MatchPoint Research Working Paper No. 2026/T05

LLM Agents in Corporate Finance: Automating the Deal Workflow

A controlled operating model for evidence, tools, deterministic analysis, human authority and measurable release

Chennakeshav (CK) Adya
Independent Researcher
July 2026

ABSTRACT

Background. Corporate-finance execution spans fragmented documents, systems, calculations, communications and approval forums. Language-model agents can coordinate multi-step work, while current workplace benchmarks show material reliability limits. **Objective.** This paper develops a controlled operating model for agents across the deal workflow. **Approach.** The analysis reviews 18 primary, authoritative, research and provider sources. Survey and benchmark results retain their sample, domain and publication boundaries. **Findings.** In a 2024 Bank of England and FCA survey of 118 firms, 75% reported current AI use and only 2% of use cases had fully autonomous decision-making. TheAgentCompany reported 24% autonomous task completion in its simulated workplace, while the June 2026 OSWorld 2.0 preprint reported 20.6% completion under its primary long-horizon metric. The proposed architecture decomposes the deal into bounded work packets, routes tools through least-privilege policy, retains deterministic calculations, records evidence and exceptions, and requires named approval for material or external actions. **Implications.** Autonomy should expand by task-specific evidence. Acceptance, citation validity, calculation fidelity, tool precision, reviewer time, rework, elapsed time, cost and incidents form the release record.

Highlights

- The work packet is the operating unit of controlled agent delegation.
- Agents coordinate retrieval, comparison, drafting, routing and exception handling.
- Financial calculations remain deterministic and externally released actions remain approval-bound.
- Every material output carries evidence, tool, model, validation and authority records.
- Productivity and autonomy require a defined baseline and observed task-specific evidence.

JEL Classification: G24, G32, G34, M15, O33

Keywords: artificial intelligence, LLM agents, corporate finance, investment banking, M&A, fund management, family offices, workflow automation, human oversight, model risk, transaction execution

1. INTRODUCTION

Corporate-finance execution is a chain of document, data, judgement and communication tasks. A transaction team identifies an opportunity, qualifies it, agrees a mandate, assembles an evidence base, builds valuation and financing analyses, prepares investor materials, manages outreach, responds to diligence, negotiates terms and preserves a defensible record. Work moves repeatedly among email, customer relationship management systems, data rooms, spreadsheets, presentations, market databases and approval forums. Each hand-off can create delay, version ambiguity or a missing decision record.

Large-language-model agents add a new execution layer. In this paper, an agent is a software system in which a language model selects and sequences tools to pursue a defined goal within explicit instructions, permissions and exit conditions [1]. This definition excludes a single prompt that produces text and excludes deterministic software that follows a fixed rule sequence. It also separates agent autonomy from authority. An agent can plan and execute multiple steps; a named professional retains authority for material commercial, financial, legal, regulatory and external-communication decisions.

The opportunity is practical. Many deal tasks depend on unstructured documents, changing evidence and exceptions that are expensive to encode as traditional rules. The control problem is equally practical. A plausible paragraph can carry an unsupported assertion. A correct tool call can act on the wrong entity. A long workflow can lose an earlier constraint. A third-party model can change without a transaction team changing its internal procedure. Corporate finance compounds these exposures because confidentiality, valuation, suitability, conflicts, recordkeeping and negotiation positions are material.

This paper asks four questions:

- Which corporate-finance tasks are suitable for agentic execution?
- Which decisions must remain behind named human approval gates?
- What architecture connects agents to transaction systems while retaining evidence, permissions and reproducibility?
- How should a firm measure productivity, quality, risk and economic value before expanding autonomy?

The principal conclusion is that the deal workflow should be decomposed into bounded work packets with acceptance tests. Agents can coordinate evidence retrieval, comparison, drafting, workflow administration and exception routing. Deterministic services should perform calculations, reconciliations and rules-based checks. Human professionals should approve mandate acceptance, valuation judgements, external claims, recipients, commercial terms, regulated communications and release. This three-part structure supports useful automation while preserving accountable deal judgement.

The paper addresses advisers, family-office investment teams, fund managers, sponsors and transaction teams. It is a design and evaluation framework. It does not report an observed Matchpoint deployment, observed time saving or realised investment return.

2. DEFINITIONS, SCOPE AND METHOD

2.1 Workflow, agent and authority

A **workflow** is the ordered set of activities, inputs, decisions, outputs and control points required to achieve a business objective. A **language-model application** may classify, extract or draft within one step. An **agent** uses a language model to control workflow execution, choose among approved tools, assess progress and stop or hand control to a human under defined conditions [1]. A **deterministic service** produces the same result from the same inputs under a specified version, subject to the behaviour of its dependencies. An **approval gate** is a state transition that requires an authorised person or control function.

Agentic capability has several dimensions:

- planning depth; the number and dependency of steps the system can organise;

- tool breadth; the data and action systems available to it;
- state duration; how long it retains and reconciles workflow state;
- action reversibility; whether an action can be withdrawn without material consequence;
- decision materiality; the possible financial, legal, regulatory or reputational effect;
- externality; whether the action changes a third-party system or communicates outside the controlled workspace.

These dimensions should be assessed separately. An agent may have broad retrieval access and narrow action authority. Another agent may act autonomously on low-risk internal administration while requiring approval for every external message.

2.2 Corporate-finance scope

The operating model covers the following transaction phases:

- origination and opportunity monitoring;
- qualification, conflicts and mandate screening;
- information request and data-room administration;
- commercial, financial and technical analysis;
- valuation, capital-structure and scenario analysis;
- preparation of teasers, information memoranda, presentations and committee papers;
- investor or counterparty mapping and outreach preparation;
- question-and-answer management and diligence follow-up;
- term-sheet comparison, negotiation support and closing administration;
- record retention, post-mortem and reusable institutional knowledge.

Legal advice, tax advice, accounting opinions, regulatory determinations, investment decisions and binding communications remain outside agent authority unless an applicable professional separately reviews and authorises the output.

2.3 Evidence method

The analysis uses four evidence classes.

Authoritative public guidance. NIST, the Financial Stability Board, the Central Bank of the UAE, the Bank of England, the Financial Conduct Authority, FINRA, official UAE data-protection material and the EU AI Act supply governance, risk, data, oversight and regulatory context [7-16].

Primary research. TheAgentCompany, OSWorld 2.0 and FinRobot papers supply benchmark or design evidence. TheAgentCompany evaluated agents in a simulated software-company environment; it did not evaluate an investment bank [3]. OSWorld 2.0 is a June 2026 preprint covering long-horizon computer-use tasks; it is current research evidence rather than a finance-specific production study [4]. FinRobot papers describe research architectures and prototypes; they do not establish production outcomes for Matchpoint or any named transaction firm [5,6].

