MatchPoint Research Working Paper No. 2026/T08

Retrieval-Augmented Generation for Financial Research: Grounded, Auditable AI

A governed source-to-claim architecture for family offices and institutional allocators

Chennakeshav (CK) Adya
Independent Researcher
July 2026

ABSTRACT

Background. Financial research requires authoritative sources, correct versions, permissions, numerical reconciliation, claim-level support and human authority. Retrieval-augmented generation connects a language model to external evidence but does not by itself establish those controls. **Objective.** This paper develops a grounded and auditable RAG architecture for family offices and institutional allocators. **Approach.** The analysis reviews 22 primary technical, official-data, regulatory and standard-setting sources available through 31 July 2026. Benchmark, survey, enforcement and consultation boundaries are retained. **Findings.** The proposed system uses a registered source corpus, structure-aware parsing, metadata and permission filters, hybrid retrieval, reranking, evidence-pack assembly, deterministic facts and calculations, claim-evidence mapping, layered evaluation, an evidence ledger and named human release. Public financial benchmarks support task-specific testing while providing no general proof of institutional production accuracy. **Implications.** Firms can begin with bounded assistive research, build labelled retrieval and numerical test sets, run cited drafts in shadow mode and enter controlled production after approved evidence, quality, authority, service, cost and incident thresholds are met. Economics should be measured per accepted research packet; revenue and investment value remain zero until supported by observed and approved attribution evidence.

Highlights

- Authority, version, permissions and evidence cut-off are retrieval inputs.
- Tables and XBRL require structured facts and deterministic calculations.
- Retrieval, extraction, calculation, claims and workflow receive separate tests.
- Every material claim links to exact evidence, calculation or visible assumption.
- Productivity is measured on accepted research packets after human review.

JEL Classification: G11, G23, G24, G32, L86, M15, O32, O33

Keywords: retrieval-augmented generation, financial research, family office, institutional allocator, evidence, citations, hybrid retrieval, XBRL, numerical reasoning, model risk, operational resilience

1. INTRODUCTION

Financial research is an evidence-constrained activity. An analyst must locate authoritative material, determine which version and reporting period applies, extract text and numbers, reconcile definitions, calculate derived measures, distinguish fact from assumption, form a conclusion and preserve the basis for review. A fluent answer without that chain is an unsupported draft.

Retrieval-augmented generation, or RAG, connects a language model to an external corpus at query time. The original RAG formulation combined parametric model memory with retrieved non-parametric memory and was evaluated on knowledge-intensive natural-language tasks [1]. Dense Passage Retrieval showed that learned dense representations could improve top-20 retrieval accuracy over a strong BM25 baseline on the open-domain question-answering datasets studied by the authors [2]. These results established useful technical building blocks. They did not establish institutional-grade accuracy for investment research, audited financial statements, private documents or numerical analysis.

Financial material adds difficulties that general question answering can obscure. The relevant evidence may sit in a filing table, footnote, amendment, covenant schedule, investor letter, data-room document or internal model. Values depend on units, periods, accounting perimeter, currency, sign and restatement status. A figure can be textually similar and financially wrong. A source can be factually relevant and contractually unavailable to the user. A statement can be supported by a document and still omit a material qualification.

Financial question-answering research confirms that the domain requires more than sentence retrieval. FinQA associates questions over financial reports with expert-written answers and executable reasoning programs [3]. TAT-QA combines tabular and textual content and requires operations such as addition, division and comparison [4]. ConvFinQA extends the task to conversational, multi-step numerical reasoning [5]. FinanceBench provides 10,231 questions with answers and evidence strings; its authors manually reviewed 2,400 answers from 16 model configurations on a 150-question sample and reported substantial limitations in the systems evaluated [6]. These are benchmark-specific findings. They support rigorous task evaluation, not a general claim about current production systems.

