MOODY'S

Connected intelligence

Where enterprise AI actually gets built

AUTHORS

Cristina Pieretti, GM and Head of Digital Content and Innovation

Trevor O'Brien, MD, Head of Engineering and Architecture, Digital Content and Innovation

Executive Summary

Enterprise AI has a context problem. Every large organisation now has access to powerful language models — but access to models has not translated into reliable, defensible AI outcomes. The organisations achieving measurably better results are those that have engineered what surrounds the model: the retrieval pipelines, evaluation frameworks, governance infrastructure, and domain knowledge that determine whether a model's output is accurate enough, auditable enough, and current enough to act on. At Moody's, we call this surrounding infrastructure connected intelligence — and this paper argues that it is where enterprise AI competition will be won or lost.

The paper opens by explaining how AI context works at a technical level: how models process tokens, why context windows create real constraints even at one million tokens, and why filling that window with the right information — in the right structure, at the right moment — is a harder engineering problem than it first appears. It then traces the industry's shift from prompt engineering to context engineering: the recognition that the model's output is determined less by the words of the prompt than by the configuration of everything surrounding it.

The paper then walks through the architecture of a production-grade context layer at Moody's — covering ingestion and normalization, retrieval architecture, evaluation frameworks, and governance and auditability — before explaining how each of these challenges is amplified in financial services across four dimensions: temporal validity, regulatory auditability, cross-entity complexity, and the stakes of the output.

A dedicated section on context management at scale introduces the four core strategies production agentic systems use — Write, Select, Compress, and Isolate — plus a fifth emerging dimension: tool management. Throughout, Moody's Agentic Credit Memo workflow is used as a concrete illustration of how these principles apply in practice: how an agent navigates financial statements, rating methodologies, and peer comparisons across multiple steps, manages context window constraints, isolates reasoning across specialised sub-agents, and produces a sourced, auditable output. The credit memo is used as the example because it is one of the most context-intensive workflows in financial services — but the same architecture underpins every workflow Moody's builds, from KYC screening and entity profiling to compliance monitoring and portfolio risk assessment.

The paper closes with an argument about durable competitive advantage: that the model layer is commoditising, and that the real differentiator is the proprietary data estate, knowledge graph, and domain expertise encoded into the intelligence layer that surrounds it. A competitor can replicate a retrieval pipeline. They cannot replicate 600 million entities, two billion ownership links, and the continuous feedback loop of real-world customer deployments that sharpens connected intelligence with every use.

Moody's Agentic Solutions (MAS) is the product layer that activates connected intelligence — delivered through Moody's MCP Servers, purpose-built Agentic Workflows, and a partnership strategy that makes both available natively inside the AI environments customers already use, including Anthropic's Claude, Microsoft 365, OpenAI, AWS, and Databricks.

Every enterprise now has access to powerful language models. Few have built the infrastructure to make them reliable.

The gap between access and infrastructure is the defining challenge of production AI in 2026. Frontier models from OpenAI, Anthropic, Google, and others continue to improve rapidly, and for many common enterprise tasks their capabilities are converging. But convergence at the model layer has not translated into convergence of outcomes. The organizations getting measurably better results are not the ones with privileged access to a particular model. They are the ones that have engineered what surrounds it: the retrieval pipelines, evaluation frameworks, governance infrastructure, and domain-specific grounding that determine whether a model's output is robust enough to act on.

At Moody's, we call this surrounding infrastructure connected intelligence. It is the set of systems, domain knowledge, and governed data architecture that determine which information reaches the model, in what structure, at what moment, and subject to what constraints. As context engineering has matured from a niche concern into a recognised discipline — with Gartner identifying it as critical enterprise infrastructure and leading AI labs including Anthropic converging on its importance — one thing has become clear: connected intelligence is where enterprise AI competition will be won or lost. And in regulated industries like financial services, where yesterday's data can be actively dangerous and every output must be auditable, the stakes are higher than anywhere else. That is the problem Moody's connected intelligence is engineered to solve — and it is what every system, workflow, and partnership described in this paper is built on.

What's in the context layer?

The infrastructure that determines what an AI model knows at the moment it responds — spanning ingestion, retrieval, evaluation, and governance. In financial services, it's not a technical detail. It's where AI reliability is built or lost.

OUTPUT

Decision

Grounded, temporally valid, auditable response to the user or downstream agent

Governance & auditability

Full provenance logging, trace per agent step, regulatory defensibility

Evaluation

Continuous evaluation scoring (relevance, groundedness, answer quality) plus completeness and pertinence

Context management

Write, select, compress, isolate, and manage tools across the window

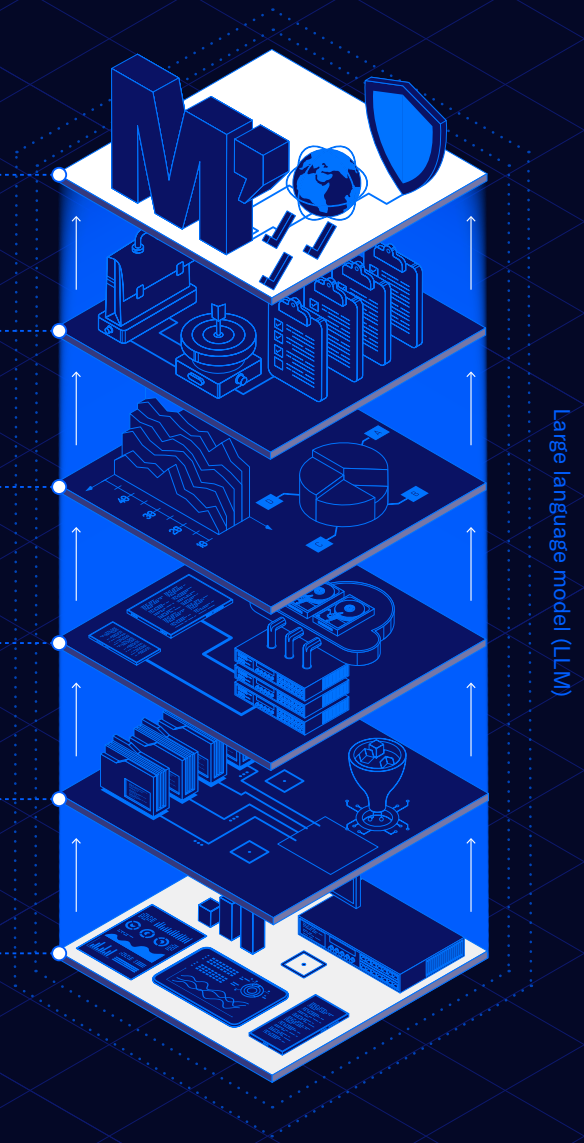
Retrieval architecture

Hybrid search (semantic + sparse), re-ranking, metadata filtering, domain-aware navigation

RAW DATA

Ingestion & normalization

Multimodal inputs transformed into structured, machine-readable form



The model is no longer the competitive advantage. In enterprise AI, outcomes are determined by how context is engineered — what data reaches the model, how it is structured, and whether it is current, grounded, and auditable.

How does AI context work?

Large language models don’t read text the way humans do.

They process information as tokens, which are fragments of text (typically a word, part of a word, or a punctuation mark) that have been converted into numerical representations the model can operate on.

Large context windows do not guarantee better outcomes. Without structure, prioritization, and signal density, more context can degrade accuracy rather than improve it.