Provider guidance. OpenAI and Anthropic material describes implementation patterns observed or recommended by the providers. It is useful practitioner guidance and is not independent comparative evidence [1,2].

Matchpoint synthesis and scenario analysis. Workflow maps, control gates, implementation stages and economic formulas are the author's synthesis. Any numerical operating scenario is labelled as illustrative and requires validation against an approved baseline.

Evidence class	Appropriate use	Boundary
Regulatory or standards guidance	Governance, data, oversight and control objectives	Application depends on entity, activity and jurisdiction
Survey evidence	Adoption and respondent-reported practices	Sample, date, geography and self-report boundary retained
Agent benchmark	Capability and failure-mode evidence	Benchmark environment differs from corporate-finance production
Research prototype	Architecture and testable design propositions	Prototype results do not prove firm-level productivity
Provider guidance	Engineering patterns and terminology	Vendor-authored; no independent performance comparison
Matchpoint scenario	Pilot design and economic sensitivity	Illustrative assumptions; no observed benefit claim

3. EVIDENCE BASE: ADOPTION, AUTONOMY AND RELIABILITY

3.1 Financial-sector adoption

The Bank of England and FCA received 118 responses to their 2024 survey of regulated financial-services firms. Seventy-five per cent of respondents reported current AI use and a further 10% planned use within three years. Foundation models represented 17% of reported AI use cases. The survey covered AI broadly and should not be read as an agent-only adoption rate [7].

The same survey provides a useful autonomy boundary. Respondents reported some degree of automated decision-making in 55% of AI use cases; only 2% of use cases had fully autonomous decision-making. One third of use cases were third-party implementations. Forty-six per cent of respondent firms reported only partial understanding of the AI technologies they used, compared with 34% reporting complete understanding. These results support a design focus on explicit accountability, third-party visibility and bounded autonomy [7].

The June 2026 FSB consultation addresses traditional AI, generative AI and agentic AI. It proposes 12 sound practices spanning strategic oversight, governance, accountability, adaptability, materiality assessment, model selection, data governance, explainability, performance management, human oversight, cyber and ICT risk, and third-party risk. The report is a consultation document; it is not an international standard and does not create binding obligations [8].

The CBUAE issued a February 2026 guidance note for licensed financial institutions concerning responsible AI and machine-learning use where consumers may be affected. It calls for documented governance, board and senior-management accountability, model inventories, data provenance, privacy and security controls, testing, meaningful human oversight, third-party due diligence, audit rights and an immediate ability for human intervention to cease use of an AI system [10]. The guidance applies to licensed financial institutions within its scope. A corporate-finance adviser or family office should confirm its own regulatory perimeter.

3.2 Agent benchmarks

TheAgentCompany created a self-contained simulated software company and tasks involving web browsing, code, programs and co-worker communication. Its reported leading baseline completed 24% of tasks autonomously. The result shows that useful workplace tasks can be completed while difficult long-horizon work remains unreliable. The domain and model vintage limit direct transfer to finance [3].

OSWorld 2.0 introduced 108 long-horizon real-world computer-use workflows. The authors report that the best evaluated configuration completed 20.6% of tasks under their primary binary metric at 500 steps, with a 54.8% partial score. The paper attributes failures to lost constraints, missed changing information, guesses in

place of clarification and skipped verification. It is a June 2026 preprint and should be treated as current research evidence subject to peer-review and benchmark revision [4].

These benchmarks do not measure a controlled corporate-finance agent using narrow tools, deterministic validators and mandatory approvals. They do establish a prudent premise: end-to-end autonomy should be earned through task-specific evidence. A model's fluent output and a benchmark score on another domain do not establish transaction reliability.

3.3 Finance-specific research architectures

FinRobot describes a layered platform for financially specialised agents, model strategies, LLMOps and DataOps, and multiple foundation models [5]. A later FinRobot paper describes specialised agents for data, concepts and thesis formation in equity research and valuation [17]. A 2025 paper extends the architecture to finance processes in enterprise-resource-planning settings, including budget planning, reporting and payment-related workflows [6]. These papers support modular decomposition and financial tool integration. They remain research systems; the paper's control model therefore adds explicit authority, evidence, release and incident records.

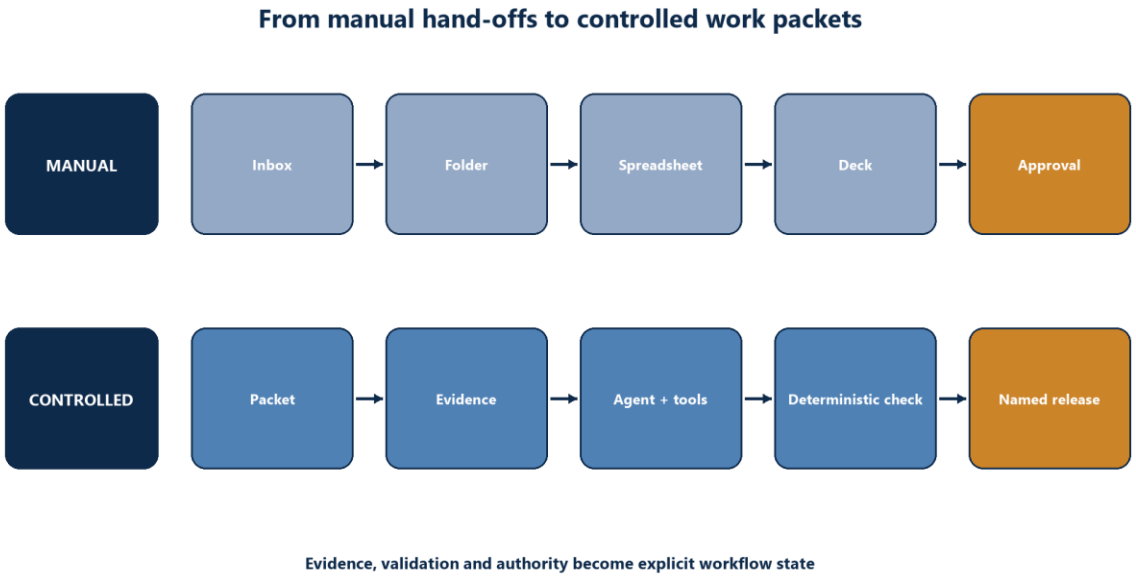

Figure 1. Before-and-after workflow map from manual hand-offs to controlled work packets and approval gates
Source: Matchpoint framework.