This paper develops a reference architecture and operating model for grounded, auditable AI in family-office and institutional-allocation research. It treats RAG as an evidence service within a governed workflow. The system retrieves candidate evidence, preserves provenance, applies deterministic extraction and calculation where appropriate, drafts claim-level outputs, exposes uncertainty and routes material conclusions to named human authority.

The proposed design has six objectives:

- retrieve the right source, version, section, table and period;
- respect identity, permissions, purpose and data rights at retrieval time;
- separate source facts, extracted values, calculations, assumptions and generated conclusions;
- attach exact evidence to material claims and disclose missing or conflicting evidence;
- evaluate retrieval, numerical reasoning, generation, workflow and human acceptance separately; and
- measure economics using accepted research packets, reviewer effort, service, cost and incidents.

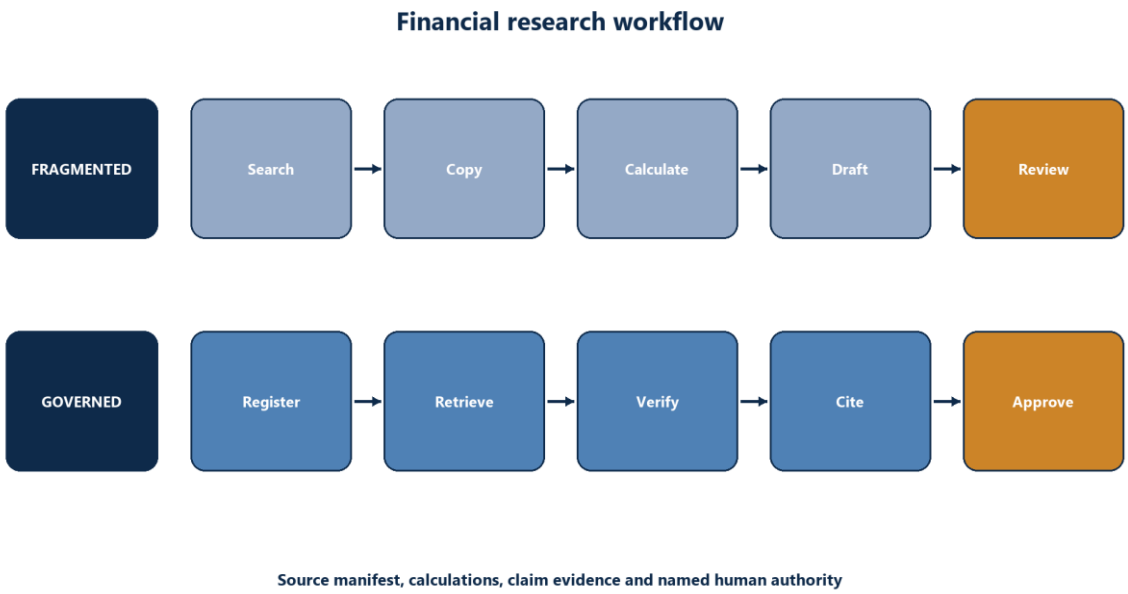

Figure 1. Financial research before and after a governed retrieval-augmented workflow
Source: Matchpoint framework.

The design does not delegate investment authority to a language model. A person or formally authorised committee owns the interpretation, recommendation and release. The architecture increases the visibility and testability of the evidence chain around that authority.

2. DEFINITIONS, SCOPE AND METHOD

2.1 Definitions

Retrieval-augmented generation means a system that retrieves external information at inference time and supplies selected context to a generative model [1]. This paper uses the term for a broader production pipeline that also includes source registration, permissions, parsing, metadata, retrieval, reranking, context assembly, citation, evaluation and records.

Grounded output means an output whose material factual claims can be traced to approved evidence or to a transparent deterministic calculation over approved inputs. Grounding is a property tested at claim level. The presence of citations alone does not establish support.

Auditable output means an output accompanied by a reproducible record of the question, user and purpose, source universe, versions, retrieval results, context, models, prompts, tools, calculations, draft, review, exceptions and release decision.