For instance, the sentence “Moody’s downgraded the issuer’s credit rating” might be split into roughly eight to ten tokens, with words like “Moody’s” and “issuer’s” each broken into two.

TOKEN BREAKDOWN:

“Moody’s downgraded the issuer’s credit rating”

Each color block = 1 token

Every model has a finite context window, the maximum number of tokens it can process in a single interaction. Current frontier models — including Claude, Gemini, and GPT — now offer context windows of one million tokens or more, with some reaching two million. That sounds enormous until you consider what a complex enterprise workflow actually requires: ratings, research, financials, ownership data, and regulatory filings assembled across dozens of entities simultaneously. At that scale, context fills fast, and what gets included, excluded, or compressed directly determines the quality of the output. When a model generates a response, it attends to the tokens in its context window to determine what is relevant, but attention is not uniform. Earlier generations of models showed a tendency to attend more strongly to tokens near the beginning and end of the window, with material in the middle receiving less weight, a phenomenon called the ‘lost in the middle’ problem. Frontier models have made significant progress on this, but production systems operating at scale, particularly those assembling context from multiple heterogeneous sources, still encounter degraded attention over long, poorly structured inputs.

This means the context window isn’t just a storage issue, but also an attention constraint. Filling it with the right information, in the right order, with the right density, is an engineering challenge that grows more complex as enterprise applications scale and as systems move from single-query assistants to agents operating across many inference turns. The context layer exists to solve that challenge.

From prompt engineering to context engineering

The first wave of enterprise AI adoption centered on prompt engineering: crafting instructions so that models interpret them correctly.

That work was valuable and remains a component of any well-designed system. But the industry has learned, often through expensive failure; that prompts are only one input into a much larger system. As Anthropic's engineering team put it in their widely cited guidance on building effective agents: the challenge is no longer finding the right words for your prompts, but answering the broader question of "what configuration of context is most likely to generate the model's desired behavior?" That configuration encompasses system instructions, tool definitions, retrieved external data, protocol integrations, multi-turn message history, and, increasingly, the accumulated state of agents operating across many inference turns and longer time horizons.

Andrej Karpathy offered a widely adopted analogy for this shift: if the LLM is the CPU, the context window is RAM, and context engineering is the operating system that decides what to load into working memory and when.

The engineering challenge is one of optimization under constraint: finding the smallest possible set of high-signal tokens that maximizes the likelihood of a desired outcome, given finite attention budgets and the well-documented phenomenon of "context rot," where model performance degrades as poorly curated information accumulates in the window.

Production experience has sharpened this insight considerably. Research from production AI agent deployments at scale has shown that KV-cache hit rate — the efficiency with which the system reuses previously computed context — is the single most important metric for a production-stage agent, because the ratio of input tokens to output tokens in agentic systems runs approximately 100:1. The overwhelming majority of cost and latency comes not from generating responses but from processing context. Separately, Anthropic's recent work on harness engineering for long-running agents has demonstrated that even frontier models struggle to maintain coherent progress across multiple sessions without structured environments that manage context state between them. Their finding, that the space of productive harness designs shift instead of shrink as models improve, is a strong signal that the infrastructure surrounding models will remain a decisive engineering challenge regardless of how capable the models themselves become.

But infrastructure alone is not enough. Connected intelligence is not just the retrieval pipelines, evaluation frameworks, and governance systems that deliver information to the model. It is also the structured domain knowledge those systems draw on: the relationships between entities, sectors, and methodologies; the temporal state of every data point; the interpretive frameworks that determine whether a given piece of evidence is relevant, current, and authoritative. In financial services, a retrieval system that returns the right document is table stakes. Connected intelligence that understands how an issuer's credit profile connects to sector-level risk factors, which were revised after a specific regulatory event, and that all three carry independent temporal validity constraints is a fundamentally different capability. It is the difference between search and judgement.

Connected intelligence, properly understood, is the production infrastructure and the structured domain knowledge it encodes, working together to solve this problem at scale. In regulated industries, where outputs must be auditable, temporally valid, and grounded in authoritative sources, both the engineering and the knowledge it surfaces can rival the models themselves in complexity.

Prompt engineering solves instructions. Context engineering determines behavior. Production AI performance depends on how instructions, data, tools, and memory are orchestrated under constraint.

The architecture of the context layer

A production grade context layer is not a single system. It is a stack of subsystems, each independently engineered, evaluated, and continuously refined. What follows is how these subsystems work at Moody's.

1.

Ingestion and normalization

Transforms raw, multimodal data into clean, structured inputs.

2.

Retrieval architecture

Finds and ranks the most relevant, authoritative evidence.

3.

Evaluation frameworks

Continuously validates for accuracy, relevance, and groundedness.

4.

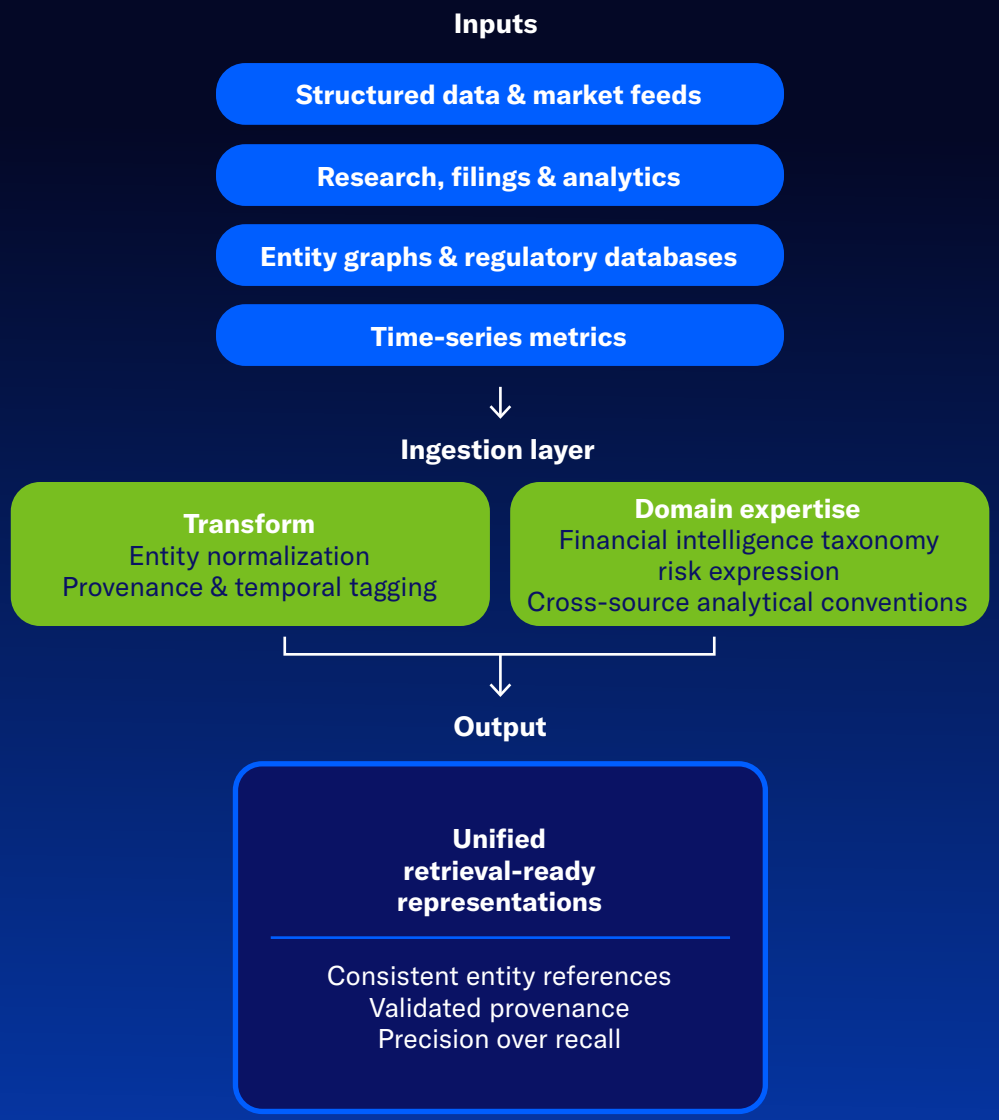
Governance and auditability

Logs every step for full traceability and regulatory compliance.