4. THE DEAL WORKFLOW AS CONTROLLED WORK PACKETS

4.1 From documents to work packets

A document-centred process asks an analyst to read an inbox, open a folder and decide what to do. An agent-ready process represents work as packets with typed inputs, authorised tools, an acceptance test, a reviewer and an expiry condition. The packet becomes the unit of delegation and audit.

A work packet should contain:

- transaction and entity identifiers;
- objective and definition of done;
- source manifest and permitted data boundary;
- approved tools and maximum permissions;

- required deterministic checks;
- materiality and risk tier;
- external-action prohibition or approval condition;
- evidence and citation requirements;
- named reviewer and escalation path;
- time, cost and tool-call budget;
- stop conditions and expiry timestamp;
- output schema and retention classification.

This structure reduces ambiguity. It also permits measurement because every packet has a start state, acceptance criteria and final disposition.

4.2 Phase map

Deal phase	Bounded agent contribution	Deterministic service	Named human authority
Opportunity monitoring	Retrieve approved sources; classify relevance; draft evidence-linked brief	Deduplication; entity matching; date checks	Origination owner decides whether to pursue
Qualification	Assemble company, sponsor, sector and transaction facts; identify missing inputs	Eligibility rules; conflicts lookup; sanctions and KYC service calls	Partner accepts, rejects or escalates
Mandate preparation	Draft scope, workplan and information request from approved templates	Fee arithmetic; version comparison; required-clause check	Partner and counsel approve terms
Data room	Index files; detect versions; extract fields; map questions to evidence	Hashing; completeness reconciliation; spreadsheet tests	Workstream lead accepts evidence status
Analysis	Compare operating, market and financial evidence; draft exception notes	Valuation calculations; model checks; reconciliations	Analyst and senior reviewer approve conclusions
Materials	Assemble cited drafts, charts and disclosure register	Numbers-to-model tie-out; style and cross-reference checks	Deal lead releases each external document
Counterparty mapping	Build evidence-linked candidate list; record rationale and conflicts flags	Entity resolution; exclusion rules; duplicate removal	Coverage owner approves inclusion and priority
Outreach	Prepare personalisation and response options	Recipient, consent, channel and timing checks	Authorised person approves recipient and message
Diligence Q&A	Classify questions; retrieve evidence; draft answers; track open items	Source-link validation; answer-status reconciliation	Workstream owner approves every external answer
Terms and closing	Compare clauses and economics; maintain issue list; draft decision brief	Cash-flow and ownership calculations; version diff	Partner, client and counsel decide and execute
Archive and learning	Assemble complete record; identify reusable patterns and evaluation cases	Retention rules; checksum; access revocation	Records owner closes and approves reuse boundary

4.3 The workflow state machine

Each packet moves through explicit states: **created, authorised, executing, awaiting evidence, awaiting approval, released, rejected, expired or incident**. The state machine matters because a language model

should not infer authority from conversational context. A change from internal draft to external release requires a recorded approval event. A failed verification changes the packet to an exception state. An expired market datum or recipient list returns the packet to review.

The agent's completion claim should be tested against the system state. A materials packet is complete when every referenced number reconciles, every claim has a source, every required section exists and the authorised reviewer records release. A polite final message from the model is not a completion criterion.

5. CONTROLLED AGENT ARCHITECTURE

5.1 Layers

The proposed architecture has seven layers.

1. Transaction identity and data plane. Every transaction, entity, document, model, message and decision uses stable identifiers. Original evidence is retained with source, version, timestamp, permissions and hash where applicable.

2. Work queue and policy engine. The queue stores packet status. The policy engine decides which tools and data classes are available, whether an action is reversible, which approvals are required and when execution must stop.

3. Agent orchestration. A coordinator decomposes the authorised packet. Specialised workers retrieve, compare, draft or reconcile within narrower instructions. Provider guidance from OpenAI and Anthropic recommends beginning with the simplest architecture that performs adequately and adding multi-agent complexity only when evaluation shows a need [1,2].

4. Tool gateway. Tools expose approved functions for document retrieval, CRM records, data-room indexes, spreadsheets, market data, email drafts and task management. Each function has a typed schema, least-privilege credential and test suite. The gateway blocks unapproved recipients, destructive operations and direct external release.

5. Deterministic analysis boundary. Valuation, cap-table, waterfall, interest, covenant, ownership, exchange-rate and reconciliation logic runs in versioned code or controlled spreadsheets. The agent can select an approved calculation, supply typed inputs and explain results. It cannot substitute generated arithmetic for the calculation record.

6. Evidence and observability layer. The system records sources, tool calls, prompts or instruction versions, model version, outputs, validation results, cost, latency, exceptions, approvals and released artefacts. Sensitive reasoning traces should not be treated as a required audit object. Decision-relevant inputs, actions and outputs provide the operational record.

7. Human authority and release. Named people approve material state transitions. Interfaces present the evidence, calculation, exception and delta first; generated prose follows. Reviewers can reject, amend, request evidence, narrow scope or stop the run.

