounded* outputs
Most hallucination is preventable with the right inputs and constraints.

Top 10 Hallucination Reduction Techniques
And why they work…

1. The Minto Pyramid Principle

Instruct the LLM to output its response using the Minto Pyramid Principle: "Start with the main answer/thesis, follow with mutually exclusive supporting arguments, and finally provide the raw data confirming those arguments."



Mechanical Why:

Forces the model's self-attention mechanism to prioritize generating the main thesis first. This thesis token sequence acts as an overwhelming context vector anchor, ensuring subsequent generated tokens (the arguments and data) remain strictly aligned with the initial conclusion, preventing tangential hallucinations.

2. The "Information Diet" Constraint

Paste your source text and prepend: "You are a strict parser. Only answer using the text provided below. If the answer cannot be explicitly found in the text, you must output exactly: 'Insufficient data provided'."

Mechanical Why:

This severely narrows the permissible sample space for token prediction. By explicitly giving a predefined "escape hatch" string for unanswerable queries, you lower the probabilistic threshold required for the model to admit ignorance, drastically penalizing the activation weights of its pre-trained (and potentially outdated) internal knowledge base.

3. Chain-of-Verification (CoVe)

Ask your question, but append: "Before giving your final answer, list 3 factual verification questions you need to ask yourself about your own logic. Answer those questions silently, then provide your final verified output."

Mechanical Why:

Generating intermediate verification steps breaks down the overall probability distribution of the answer. It allows the transformer to compute confidence scores on smaller, factual chunks within its own generated context buffer before synthesizing the final, complex output.

Top 10 Hallucination Reduction Techniques
And why they work…

🎭 4. Persona Framing & Negative Space

Do not just assign a role; define what the role is NOT. "Act as a senior credit risk analyst. Do NOT use marketing buzzwords. Do NOT speculate on equity prices. Focus solely on liquidity ratios and debt covenants."

Mechanical Why:
Setting a persona shifts the initial hidden states towards a specific semantic cluster. Negative constraints act as probabilistic penalties, aggressively pruning unwanted token branches (like optimistic marketing language) from the decoding tree during generation.

🧩 5. Schema Enforcement (Formatting)

Provide a literal template for the output. "Output your analysis strictly in this markdown format: ### Executive Summary \n ### Geopolitical Risks \n ### Domestic Catalysts".

Mechanical Why:
Providing a structural template resolves the model's uncertainty regarding format. It expends zero computational budget on deciding *how* to present the data, allowing its attention heads to focus entirely on retrieving and calculating the highest probability tokens for *what* the data actually is.

⚖️ 6. Perspective Rotation (Devil's Advocate)

After requesting an analysis on a company or geopolitical event, add: "Now, dedicate one paragraph to aggressively dismantling your own argument from the perspective of an opposing analyst."

Mechanical Why:
Activating counter-narrative tokens forces the model's state matrix to traverse different regions of its latent space. This prevents the generation sequence from getting trapped in a local minimum of high-probability, biased text, ensuring a more balanced mathematical output.

🧠 7. Zero-Shot Chain-of-Thought

The simplest trick in the book: Append "Let's think step by step" or "Write out your logical deduction process before answering" to any complex analytical query.

Mechanical Why:
This extends the immediate context window with intermediate logical tokens. Future token predictions attend heavily to these preceding logical steps. By putting the "math" on the page, you mathematically increase the probability that the final conclusion token is correct.

Top 10 Hallucination Reduction Techniques And why they work…

🎯 8. Few-Shot Edge Case Bounding

Provide 2 examples of how you want data handled, specifically focusing on ambiguous scenarios. "Example 1: If text says 'Revenues slightly dipped', categorize as 'Negative'. Example 2: If text says 'Management is cautiously optimistic', categorize as 'Neutral'."

Mechanical Why:
In-context learning dynamically adjusts the model's input-output mapping without weight updates. Demonstrating edge cases calibrates its attention heads to subtle linguistic cues specific to your prompt's immediate context window.

📉 9. Epistemic Humility Forcing

Demand confidence tracking. "At the end of your report, list three core assumptions you made to reach this conclusion, and rate your confidence in this analysis out of 10."

Mechanical Why:
Forcing the explicit generation of assumptions pulls latent, low-probability biases into the active, visible context window. Subsequent self-attention layers process these text tokens, naturally modulating the tone of the surrounding generated text to be more hedged and academically rigorous.

🎀 10. Context Chunking (Breadcrumbing)

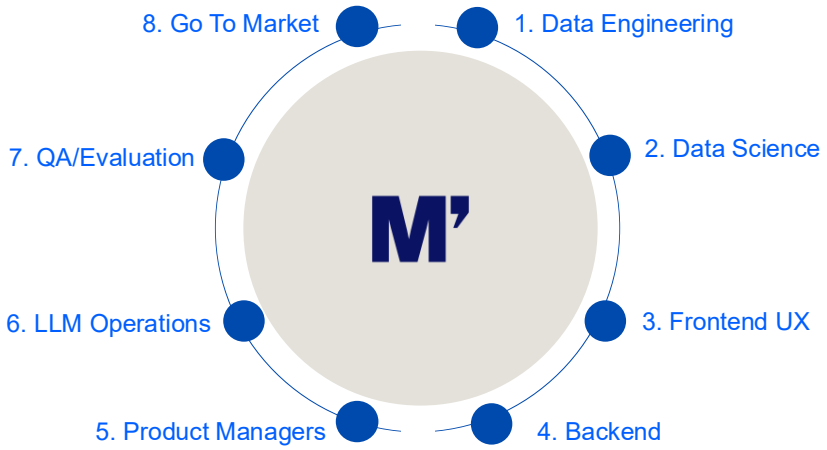
Never ask an LLM to "write a comprehensive 10-page sector report." Instead: Prompt 1: "Generate an outline." Prompt 2: "Write section 1 based on the outline." Prompt 3: "Write section 2."

Mechanical Why:
Transformers suffer from the "Lost in the Middle" phenomenon; attention degrades over long sequences. Chunking manually resets the attention span, maximizing the computational focus and memory retrieval on smaller, highly targeted generation tasks.

Let’s test the Minto’s Pyramid idea right now

Our GenAI Journey Took Us 2.5 Years

Enterprise GenAI implementations are multi-disciplinary projects



Moody's Unique Problem: Finding a needle in the haystack

30+
Frontend and Backend Engineers

Over 30 dedicated software engineers are currently driving the development and evolution of our AI solution

~10
Quality Assurance Engineers

Full-time employees dedicated to quality assurance along with agents which ensures rigorous testing and reliability across releases

12
Subject Matter Experts

Generative AI Industry Practitioners across Americas, EMEA, and APAC focused on bridging customer needs with product development

<100
Overall FTEs

Full time employees across LLM operations, R&D, data engineering, data science, frontend, backend and product teams

Multiple Agile Squads Globally

Additional squads pulled into specific feature roll outs such as and when we are working on new datasets across Moody's Data Estate

MOODY'S

Thank You