1 Ingestion and normalization

The financial intelligence that flows into Moody's connected intelligence arrives in heterogeneous forms: structured quantitative data and real-time market feeds alongside unstructured research and filings; entity relationship graphs alongside regulatory databases; time-series financial metrics alongside qualitative analytical content spanning multiple jurisdictions, source types, and publication cadences. Before retrieval can begin, the ingestion layer must transform these inputs into unified, machine-readable representations normalizing entity references, reconciling inconsistent formatting across source types, and ensuring that every data point carries the temporal and provenance metadata required to evaluate its validity at query time.

What differentiates Moody's ingestion layer is not the tooling, as parsing and normalisation have become table stakes, but the subject matter expertise encoded in how it is applied. Moody's Analytics brings deep knowledge of the taxonomy, structure, and analytical conventions of the full credit and compliance intelligence estate: how risk factors are expressed across structured and unstructured sources, where material qualifications appear, how triggers and conditions are stated and revised across source types and publication cycles. That expertise is what allows our ingestion and retrieval systems to navigate Moody's intelligence estate with the precision of a domain specialist, not the broad recall of a general-purpose search engine.



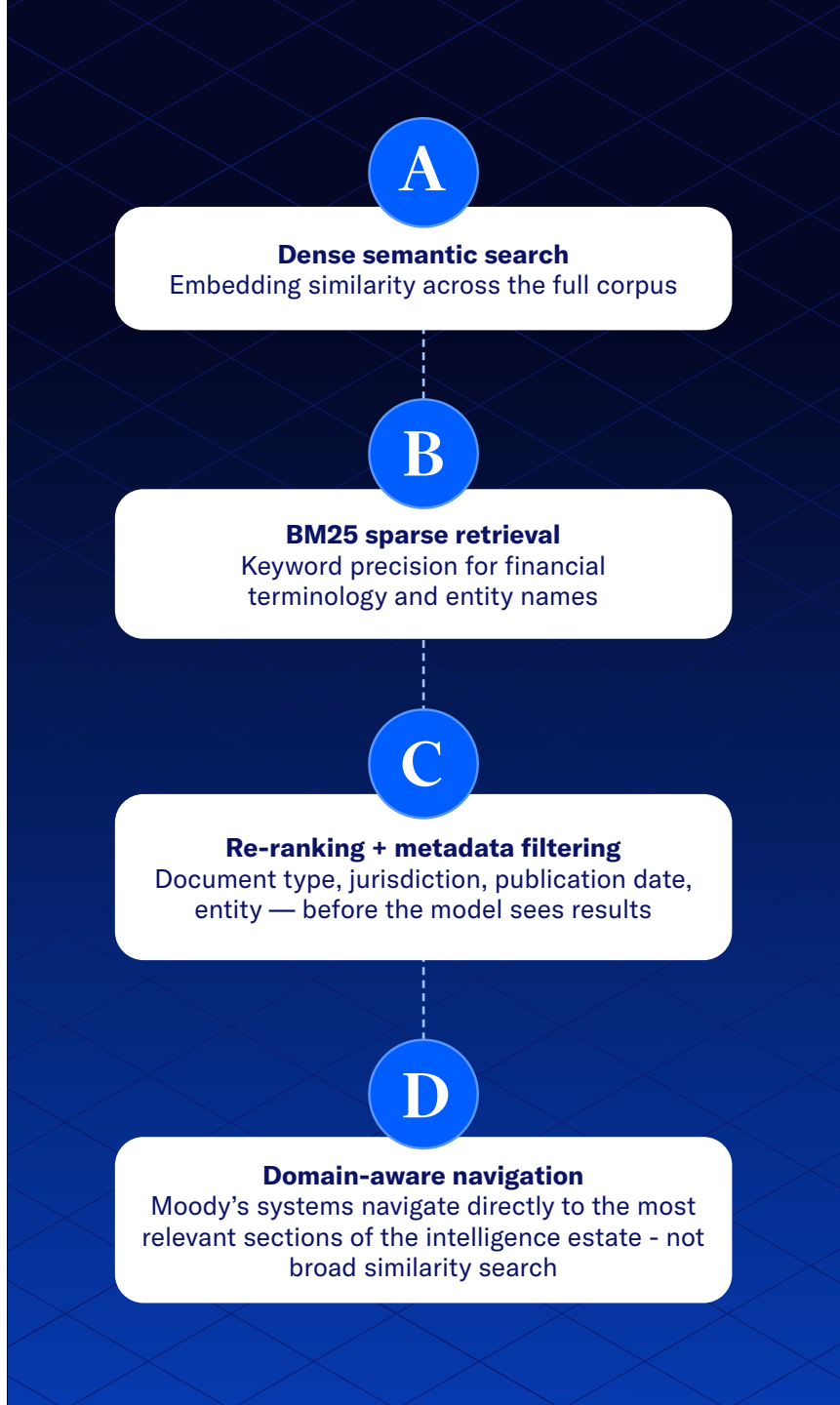
2 Retrieval architecture

Generic vector search, embedding a query and returning the most similar documents from a store, is insufficient for high-stakes enterprise applications. The failure mode is subtle: vector similarity captures semantic relatedness, but semantic relatedness is not the same as informational relevance. A passage about debt restructuring may be semantically close to a query about leverage ratios without actually containing the data needed to answer it.

Production retrieval systems address this through hybrid architectures that combine dense semantic search with sparse retrieval methods such as BM25, followed by re-ranking stages that apply learned or rule-based relevance scoring. In domain-specific applications, retrieval must also be metadata-aware, filtering by document type, jurisdiction, publication date, and entity so that temporal and regulatory constraints are enforced before the model ever sees the results.

Chunking strategy is a further critical variable. Effective chunking follows the structure of the underlying data — using section headers, paragraph boundaries, table delimiters, and data schema boundaries to produce chunks that preserve analytical coherence and ensure that quantitative and qualitative signals are not fragmented across retrieval units.

But retrieval architecture alone does not produce Moody's-grade context. What differentiates our approach is that domain knowledge is encoded directly into retrieval logic. When an agentic system assesses a company's credit profile, it retrieves and cross-references earnings call transcripts, credit ratings, regulatory filings, sector indicators, ESG disclosures, and peer comparisons — not as isolated sources but as a connected web of evidence, with each element linked through Moody's knowledge graph. A data point anchored in these structured relationships can be retrieved far more accurately than one floating in isolation. This is what we mean when we say connected intelligence encompasses knowledge, not just infrastructure. Moody's retrieval architecture layers a graph-based understanding of entities and relationships across the full intelligence estate, connecting evidence so that critical context is never lost.