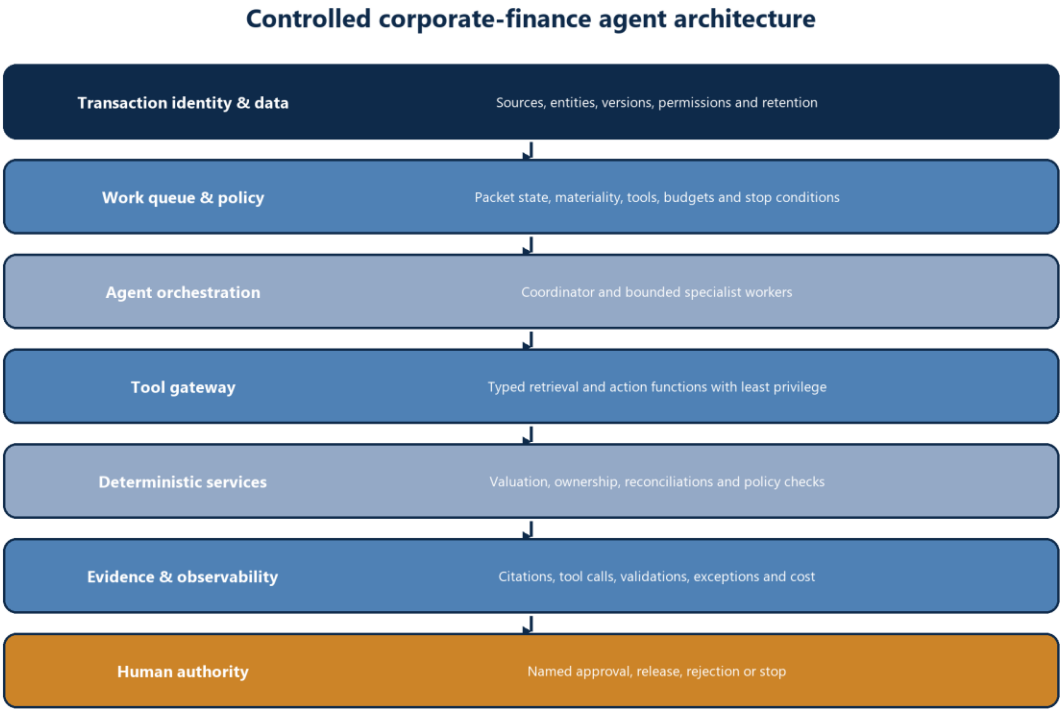

Figure 2. Controlled tool-stack architecture for transaction identity, policy, agents, deterministic services, evidence and authority

Source: Matchpoint architecture; agent patterns informed by [1,2] and control objectives by [8-11].

5.2 Single-agent and multi-agent choices

A single agent with several tools is easier to evaluate and operate. It should be the default for a bounded task such as compiling a meeting brief or reconciling an information request. Multiple agents are appropriate when instructions or tool sets conflict, specialist evaluation differs, or work can be independently cross-checked. Examples include separate evidence-retrieval, financial-analysis and release-review agents.

Agent count should not be used as a sophistication metric. Every additional hand-off introduces state, provenance and failure modes. The decision should follow evaluation evidence:

- Can one agent select the correct tools consistently?
- Do instructions exceed a manageable decision space?
- Do workstreams require different credentials or data boundaries?
- Can independent workers produce outputs that are reconciled deterministically?
- Does the added orchestration improve task acceptance after cost and latency are included?

5.3 Memory and transaction state

The system should distinguish four forms of state:

- **packet state**, which records the current task and authority;
- **transaction state**, which records approved facts, open issues and decisions for the deal;
- **institutional knowledge**, which contains approved reusable playbooks and precedent;
- **model context**, which is the temporary information supplied to a particular run.

Only approved records should become institutional knowledge. Drafts, privileged material and client-specific information require retention and reuse rules. Memory should cite its origin and effective date. A stale precedent should surface as stale rather than silently shape a live term or valuation.

6. SPECIALISED AGENTS ACROSS THE DEAL CYCLE

6.1 Origination and qualification agent

The origination agent monitors approved public and licensed sources, links signals to existing entities, prepares a dated opportunity brief and identifies missing qualification facts. It can draft CRM updates. It should not invent revenue, funding need, ownership, affordability or decision-maker status. Each field requires a source or an explicit unknown status.

Qualification combines judgement with rules. Deterministic checks can test mandate size, geography, transaction type, sector exclusions and known conflicts. The agent can assemble evidence and propose questions. A partner decides whether the opportunity merits contact, what claim can be made and which fee approach is appropriate.

6.2 Mandate and workplan agent

This agent converts an approved opportunity into a draft scope, responsibility matrix, information request and delivery plan. It uses clause and template libraries selected for jurisdiction and transaction type. It can compare a counterparty draft against an approved base and create an issue list. Fee arithmetic, dates and required clauses are checked deterministically. Counsel and the responsible partner approve contractual language.

6.3 Evidence and diligence agent

The evidence agent ingests authorised data-room content, records file identity and version, extracts structured facts with source locations, maps evidence to an assertion register and opens exceptions for missing or contradictory information. It can group questions, draft evidence-linked responses and prepare reviewer packs.

The agent should abstain when the source does not support the requested conclusion. A contradiction is an output, not a drafting inconvenience. Material assertions such as ownership, debt, cash, customer concentration, forecast assumptions and legal rights receive named review.

6.4 Financial-analysis agent

The financial-analysis agent prepares and validates inputs, selects approved models, runs versioned calculations and explains sensitivities. It can reconcile financial statements, identify outliers and map a management forecast to historical drivers. It may propose scenarios. The scenario assumptions remain visible and approved.

Valuation remains a professional judgement supported by calculations. Comparable-company selection, precedent relevance, normalisation, discount rate, terminal assumptions, control, liquidity and transaction-specific risk require reviewer authority. The agent's role is to make the evidence and sensitivity structure legible.

6.5 Materials agent

The materials agent assembles teasers, information memoranda, presentations, committee papers and lender packs from an approved outline and evidence register. It creates a claim ledger linking each external assertion to its source, owner and review status. It checks consistency among prose, tables, charts and model outputs. It flags promotional language whose basis is missing.

Every external artefact has a release certificate listing the file hash, source cut-off, model version, unresolved limitations, approvals and intended audience. The SEC's AI-washing enforcement illustrates that claims about AI capability require a factual basis [15]. The same principle applies to claims about a target, fund, strategy or adviser capability.

6.6 Counterparty and outreach agent

The counterparty agent builds an evidence-linked target list from approved data. It records investment mandate, sector, geography, ticket, stage, recent activity, conflicts and the source date. Unknown values remain unknown. The outreach agent can draft personalisation and follow-up options.

Recipient selection and message release remain approved actions. The gateway checks identity, consent or lawful basis where applicable, channel restrictions, confidentiality, conflicts, prior contact and do-not-contact status. Messages are retained in approved systems. FINRA's technology-neutral reminder and SEC recordkeeping actions illustrate that existing supervision and record obligations remain relevant when new tools are used [13,14]. Applicability requires jurisdiction-specific advice.

6.7 Q&A, negotiation and closing agent

The Q&A agent classifies incoming questions, maps them to evidence and owners, drafts answers, records approvals and reconciles open items. The negotiation agent compares term-sheet and document versions, builds an issue tree and calculates economic effects through approved models. It can prepare decision briefs that separate fact, assumption, alternative and unresolved issue.

The agent does not accept terms, waive rights, make representations or transmit execution copies. Those actions require named authority. Closing administration can be more autonomous where it concerns reversible reminders, checklist updates and collection of already authorised documents.