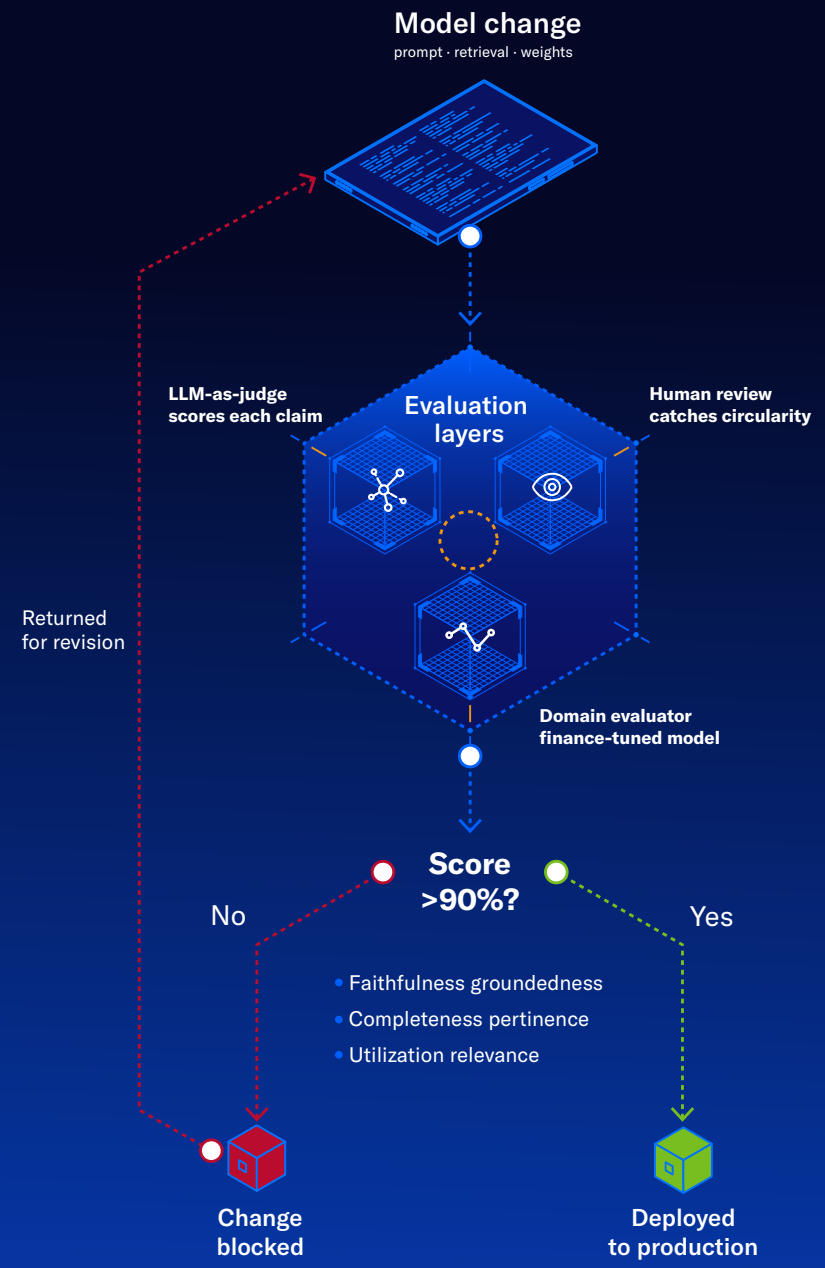
3 Evaluation frameworks

In enterprises, AI systems are scored across multiple dimensions. The industry has converged on a core set of evaluation dimensions: context relevance (do the retrieved sources actually contain answer-bearing information?), groundedness (is every claim in the response directly supported by the retrieved context, with no hallucinated additions?), and answer relevance (does the response address the user's actual question?). Each captures a distinct failure mode. A system can score highly on groundedness while failing on relevance if it faithfully reports content that does not address the query.

Evaluation must be granular enough to decompose generated answers into individual claims and verify each against retrieved source. This is commonly implemented through LLM-as-a-judge pipelines, where a separate model scores the primary model's output. It is worth acknowledging that this introduces a degree of circularity, using an LLM to evaluate an LLM's output carries its own failure modes, which is why Moody's supplements automated scoring with human-in-the-loop validation and purpose-built evaluator models tuned to financial domain accuracy. Evaluation results feed directly into CI/CD gates: if a retrieval or prompt change causes faithfulness to drop below a defined threshold, the change is blocked before it reaches production.

Beyond these core dimensions, production systems also track utilization (what fraction of retrieved context was actually used in the response?), completeness (did the response address all parts of the query?), and pertinence (is the response meaningful in context, or technically correct but analytically useless?). In finance especially, a single data point without surrounding analytical context is often less useful than a short paragraph that situates the figure within a trend. Our evaluation framework is designed to catch instances where expanding context encourages models to produce overly long answers, lose precision, or drift off-topic. Each update is measured against our benchmarks so we can demonstrate when accuracy is advancing and ensure it never slips.

Finding a document is not the same as producing insight. Judgment requires understanding relationships, relevance, and temporal validity — not just semantic similarity.



4 Governance and auditability

In regulated industries, every context decision must be traceable. Which documents were retrieved, which were excluded, how were they ranked, what evaluation scores did the output receive, and what model version generated the response. At Moody's, governance is not an afterthought applied to a working system, it is a prerequisite for deployment, built into the pipeline from the outset through purpose-built observability infrastructure that monitors every stage from query to response.

This matters more as systems become agentic. When an AI agent autonomously executes a multi-step credit assessment, retrieving earnings transcripts, cross-referencing rating methodologies, pulling sector benchmarks, and generating a synthesized output, the governance layer must track not just the final answer but the reasoning chain that produced it. Moody's Agentic Solutions are built with this requirement embedded: every tool call, every retrieval decision, every intermediate reasoning step is logged within an auditable framework.

Each agentic workflow produces a structured trace that records sources and data points retrieved at each step, the content used and can include other factors such as scores assigned by the re-ranker, the evaluation metrics for the generated output, and the model version used, creating a complete provenance chain from query to response that can be reviewed by compliance teams or presented to regulators.

In a regulatory environment where "what did the AI say?" is increasingly followed by "why did it say it?", this infrastructure is not optional.

Managing context at scale

The industry has converged on four core strategies for managing the context window at production scale, taxonomized by LangChain's widely adopted research on agent context engineering. Each addresses a different constraint; in practice, production systems use all four simultaneously.

WRITE

Persist information outside the context window for later retrieval. Agents working on complex tasks need to offload intermediate reasoning, plans, and accumulated state to external memory rather than relying on the window to hold everything. At Moody's, this is one of the techniques our agentic systems use to maintain continuity across multi-step credit assessments. As an agent works through financial statements, rating methodologies, and peer comparisons, intermediate findings can be persisted to external memory, so they survive context window limits and remain available for the final synthesis, rather than relying on the window alone to hold an increasingly complex analytical chain.

SELECT

Retrieve the most relevant fragments from a larger information store. As described in the retrieval architecture section, production-grade selection goes well beyond naive similarity search — it requires hybrid retrieval, re-ranking, metadata filtering, and in our case, domain-aware navigation that knows where material information lives across Moody's intelligence estate.

COMPRESS

Retain only the tokens required for the current task. This is where the 100:1 input-to-output ratio in agentic systems becomes a practical engineering problem. When a Moody's agent is processing a multi-year credit history spanning years of financial data, ratings history, and analytical content, raw context can easily exceed any model's effective window. Compression at agent-to-agent boundaries, where one agent's output is summarized into a concise payload before being passed to the next, is essential. The key insight is that what you remove from the context matters as much as what you include. A focused 300-token context grounded in the right evidence routinely outperforms an unfocused 100,000-token context that merely contains it.

ISOLATE

Partition context across specialized subsystems rather than loading everything into a single model's window. Moody's Agentic Solutions use this pattern extensively: when assessing a company's credit profile, separate agents handle financial analysis, sector dynamics, ESG disclosures, and peer benchmarking, each operating with a focused, task-specific context. Only compressed, verified outputs are shared between them. This isolation improves both accuracy and auditability, because each agent's reasoning chain can be inspected independently.

A fifth dimension is also emerging alongside these four: **tool management**. As agentic systems gain capabilities, their available action space, the set of tools, APIs, and data sources they can call on, grows rapidly. Production research on agentic systems has shown that dynamically adding or removing tools mid-workflow degrades both cache efficiency and output quality; the recommended approach is to keep tool definitions stable and instead mask availability based on the current stage of the task. For Moody's, where a credit assessment agent may need access to financial databases, rating methodologies, regulatory filing APIs, and news feeds at different stages, managing which tools are active at each step, without disrupting the context that has already been built, is a non-trivial engineering challenge and an active area of investment.

Why financial services amplify every challenge

Every challenge described in the preceding sections, ingestion, retrieval, evaluation, compression, isolation, tool management, is amplified in financial services. The data is multimodal, inconsistently structured, and time-sensitive in ways that are unforgiving of error.

A refinancing reported this morning can invalidate last week's debt ratios entirely. Yesterday's data is not merely stale; if it contradicts current reality, it is actively dangerous. This is the domain where context engineering is hardest, and where getting it wrong has the most immediate consequences.

In financial services, stale data isn't just neutral; it's actively dangerous. Every output must reflect the most current, authoritative evidence available at the moment of decision.

Amplifier 1

The first amplifier is **temporal validity**. Corpus freshness in financial services must be measured in hours, not days. Ingestion pipelines must detect when new filings, ratings actions, or market events have been published and update the retrieval index before the next query arrives. Failure to do so means the system may ground its response in outdated evidence while newer, contradictory information exists in the source corpus but has not yet been indexed. Standard context layer evaluation frameworks assume a relatively stable knowledge base; in financial services, the correct answer to a question can change between morning and afternoon. Evaluation must account for this, penalizing responses grounded in information that was accurate at the time of indexing but has since been superseded. At Moody's, this is not a theoretical concern, it is a daily operational requirement that shapes how we design every stage of the pipeline.

Amplifier 2

The second amplifier is **regulatory auditability**. Financial regulators across jurisdictions are converging on a clear expectation: AI systems used in material financial decisions must be explainable and auditable. In the UK, the FCA launched the Mills Review in January 2026 to examine how AI reshapes financial services, with recommendations due to the FCA Board in summer 2026 — and has confirmed that practical guidance on audit trails, explainability, and senior manager accountability under SMCR is expected by the end of 2026. The US SEC's 2026 examination priorities explicitly include reviewing whether firms can demonstrate how AI-driven decisions were reached, with examiners focused on whether AI representations are accurate and whether technology-driven recommendations align with regulatory obligations. In the EU, the AI Act imposes direct obligations on financial institutions: AI systems used in credit scoring and lending decisions are classified as high-risk, requiring full technical documentation, automatic logging of every decision, and human oversight mechanisms with compliance required from August 2026. Across jurisdictions, the direction of travel is consistent: governance is not a feature, it is a prerequisite. Every retrieval decision, every source ranking, every evaluation score must be logged and defensible. This is why Moody's connected intelligence is built with auditability embedded from the outset, not applied after the fact.

Amplifier 3

The third amplifier is **cross-entity complexity**. A credit assessment does not exist in isolation. An issuer's profile is connected to its sector's risk factors, which are influenced by macroeconomic conditions, which may be subject to regulatory actions that vary by jurisdiction. When an agentic system processes this web of dependencies, a flawed retrieval at any node can cascade through the entire analytical chain. This is why the domain knowledge encoded in the context layer, the structured relationships between entities, sectors, methodologies, and temporal states described earlier in this paper, is not an enhancement, it is a structural requirement. A general-purpose retrieval system has no way to know that a change in sovereign risk methodology should trigger re-evaluation of every corporate issuer in that jurisdiction. Connected intelligence built on a century of accumulated domain expertise does.

Amplifier 4

The fourth amplifier is the **stakes of the output**. In general-purpose enterprise AI, a flawed response is an inconvenience that can be corrected. In financial services and beyond, where the data is deep, proprietary, and temporally complex, that asymmetry between what can be built quickly and what can only be accumulated through validated data, domain-specific benchmarks, and years of real-world deployment is the definition of a durable advantage. The cost of error is not reputational embarrassment; it is financial, legal, and regulatory exposure. This is why faithfulness targets in financial services typically exceed 90%, why evaluation frameworks must decompose generated answers into individual claims and verify each against retrieved sources, and why systems must be designed to catch errors before outputs reach the user. In this domain, connected intelligence is not infrastructure. It is risk management.

If an AI system can't explain why it produced an output, it can't be deployed at scale in regulated markets.

Why the advantage compounds

The argument that the model layer is commoditizing is no longer a contrarian position. It is industry consensus. IBM's Chief AI Architect has described 2026 as a "buyer's market" where the model itself is not the main differentiator. Harvard Business Review published an analysis in February arguing that when every company has access to the same AI models, organizational context becomes the competitive advantage. InformationWeek's enterprise AI predictions counselled organizations to "architect your stack so you can swap in vendor innovations as they mature, while your real differentiation lives in the domain models, policies and evaluation data that no platform vendor can ship for you." For Moody's, this is not a future scenario to prepare for. It is the architectural principle we have already built against: our systems are model-agnostic by design, matching different models to different tasks across retrieval, summarization, classification, and structured extraction.

The more important insight is that Moody's connected intelligence - the proprietary data estate, knowledge graph, entity relationships, and temporal validity rules encoded across credit, compliance, ESG, and macroeconomic domains - is where competitive advantage compounds. Every iteration of retrieval logic, every refinement to evaluation benchmarks, every new data source integrated into the pipeline makes the system harder to replicate. But the compounding effect runs deeper than engineering. It extends to the structured domain knowledge that the engineering draws on: the relationships between 600 million entities, sectors, and methodologies; the temporal validity rules that govern when evidence expires; the interpretive frameworks that determine whether a retrieval is relevant to a decision or merely semantically adjacent. A competitor can replicate a retrieval pipeline. They cannot replicate the connected intelligence that pipeline draws on, nor the continuous feedback loop generated by customer interactions and real-world workflow deployments that sharpens the system with every use. In financial services and beyond, where the data is deep, proprietary, and temporally complex, that asymmetry between what can be built quickly and what can only be accumulated through validated data, domain-specific benchmarks, and years of real-world deployment is the definition of a durable advantage.