7. AUTONOMY AND DECISION RIGHTS

7.1 A five-tier authority ladder

Tier	Agent permission	Typical corporate-finance example	Release condition
0: Observe	Read approved sources and produce an internal trace	Index a data room; classify email; benchmark current process	No external action; reviewer sees output
1: Recommend	Draft or rank options without changing systems	Draft opportunity brief; propose diligence questions	Professional approves any use
2: Prepare	Write to controlled draft systems; run approved calculations	Create CRM draft; populate model inputs; assemble deck draft	Deterministic checks and named review
3: Act reversibly	Execute approved low-risk actions with monitoring	Create internal task; request missing document from an approved internal owner	Policy permits action; human can reverse promptly
4: Act externally	Change an external system or communicate with a counterparty	Send outreach; publish materials; update a binding term process	Explicit per-action approval and complete release record

Tier 4 is an authority category rather than a maturity destination. Many transaction activities should remain approval-bound even after excellent task performance. The appropriate tier depends on materiality, reversibility, confidentiality, recipient impact and regulatory scope.

7.2 Materiality and reversibility

An autonomy decision should combine at least five factors: financial impact, legal or regulatory effect, confidentiality, reversibility and external reach. A simple internal reminder may receive Tier 3 permission. A short external email may remain Tier 4 because the recipient, timing and statement can affect a live process.

The CBUAE guidance describes human-in-the-loop, human-on-the-loop and human-out-of-the-loop models and links oversight to risk. It states that human-out-of-the-loop use should be confined to low-risk, non-

material processes with appropriate controls within its scope [10]. The FSB consultation similarly links human oversight to materiality, risk, autonomy, complexity and explainability [8].

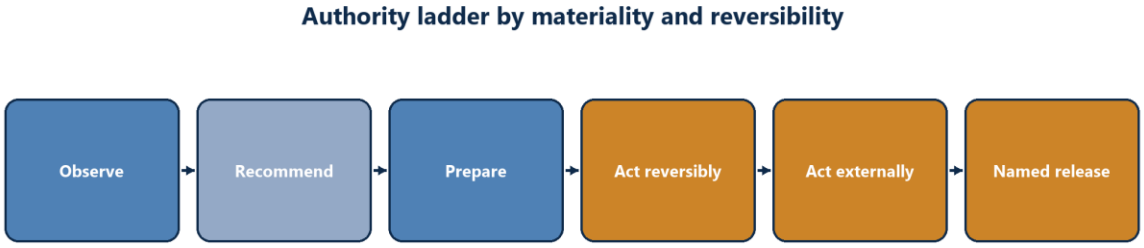

Figure 3. Authority ladder from observation to external action and named release
Source: Matchpoint framework; oversight concepts informed by [8,10].

8. RISK AND CONTROL FRAMEWORK

8.1 Information accuracy and provenance

Every material claim should carry source identity, location, version, date and extraction method. Generated summaries should preserve limitations and disagreements. Retrieval tests should measure whether the correct source passage was found; generation tests should measure whether the output is supported by that passage. A complete citation does not cure an incomplete source population.

Controls include source manifests, evidence cut-off dates, duplicate and version detection, citation validators, claim-to-source sampling, contradiction queues and reviewer abstention options. NIST's Generative AI Profile calls for risk management across governance, measurement and management activities, including attention to confabulation, information integrity, privacy, security and third-party components [9].

8.2 Calculation and model risk

Financial arithmetic should be removed from free-form generation. Typed inputs enter a versioned calculation. Outputs include units, currency, dates, formula version and reconciliation status. The agent explains and compares results after the calculation succeeds.

Model controls include approved templates, locked formula regions, input validation, independent recomputation, balance and cash-flow checks, scenario labels, review thresholds and model-change records. An agent may detect a possible error; a passed deterministic test provides the release evidence.

8.3 Confidentiality, privacy and privilege

Tool permissions follow least privilege, transaction segregation and purpose limitation. Data-class policies determine whether material may be sent to a model provider, processed in a particular region, retained, used for training, included in logs or reused. Prompts and outputs can contain personal or confidential data and require the same classification discipline as source files.

The UAE Personal Data Protection Law provides a federal framework for personal-data processing, rights and cross-border transfers [12]. DIFC, ADGM, UK, EU and other regimes may apply depending on entity, establishment, data subject and processing. A data-protection and legal assessment should precede production use.

8.4 Prompt injection and tool abuse

Transaction documents and web pages are untrusted inputs. Text embedded in a file can attempt to change an agent's instructions or induce a tool call. The ingestion layer should separate content from control instructions, screen active content, restrict tools by packet policy and require approvals independent of document text.

Action tools need allow-lists, parameter validation, recipient controls, idempotency where possible, rate limits and dry-run modes. Destructive actions should be absent from ordinary agent credentials. NIST SP 800-218A extends secure-development practices to generative-AI and dual-use foundation-model systems, including risk-oriented development, provenance and secure lifecycle considerations [11].

8.5 Third-party and concentration risk

Vendor assessment covers model and service changes, training and retention terms, location, sub-processors, uptime, incident notification, audit rights, evaluation access, portability and termination. A firm should know which workflow depends on which provider and what happens when that provider is unavailable or materially changed.

The Bank of England/FCA survey reported that one third of AI use cases were third-party implementations; the top three named model providers represented 44% of named model providers [7]. The FSB consultation includes third-party performance, transparency, data quality, supply-chain, concentration and business-continuity considerations [8]. These findings support an inventory and an exit plan.

8.6 Supervision, communications and records

Agent activity should occur through approved systems. External communications require recipient, content and channel approval. The record should include the released version and approval, rather than only the model draft. FINRA states that its existing technology-neutral rules and securities laws continue to apply when member firms use generative AI [13]. The SEC's 2024 recordkeeping actions concerned widespread failures to preserve required electronic communications [14]. Entity-specific obligations require legal and compliance advice.

Figure 4. Risk-control map for evidence, calculations, data, security, third parties and authority

Source: Matchpoint synthesis based on [7-11,13,14].

9. EVALUATION AND RELEASE ENGINEERING

9.1 Evaluation unit

The evaluation unit should be the work packet. A task passes when its acceptance criteria pass, its prohibited actions remain absent and required approvals are present. Evaluation sets should use representative transaction artefacts with confidential information removed or appropriately controlled.

Four test levels are required:

- **Component tests:** retrieval, extraction, entity matching, calculation input, tool schema and policy checks.
- **Packet tests:** complete a bounded task with known acceptance criteria and hidden edge cases.
- **Workflow tests:** coordinate several packets across changing state, exceptions and approvals.
- **Production monitoring:** compare shadow or controlled-live outputs with the current human baseline and investigate drift.