The more important insight is that the connected intelligence is where competitive advantage compounds.

How the context layer fails

Building the context layer also means understanding how it breaks. The strategies described above, write, select, compress, isolate, manage tools, are defenses against a set of failure modes that are now well-documented in production systems. The OWASP Top 10 for Agentic Applications 2026, the first industry-standard security framework for autonomous AI, formally catalogues many of these risks. Knowing the taxonomy matters, because each failure requires a different fix. Treating the wrong one wastes engineering effort and leaves the real problem unaddressed.

RETRIEVAL POISONING

An incorrect or outdated document retrieved early in a multi-turn interaction can cascade through subsequent reasoning steps. Because models treat retrieved context as authoritative, a single bad retrieval can corrupt an entire analytical chain. In financial services, this is particularly dangerous: a stale research document or outdated rating action retrieved alongside current data can lead an agent to ground its analysis in superseded evidence without flagging the contradiction. Research has demonstrated that as few as five carefully crafted documents injected into a RAG knowledge base can manipulate AI responses with a 90% success rate. For Moody's, this is why retrieval must be metadata-aware and temporally filtered, ensuring the system never treats an outdated document as current evidence.

EVALUATION DRIFT

As the underlying corpus changes, evaluation benchmarks go stale. A system that scored well on faithfulness last quarter may be performing materially worse today if the evaluation dataset no longer reflects the current distribution of queries and documents. Production data suggests that 67% of retrieval-augmented systems experience significant retrieval accuracy degradation within 90 days of deployment. In financial services, where the corpus changes daily, evaluation drift is not a risk to monitor, it is a certainty to engineer against. This is why evaluation at Moody's is a continuous operational process, with benchmarks recalibrated as the underlying data evolves.

Reliable agentic systems are built by design — through isolation, evaluation, and traceability.

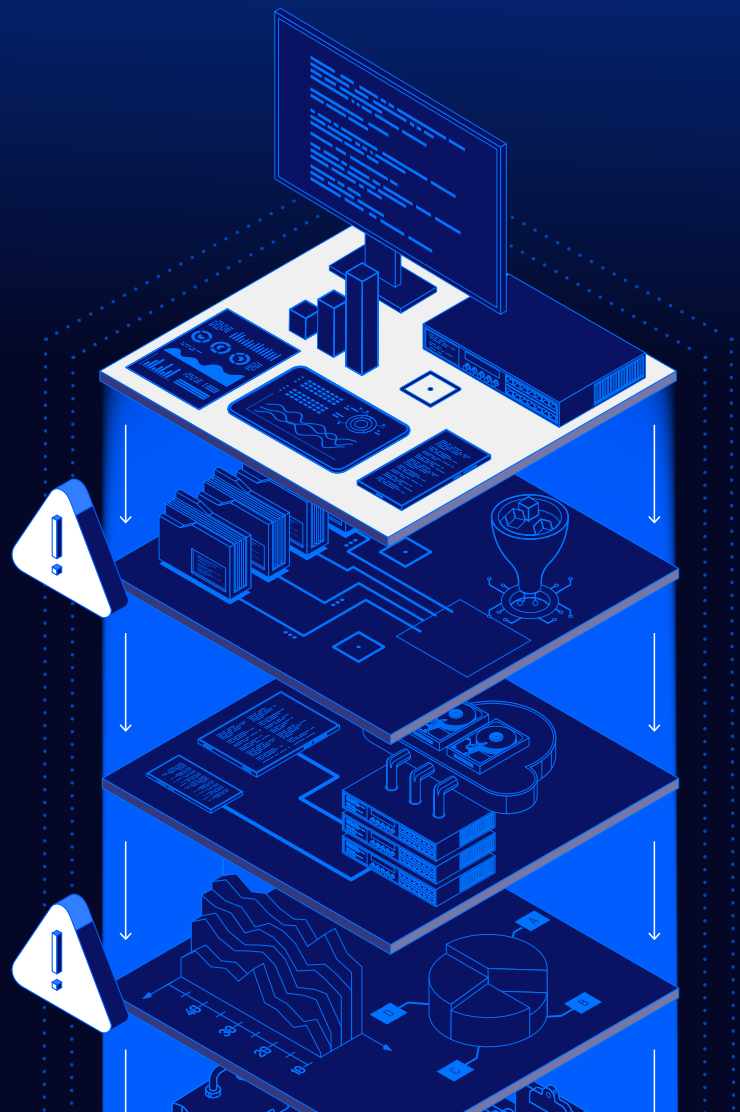
CASCADING FAILURE ACROSS AGENTS

As systems move from single-model architectures to multi-agent workflows, a new class of failure emerges: errors that propagate across agent boundaries. When one agent's flawed output becomes another agent's input context, the error compounds through the chain. Simulation research has found that a single compromised or malfunctioning agent can corrupt downstream decision-making across an entire agent network within hours. In Moody's agentic workflows, where separate agents handle financial analysis, sector assessment, and peer benchmarking before their outputs are synthesized, the isolation strategy described earlier is not just an efficiency measure, it is a containment mechanism. Each agent's reasoning chain can be inspected and validated independently before its output enters the next stage.

MEMORY AND CONTEXT POISONING AS A SECURITY VECTOR

As agentic systems gain persistent memory and access to external data sources, a new threat has emerged: deliberate contamination of an agent's knowledge base or long-term memory. Unlike a prompt injection that affects a single session, memory poisoning persists across sessions and can cause subtle behavioral drift that is extremely difficult to detect. The OWASP Top 10 for Agentic Applications 2026, a globally peer-reviewed framework developed by more than 100 security researchers and practitioners, identifies this class of risk as part of its new taxonomy for autonomous AI systems, alongside agent goal hijacking, tool misuse, and cascading trust failures. Microsoft's own taxonomy of agentic AI failure modes classifies it separately from traditional prompt injection, noting that it specifically targets systems that autonomously decide what to store and retrieve. For any context layer operating in regulated financial services, this reinforces the case for strict provenance tracking on every document in the retrieval corpus and governance controls over what enters the knowledge base. When the industry's leading security framework formally catalogues these risks, they are not theoretical concerns, they are design constraints.

These failure modes are not one-time problems to solve. They are ongoing operational realities that require continuous monitoring, benchmark recalibration, and structured engineering investment. At Moody's, this work is embedded in how we operate, not as periodic audits but as part of the production pipeline itself.



What we've built

The models will keep getting better. That is not the question. The question is whether the infrastructure and knowledge surrounding those models is engineered to capture the improvement, and in financial services, whether it can do so at the standard of accuracy, auditability, and temporal validity the domain demands.

For over a century, Moody's has built one of the deepest reservoirs of financial intelligence in the world, spanning credit ratings, risk analytics, regulatory and compliance data, ESG and climate assessments, economic forecasting, and structured finance. That intelligence is not siloed by product line. It is interconnected: entities linked to sectors, sectors linked to sovereigns, risk factors linked to regulatory regimes and climate exposures. A sovereign downgrade affects every corporate issuer in that jurisdiction. A climate event changes the risk profile of an entire sector, which reprices the structured finance instruments built on it, which affects the portfolios holding those instruments, which changes the compliance risk profile of the institutions exposed to them. Understanding how those cascades propagate, and encoding that understanding into retrieval logic, evaluation frameworks, and governance infrastructure, is what Moody's has engineered into connected intelligence. This is not a model. It is the accumulated knowledge of interconnected risk, and the infrastructure that makes it available to any model, at the moment of decision.

At the foundation of everything Moody's has built is connected intelligence: a unified knowledge graph spanning more than 600 million entities and two billion ownership links, interconnecting credit ratings, research, financials, ownership structures, news sentiment, ESG data, and risk signals across credit, compliance, and operational domains. This is not a data warehouse. It is a living architecture of interconnected risk knowledge, continuously updated, rigorously validated, and encoded with the interpretive frameworks that only a century of credit and compliance expertise can produce. Connected intelligence is what makes Moody's context layer different in kind, not just degree, from general-purpose alternatives. Moody's Agentic Solutions (MAS) is the product layer that activates connected intelligence. MAS comprises two integrated components: Moody's MCP Servers, which give AI systems direct, protocol-level access to Moody's proprietary intelligence estate in real time; and purpose-built Agentic Workflows - coordinated systems of domain-specific agents that automate end-to-end processes across credit risk, lending, and KYC and compliance, engineered to replicate the rigour of specialised analytical processes with outputs that are sourced, explainable, and auditable. MAS is available through two distribution channels. Directly, through Moody's own platforms, where customers can access MCP Servers and Agentic Workflows in the environments Moody's operates and controls. And through partnerships, embedded natively inside the AI platforms where customers already work, with each integration designed around what that environment can uniquely deliver.

Through Anthropic's Claude, Moody's became the first financial services firm in the world to launch an MCP App: fully built, end-to-end credit and compliance workflows, including credit memo generation, peer comparisons, scorecard assessments, entity profiling, ownership mapping, adverse media screening, and sanctions checks running natively inside Claude Desktop, Claude.ai, and Claude Enterprise, rendered as interactive, auditable reports without leaving the interface. Through Microsoft, Moody's MCP Servers and a dedicated Moody's agent are embedded directly inside Microsoft 365 Copilot, Researcher, and Excel, putting connected intelligence inside the productivity environment used by over a billion people. Through OpenAI, Moody's MCP Servers are available inside ChatGPT Enterprise, giving customers building AI in OpenAI environments direct access to Moody's proprietary intelligence. Through AWS Marketplace, Moody's Agentic Credit Memo workflow is deployable directly into customers' existing cloud infrastructure. And through the Databricks Data Intelligence Platform, Moody's GenAI-ready data is available to teams building and scaling AI solutions on Databricks. The distribution strategy is deliberate: connected intelligence and Moody's Agentic Solutions wherever customers choose to work, without friction, without a new login, without asking them to move.

For business and technology leaders evaluating AI investments, the implication of everything in this paper is concrete: assess your AI strategy not by the sophistication of the model you have deployed, but by the maturity of the connected intelligence - the infrastructure and domain knowledge - that surrounds it. In regulated financial services, connected intelligence is not a technical enhancement. It is risk management. And that is precisely what Moody's has built.

MOODY'S

CREDIT RATINGS ISSUED BY MOODY'S CREDIT RATINGS AFFILIATES ARE THEIR CURRENT OPINIONS OF THE RELATIVE FUTURE CREDIT RISK OF ENTITIES, CREDIT COMMITMENTS, OR DEBT OR DEBT-LIKE SECURITIES, AND MATERIALS, PRODUCTS, SERVICES AND INFORMATION PUBLISHED OR OTHERWISE MADE AVAILABLE BY MOODY'S (COLLECTIVELY, "MATERIALS") MAY INCLUDE SUCH CURRENT OPINIONS. MOODY'S DEFINES CREDIT RISK AS THE RISK THAT AN ENTITY MAY NOT MEET ITS CONTRACTUAL FINANCIAL OBLIGATIONS AS THEY COME DUE AND ANY ESTIMATED FINANCIAL LOSS IN THE EVENT OF DEFAULT OR IMPAIRMENT. SEE APPLICABLE MOODY'S RATING SYMBOLS AND DEFINITIONS PUBLICATION FOR INFORMATION ON THE TYPES OF CONTRACTUAL FINANCIAL OBLIGATIONS ADDRESSED BY MOODY'S CREDIT RATINGS. CREDIT RATINGS DO NOT ADDRESS ANY OTHER RISK, INCLUDING BUT NOT LIMITED TO: LIQUIDITY RISK, MARKET VALUE RISK, OR PRICE VOLATILITY. CREDIT RATINGS, NON-CREDIT ASSESSMENTS ("ASSESSMENTS"), AND OTHER OPINIONS INCLUDED IN MOODY'S MATERIALS ARE NOT STATEMENTS OF CURRENT OR HISTORICAL FACT. MOODY'S MATERIALS MAY ALSO INCLUDE QUANTITATIVE MODEL-BASED ESTIMATES OF CREDIT RISK AND RELATED OPINIONS OR COMMENTARY PUBLISHED BY MOODY'S ANALYTICS, INC. AND/OR ITS AFFILIATES. MOODY'S CREDIT RATINGS, ASSESSMENTS, OTHER OPINIONS AND MATERIALS DO NOT CONSTITUTE OR PROVIDE LEGAL, COMPLIANCE, INVESTMENT, FINANCIAL OR OTHER PROFESSIONAL ADVICE, AND MOODY'S CREDIT RATINGS, ASSESSMENTS, OTHER OPINIONS AND MATERIALS ARE NOT AND DO NOT PROVIDE RECOMMENDATIONS TO PURCHASE, SELL, OR HOLD PARTICULAR SECURITIES. MOODY'S CREDIT RATINGS, ASSESSMENTS, OTHER OPINIONS AND MATERIALS DO NOT COMMENT ON THE SUITABILITY OF AN INVESTMENT FOR ANY PARTICULAR INVESTOR. MOODY'S ISSUES ITS CREDIT RATINGS, ASSESSMENTS AND OTHER OPINIONS AND PUBLISHES OR OTHERWISE MAKES AVAILABLE ITS MATERIALS WITH THE EXPECTATION AND UNDERSTANDING THAT EACH INVESTOR WILL, WITH DUE CARE, MAKE ITS OWN STUDY AND EVALUATION OF EACH SECURITY THAT IS UNDER CONSIDERATION FOR PURCHASE, HOLDING, OR SALE.

MOODY'S CREDIT RATINGS, ASSESSMENTS, OTHER OPINIONS, AND MATERIALS ARE NOT INTENDED FOR USE BY RETAIL INVESTORS AND IT WOULD BE RECKLESS AND INAPPROPRIATE FOR RETAIL INVESTORS TO USE MOODY'S CREDIT RATINGS, ASSESSMENTS, OTHER OPINIONS OR MATERIALS WHEN MAKING AN INVESTMENT DECISION. IF IN DOUBT YOU SHOULD CONTACT YOUR FINANCIAL OR OTHER PROFESSIONAL ADVISER.