9.2 Metric dictionary

Metric	Definition	Why it matters
Task acceptance	Packets passing all required criteria divided by completed packets	Measures usable completion
Citation validity	Supported sampled claims divided by sampled claims	Measures grounding
Source coverage	Required source classes represented divided by required classes	Detects narrow evidence
Calculation fidelity	Outputs matching approved deterministic result	Measures numerical integrity
Tool precision	Correct authorised tool calls divided by all tool calls	Detects unsafe or wasteful action
Exception recall	Known material exceptions surfaced divided by known exceptions	Measures risk detection
External-action precision	Approved external actions divided by attempted external actions	Measures authority control
Reviewer time	Median active review minutes per accepted packet	Measures human capacity effect
Rework	Packets reopened after release divided by released packets	Measures downstream quality
Elapsed time	Median authorised-start to accepted-release time	Measures process speed
Cost per accepted packet	Model, tool and review cost divided by accepted packets	Measures economics
Incident rate	Material incidents per defined packet volume	Measures operational risk

Success-rate metrics should be accompanied by sample size, task mix and confidence intervals where practical. Average performance can conceal a material failure class. A release gate may require zero unauthorised external actions and zero calculation mismatches even when other quality metrics use thresholds.

9.3 Baseline and experiment design

The current workflow should be observed before automation. Baseline data includes active analyst time, senior-review time, elapsed time, rework, error and exception rates, systems touched and output acceptance. Historical estimates should be labelled as estimates if direct observation is unavailable.

A controlled pilot runs the agent in shadow mode against the same packet. Reviewers score both outputs using the same rubric without relying on the agent's self-assessment. Difficult cases, negative cases and changing

information belong in the test set. Repeated runs measure variance. Provider, model, prompt, tool and policy versions are retained.

Expansion follows evidence. A task can move from observe to recommend, prepare and reversible action after its thresholds pass and control owners approve the new tier. Model updates return affected tasks to regression testing.

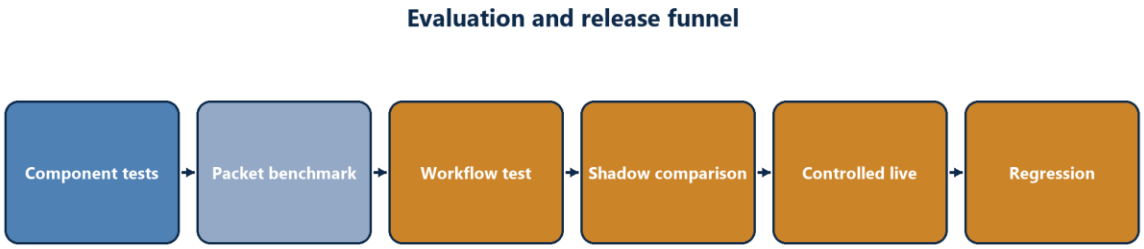

Figure 5. Evaluation funnel from component tests through controlled production and regression monitoring
Source: Matchpoint framework; benchmark limitations informed by [3,4].

10. PRODUCTIVITY AND ECONOMIC CASE

10.1 Value equation

The economic case should be calculated at accepted-output level:

Net annual value = released human capacity value + avoided rework and external cost + incremental contribution from additional accepted work - model and tool cost - build and control cost - additional review and incident cost.

Released capacity is valuable only when it can be redeployed to higher-value work, used to increase accepted throughput or removed from cost. A reduction in drafting minutes with unchanged reviewer time may have limited economic value. A faster first draft that creates more rework can have negative value.

10.2 Illustrative scenario

The following scenario is unverified and uses management assumptions solely to show the calculation structure.

Input	Conservative	Reference	Expansion	Status
Eligible packets per year	1,200	2,400	4,000	Illustrative management assumption
Baseline active minutes per packet	45	60	75	Illustrative management assumption
Accepted active minutes after deployment	36	39	41	Illustrative management assumption
Loaded blended hourly cost	USD 100	USD 140	USD 180	Illustrative management assumption
Model and tool cost per attempted packet	USD 3	USD 5	USD 8	Illustrative management assumption
First-year build and control cost	USD 180,000	USD 300,000	USD 500,000	Illustrative management assumption
Acceptance rate	70%	82%	90%	Illustrative management assumption

The table does not establish a business case. Actual inputs require observed packet volumes, times, acceptance, reviewer effort, costs and incident performance. The model should include downside cases for lower acceptance, higher review, provider price changes, rework and delayed deployment.

10.3 Productivity tree

Productivity can arise through six mechanisms:

- fewer search and hand-off minutes;
- greater reuse of approved facts and calculations;
- reduced version and reconciliation work;
- earlier exception detection;
- shorter elapsed time while work waits for routing;
- additional accepted throughput from the same authorised team.

Each mechanism needs a metric and owner. Revenue uplift should be recorded only when an accepted opportunity, mandate or closing can be attributed through the firm's normal commercial records. Activity volume is not revenue evidence.

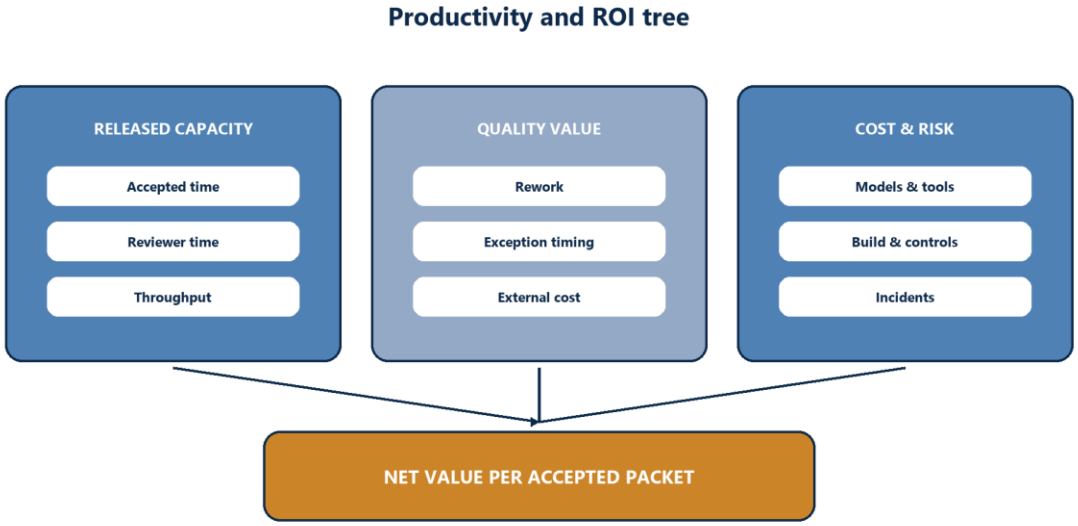

Figure 6. Productivity and ROI tree for accepted capacity, quality value, cost and risk
Source: Matchpoint scenario analysis. All operating inputs require observed baseline and pilot evidence.