ALL INFORMATION CONTAINED HEREIN IS PROTECTED BY LAW, INCLUDING BUT NOT LIMITED TO, COPYRIGHT LAW, AND NONE OF SUCH INFORMATION MAY BE COPIED OR OTHERWISE REPRODUCED, REPACKAGED, FURTHER TRANSMITTED, TRANSFERRED, DISSEMINATED, REDISTRIBUTED OR RESOLD, OR STORED FOR SUBSEQUENT USE FOR ANY SUCH PURPOSE, IN WHOLE OR IN PART, IN ANY FORM OR MANNER OR BY ANY MEANS WHATSOEVER, BY ANY PERSON WITHOUT MOODY'S PRIOR WRITTEN CONSENT. FOR CLARITY, NO INFORMATION CONTAINED HEREIN MAY BE USED TO DEVELOP, IMPROVE, TRAIN OR RETRAIN ANY SOFTWARE PROGRAM OR DATABASE, INCLUDING, BUT NOT LIMITED TO, FOR ANY ARTIFICIAL INTELLIGENCE, MACHINE LEARNING OR NATURAL LANGUAGE PROCESSING SOFTWARE, ALGORITHM, METHODOLOGY AND/OR MODEL.

MOODY'S CREDIT RATINGS, ASSESSMENTS, OTHER OPINIONS AND MATERIALS ARE NOT INTENDED FOR USE BY ANY PERSON AS A BENCHMARK AS THAT TERM IS DEFINED FOR REGULATORY PURPOSES AND MUST NOT BE USED IN ANY WAY THAT COULD RESULT IN THEM BEING CONSIDERED A BENCHMARK.

All information contained herein is obtained by MOODY'S from sources believed by it to be accurate and reliable. Because of the possibility of human or mechanical error as well as other factors, however, all information contained herein is provided "AS IS" without warranty of any kind. MOODY'S adopts all necessary measures so that the information it uses in assigning a credit rating or assessment is of sufficient quality and from sources MOODY'S considers to be reliable including, when appropriate, independent third-party sources. However, MOODY'S is not an auditor and cannot in every instance independently verify or validate information received in the credit rating or assessment process or in preparing its Materials.

To the extent permitted by law, MOODY'S and its directors, officers, employees, agents, representatives, licensors and suppliers disclaim liability to any person or entity for any indirect, special, consequential, or incidental losses or damages whatsoever arising from or in connection with the information contained herein or the use of or inability to use any such information, even if MOODY'S or any of its directors, officers, employees, agents, representatives, licensors or suppliers is advised in advance of the possibility of such losses or damages, including but not limited to: (a) any loss of present or prospective profits or (b) any loss or damage arising where the relevant financial instrument is not the subject of a particular credit rating or assessment assigned by MOODY'S.

To the extent permitted by law, MOODY'S and its directors, officers, employees, agents, representatives, licensors and suppliers disclaim liability for any direct or compensatory losses or damages caused to any person or entity, including but not limited to by any negligence (but excluding fraud, willful misconduct or any other type of liability that, for the avoidance of doubt, by law cannot be excluded) on the part of, or any contingency within or beyond the control of, MOODY'S or any of its directors, officers, employees, agents, representatives, licensors or suppliers, arising from or in connection with the information contained herein or the use of or inability to use any such information.

NO WARRANTY, EXPRESS OR IMPLIED, AS TO THE ACCURACY, TIMELINESS, COMPLETENESS, MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OF ANY CREDIT RATING, ASSESSMENT, OTHER OPINION OR INFORMATION IS GIVEN OR MADE BY MOODY'S IN ANY FORM OR MANNER WHATSOEVER.

Moody's Investors Service, Inc., a wholly-owned credit rating agency subsidiary of Moody's Corporation ("MCO"), hereby discloses that most issuers of debt securities (including corporate and municipal bonds, debentures, notes and commercial paper) and preferred stock rated by Moody's Investors Service, Inc. have, prior to assignment of any credit rating, agreed to pay Moody's Investors Service, Inc. for credit ratings opinions and services rendered by it. MCO and all MCO entities that issue ratings under the "Moody's Ratings" brand name ("Moody's Ratings"), also maintain policies and procedures to address the independence of Moody's Ratings' credit ratings and credit rating processes. Information regarding certain affiliations that may exist between directors of MCO and rated entities, and between entities who hold credit ratings from Moody's Investors Service, Inc. and have also publicly reported to the SEC an ownership interest in MCO of more than 5%, is posted annually at www.ir.moodys.com under the heading "Investor Relations — Corporate Governance — Charter and Governance Documents - Director and Shareholder Affiliation Policy."

Moody's SF Japan K.K., Moody's Local AR Agente de Calificación de Riesgo S.A., Moody's Local BR Agência de Classificação de Risco LTDA, Moody's Local MX S.A. de C.V, I.C.V., Moody's Local PE Clasificadora de Riesgo S.A., Moody's Local PA Calificadora de Riesgo S.A., Moody's Local CR Calificadora de Riesgo S.A., Moody's Local ES S.A. de CV Clasificadora de Riesgo, Moody's Local RD Sociedad Calificadora de Riesgo S.R.L. and Moody's Local GT S.A.(collectively, the "Moody's Non-NRSRO CRAs") are all indirectly wholly-owned credit rating agency subsidiaries of MCO. None of the Moody's Non-NRSRO CRAs is a Nationally Recognized Statistical Rating Organization.

Additional terms for Australia only: Any publication into Australia of this document is pursuant to the Australian Financial Services License of MOODY'S affiliate, Moody's Investors Service Pty Limited ABN 61 003 399 657AFSL 336969 and/or Moody's Analytics Australia Pty Ltd ABN 94 105 136 972 AFSL 383569 (as applicable). This document is intended to be provided only to "wholesale clients" within the meaning of section 761G of the Corporations Act 2001. By continuing to access this document from within Australia, you represent to MOODY'S that you are, or are accessing the document as a representative of, a "wholesale client" and that neither you nor the entity you represent will directly or indirectly disseminate this document or its contents to "retail clients" within the meaning of section 761G of the Corporations Act 2001. MOODY'S credit rating is an opinion as to the creditworthiness of a debt obligation of the issuer, not on the equity securities of the issuer or any form of security that is available to retail investors.

Additional terms for India only: Moody's credit ratings, Assessments, other opinions and Materials are not intended to be and shall not be relied upon or used by any users located in India in relation to securities listed or proposed to be listed on Indian stock exchanges.

Additional terms with respect to Second Party Opinions and Net Zero Assessments (as defined in Moody's Ratings Rating Symbols and Definitions): Please note that neither a Second Party Opinion ("SPO") nor a Net Zero Assessment ("NZA") is a "credit rating". The issuance of SPOs and NZAs is not a regulated activity in many jurisdictions, including Singapore. EU: In the European Union, each of Moody's Deutschland GmbH and Moody's France SAS provide services as an external reviewer in accordance with the applicable requirements of the EU Green Bond Regulation. JAPAN: In Japan, development and provision of SPOs and NZAs fall under the category of "Ancillary Businesses", not "Credit Rating Business", and are not subject to the regulations applicable to "Credit Rating Business" under the Financial Instruments and Exchange Act of Japan and its relevant regulation. PRC: Any SPO: (1) does not constitute a PRC Green Bond Assessment as defined under any relevant PRC laws or regulations; (2) cannot be included in any registration statement, offering circular, prospectus or any other documents submitted to the PRC regulatory authorities or otherwise used to satisfy any PRC regulatory disclosure requirement; and (3) cannot be used within the PRC for any regulatory purpose or for any other purpose which is not permitted under relevant PRC laws or regulations. For the purposes of this disclaimer, "PRC" refers to the mainland of the People's Republic of China, excluding Hong Kong, Macau and Taiwan.