11. IMPLEMENTATION ROADMAP

11.1 Stage 1: mandate and inventory

Appoint an accountable sponsor, transaction owner, risk owner and technical owner. Inventory candidate workflows, systems, data classes, providers, regulatory perimeter and existing controls. Select a low-materiality packet with meaningful unstructured work and clear acceptance criteria.

Exit gate: documented purpose, authority, data boundary, baseline plan, risk tier and stop conditions.

11.2 Stage 2: evidence and deterministic foundation

Create stable transaction identifiers, source manifests, approved calculation services, template libraries and a release record. Define the policy engine and remove unnecessary action permissions.

Exit gate: representative packet can be completed manually using the same structured evidence and acceptance criteria.

11.3 Stage 3: observe and recommend

Run the agent in a controlled environment. Test components, adversarial inputs, negative cases and failure recovery. Compare outputs with the baseline. Record reviewer amendments and convert them into evaluation cases where appropriate.

Exit gate: agreed acceptance, citation, calculation, exception, cost and unauthorised-action thresholds pass on a representative sample.

11.4 Stage 4: prepare and shadow

Allow writes to controlled draft systems and approved calculations. Maintain human release. Test provider outage, credential revocation, stale data, prompt injection, duplicate messages and interrupted workflows.

Exit gate: operational resilience, rollback, incident handling and full release record pass.

11.5 Stage 5: controlled production

Deploy to a limited user and transaction population. Monitor every packet, review incidents and compare accepted output with the baseline. Keep model and policy changes behind regression testing.

Exit gate: the control committee approves expansion based on observed performance and economic evidence.

11.6 Stage 6: portfolio learning

Add work packets or autonomy one class at a time. Use shared evidence, tool and control services. Maintain separate evaluation thresholds for each task and materiality tier. Review third-party concentration, data reuse and regulatory change at defined intervals.

Timing, staffing and thresholds require an approved implementation plan. The roadmap supplies sequence and gates rather than a universal duration.

12. APPLICATION TO FAMILY OFFICES AND FUND MANAGERS

12.1 Family-office CIO and alternatives team

A family-office team can use bounded agents to monitor portfolio-company reporting, assemble investment-committee packs, reconcile capital calls, compare manager communications, prepare diligence trackers and maintain decision records. Portfolio allocation, manager selection, conflicts, liquidity decisions and investment approval remain authorised judgements.

The highest-value starting packet is often evidence-heavy and internally consumed. Examples include a quarterly manager brief with source-linked performance, exposure and open-action reconciliation; or a co-investment diligence register that maps every assertion to a data-room source and owner. These tasks have clear acceptance criteria and avoid direct trading authority.

12.2 Fund manager or GP raising capital

A fund manager can use agents to maintain a data-room index, prepare a claim ledger, reconcile track-record presentations, draft due-diligence questionnaires, classify LP questions and assemble evidence-linked responses. An agent can also prepare a prospect list using approved mandate data and identify information gaps.

Fund terms, performance presentation, track-record attribution, suitability, target-list inclusion and every external communication require authorised review. Claims about AI-assisted processes also require a factual

basis. The operating record should show what the system actually does, where humans decide and how outputs are tested.

12.3 Shared service opportunity

Both ICPs benefit from a shared transaction knowledge layer: entities, documents, sources, calculations, decisions, approvals and communications. The tool and control foundation can serve multiple packets while permissions keep mandates separated. This shared service creates scale through reused controls and test infrastructure, while each workflow retains its own evidence thresholds.

13. LIMITATIONS AND RESEARCH AGENDA

This paper has six principal limitations.

First, available workplace-agent benchmarks are not corporate-finance production studies. TheAgentCompany uses a simulated software firm, and OSWorld 2.0 covers general computer-use workflows [3,4]. Their results support caution and test design; they do not forecast deal-team performance.

Second, finance-agent papers largely describe research architectures and demonstrations [5,6,17]. Production data on confidentiality, supervision, review effort, accepted throughput and incident rates remains limited in the public evidence reviewed.

Third, model, tool and benchmark performance changes quickly. All deployment decisions require current evaluations using the chosen model, tool versions, data boundary and task population.

Fourth, the regulatory discussion spans several jurisdictions and entity types. Applicability depends on legal status, activity, client, data and location. This paper is not a regulatory opinion.

Fifth, the economic scenario uses unverified management assumptions. It is a calculation template and does not support a productivity or return claim.

Sixth, successful task completion does not capture every form of risk. A workflow can produce a correct artefact while using an unauthorised source, disclosing confidential data or creating an incomplete record. Evaluation must score both output and process.

Future research should build a finance-specific agent benchmark using realistic, permissioned transaction artefacts and deterministic graders. It should measure long-horizon state, changing evidence, entity identity, calculation fidelity, citation validity, recipient controls, conflict escalation and human review. Multi-period field studies should compare accepted throughput, reviewer time, rework, incident rate and commercial attribution.

14. CONCLUSION

LLM agents can become a useful execution layer for corporate finance when the deal process is represented as bounded work packets with typed tools, deterministic calculations, evidence records and explicit authority. The strongest use cases coordinate unstructured evidence and repetitive workflow administration. Material judgement and external release remain named professional responsibilities.

The implementation sequence begins with identity, evidence, policy and measurement. A firm should establish its baseline, select a low-materiality packet, test component and workflow performance, run in shadow, observe accepted output and expand only through documented gates. The resulting operating model can support faster routing, broader evidence coverage and more consistent records. Actual benefits remain empirical questions for each workflow.

For family offices and fund managers, the practical destination is a controlled transaction workbench: agents prepare, retrieve, compare and reconcile; deterministic services calculate and validate; professionals decide, approve and communicate. That division of labour converts agent capability into accountable execution.

DECLARATIONS

Author. Chennakeshav (CK) Adya.

Funding. No external funding was received for this paper.

Conflicts of interest. The author is Managing Partner of Matchpoint Partners, which may advise clients on corporate-finance, transaction-execution and technology-enabled operating-model matters.

Data availability. The paper uses public sources listed below. The illustrative economic scenario is generated for framework demonstration and is not observed firm data.

Use of AI. AI tools supported research retrieval, drafting, consistency checking, document production and visual generation under human direction. The author remains responsible for the paper's conclusions.

REFERENCES

[1] OpenAI. (2025). *A practical guide to building agents*. https://openai.com/business/guides-and-resources/a-practical-guide-to-building-ai-agents/

[2] Anthropic. (2024). *Building effective agents*. https://www.anthropic.com/engineering/building-effective-agents

[3] Xu, F. F., Song, Y., Li, B., et al. (2025). TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2412.14161

[4] Yuan, M., Zhou, Z., Xiong, X., et al. (2026). OSWorld 2.0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*. https://arxiv.org/abs/2606.29537

[5] Yang, H., Zhang, B., Wang, N., et al. (2024). FinRobot: An open-source AI agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*. https://arxiv.org/abs/2405.14767

[6] Lin, L., et al. (2025). FinRobot: Generative business process AI agents for enterprise resource planning in finance. *arXiv preprint arXiv:2506.01423*. https://arxiv.org/abs/2506.01423

[7] Bank of England and Financial Conduct Authority. (2024). *Artificial intelligence in UK financial services: 2024*. Survey of 118 responding firms. https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024

[8] Financial Stability Board. (2026). *Sound practices for responsible adoption of artificial intelligence: Consultation report*. Consultation dated 10 June 2026. https://www.fsb.org/2026/06/sound-practices-for-responsible-adoption-of-artificial-intelligence-ai-consultation-report/

[9] National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[10] Central Bank of the UAE. (2026). *Guidance Note on the Consumer Protection and Responsible Adoption and Use of Artificial Intelligence and Machine Learning by Licensed Financial Institutions in the U.A.E.* Issued 11 February 2026. https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence

[11] National Institute of Standards and Technology. (2024). *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile*. NIST SP 800-218A. https://doi.org/10.6028/NIST.SP.800-218A

[12] UAE Government. (2021). *Federal Decree-Law No. 45 of 2021 Regarding the Protection of Personal Data*. Official portal summary. https://u.ae/en/about-the-uae/digital-uae/data/data-protection-laws

[13] Financial Industry Regulatory Authority. (2024). *Regulatory Notice 24-09: FINRA reminds members of regulatory obligations when using generative artificial intelligence and large language models*. https://www.finra.org/rules-guidance/notices/24-09

[14] U.S. Securities and Exchange Commission. (2024). *Twenty-six firms to pay more than $390 million combined to settle SEC's charges for widespread recordkeeping failures*. Press Release 2024-98. https://www.sec.gov/newsroom/press-releases/2024-98

[15] U.S. Securities and Exchange Commission. (2024). *SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence*. Press Release 2024-36. https://www.sec.gov/newsroom/press-releases/2024-36

[16] European Parliament and Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[17] Zhou, T., Wang, P., Wu, Y., and Yang, H. (2024). FinRobot: AI agent for equity research and valuation with large language models. *arXiv preprint arXiv:2411.08804*. https://arxiv.org/abs/2411.08804

[18] Financial Stability Board. (2024). *The financial stability implications of artificial intelligence*. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

APPENDIX A. WORK-PACKET RECORD

Field	Required content
Packet ID	Stable identifier linked to transaction and workstream
Objective	One observable business outcome
Definition of done	Testable acceptance criteria
Inputs	Authorised source manifest and effective date
Tools	Approved functions and permission scope
Materiality	Financial, legal, confidentiality and external-impact tier
Deterministic checks	Calculations, reconciliations and policy tests
Reviewer	Named person and required competence
Stop conditions	Missing evidence, contradiction, uncertainty, cost or time limit
Output	Typed schema, citation and retention class
Approval	Person, timestamp, decision and released artefact hash

APPENDIX B. APPROVAL MATRIX

Action	Agent may prepare	Agent may execute	Required authority
Internal source-linked brief	Yes	Save to controlled draft workspace	Workstream owner accepts
CRM factual draft	Yes	Save as draft or proposed update	Record owner approves material fields
Valuation calculation	Supply typed inputs to approved model	Run approved version	Analyst reviews; senior professional approves judgement
Investor or lender target	Rank with evidence	No external use without approval	Coverage owner approves inclusion and priority
External email	Draft	Only after per-message approval	Authorised sender approves recipient and content
Information memorandum	Assemble and reconcile	No publication without release	Deal lead, client and relevant advisers approve
Term acceptance	Compare and calculate	No	Client and authorised professional decide
Destructive file or system action	No ordinary permission	No	Separate controlled administrative procedure

APPENDIX C. PILOT SCORECARD

Dimension	Test	Release evidence
Evidence	Required sources, citations and contradictions	Source manifest and sampled claim record
Numbers	Reconciliation and deterministic equivalence	Calculation version and passed tests
Workflow	Correct tools, states and stop conditions	Tool log and packet-state history

Dimension	Test	Release evidence
Authority	No prohibited action; approvals present	Policy log and approval record
Quality	Rubric score and reviewer amendments	Blind comparison where practical
Economics	Accepted time, review, rework and cost	Baseline and pilot measurement file
Resilience	Provider outage, interruption and recovery	Exercise record and restored state
Security	Injection, permission and data-boundary tests	Test cases, findings and remediation

APPENDIX D. INCIDENT RECORD

An agent incident record should include the packet and transaction identifiers; discovery time; model, tool, policy and source versions; attempted and completed actions; affected data or counterparties; financial, legal, regulatory and reputational assessment; containment; rollback or correction; required notifications; root cause; control changes; regression cases; owner; and closure approval.

APPENDIX E. GLOSSARY

Acceptance criterion. An observable condition required for a work packet to pass.

Agent. A system in which a language model controls workflow execution and selects among approved tools within instructions and guardrails.

Approval gate. A state transition requiring an authorised person or control function.

Deterministic service. Versioned calculation or rule logic that supplies reproducible operational evidence.

Evidence ledger. Record linking material claims, calculations and decisions to source, version, owner and review status.

External action. A communication or system change that affects a party or environment outside the controlled internal workspace.

Human in the loop. A person approves or rejects a material output or action before completion.

Human on the loop. A person monitors defined activity and can intervene under documented conditions.

Materiality. The potential financial, legal, regulatory, confidentiality, client or reputational effect of a use case or action.

Policy engine. The control service that grants tools, data and actions according to packet identity, risk and authority.

Release certificate. The record of the exact artefact, evidence cut-off, checks, exceptions, intended audience and approvals associated with external release.

Shadow mode. Execution in which the agent produces an output for comparison and cannot change production systems or communicate externally.

Tool gateway. The typed, permissioned interface through which an agent retrieves data or performs an approved action.

Work packet. A bounded unit of delegated work with inputs, tools, acceptance criteria, authority, reviewer and stop conditions.