Research packet means the controlled unit delivered for review. It contains a defined question, source manifest, extracted facts, calculations, assumptions, draft analysis, claim-evidence map, unresolved exceptions and approval state.

Authoritative source means a source approved for a defined purpose. Authority is contextual. A filed annual report may be authoritative for a reported balance; an internal portfolio system may be authoritative for current holdings; counsel may own a legal interpretation. Search ranking does not confer authority.

Evidence cut-off means the timestamp through which the source set has been checked. A cut-off creates an explicit boundary for currentness and later reproduction.

2.2 Scope

Primary ICP A2 is the family office. Secondary ICP A1 is the institutional allocator. The architecture also applies to fund managers and advisory teams with comparable research workflows. Regulatory references retain their jurisdiction and entity scope. A family office or allocator should obtain its own legal, regulatory, privacy, data-rights and records analysis.

The use cases include public-company research, manager diligence, private-market underwriting, portfolio monitoring, investment-committee preparation and investor or principal reporting. The paper does not evaluate securities selection, predict returns, provide investment advice or assert that RAG improves portfolio performance.

2.3 Evidence method

The analysis uses 22 primary and authoritative sources available through 31 July 2026. The technical literature includes peer-reviewed conference or journal papers and a primary benchmark preprint. The institutional material includes regulator or standard-setter publications and official financial-data documentation.

Evidence class	Permitted use	Boundary retained
Peer-reviewed technical paper	Architecture, evaluation method and reported benchmark result	Dataset, model, date and experiment scope retained
Primary benchmark preprint	Dataset design and authors' reported evaluation	Preprint status and sample boundary retained
Official survey	Adoption, control and third-party context	Respondent-reported results; population retained
Regulation or rule text	Applicable obligation within stated legal perimeter	Jurisdiction and entity scope retained
Supervisory or enforcement material	Control, record and representation risk	Case facts and status retained
Consultation	Emerging practices and risk framing	Consultation status; not presented as binding
Matchpoint framework	Architecture, operating model and scorecards	Requires organisation-specific validation
Matchpoint scenario	Economic calculation structure	Unverified illustrative management assumptions only

No provider marketing benchmark is used as evidence of institutional accuracy or productivity. Technical results from earlier model generations are not extrapolated to current systems. Current systems still require task-specific evaluation because corpus, parser, retriever, model, prompt, tooling and workflow choices alter performance.

3. THE FINANCIAL-RESEARCH PROBLEM

3.1 Research is a chain of evidence transformations

A research conclusion rarely comes directly from one passage. A typical question such as whether leverage has increased may require identification of the right issuer and period, retrieval of a filing and amendment, extraction from financial statements and notes, reconciliation of debt definitions, normalisation of currency and units, a deterministic calculation, comparison with a prior period, review of management commentary and disclosure of perimeter changes.

Each transformation can fail independently. Retrieval may miss a footnote. Parsing may detach a header from a table. Entity resolution may select a similarly named issuer. A model may copy a number while losing its unit. A calculation may mix fiscal and calendar periods. Generation may state a causal conclusion unsupported by the evidence.

The architecture therefore treats the output as a graph:

question -> source -> segment -> fact -> calculation -> claim -> conclusion -> decision.

The arrows are recorded transformations, not merely semantic similarity.

3.2 Authority, currentness and completeness

Relevant evidence and authoritative evidence overlap imperfectly. A news article may describe an acquisition earlier than a filing. The filing may define legal terms more precisely. A data vendor may normalise values for comparison. The issuer source remains necessary to inspect exclusions, restatements and footnotes.

Every source record should include:

- source owner and publisher;
- document or dataset identifier;
- entity and instrument identifiers;
- document type and reporting period;
- publication, filing, effective and ingestion timestamps;
- version, amendment and supersession relationships;
- jurisdiction and language;
- permission, licence and purpose restrictions;
- parser and extraction version; and
- quality and currentness status.

Completeness requires a declared source universe. If the question concerns a public issuer, the universe may include all specified filing types through the cut-off, investor materials and approved market data. If the question concerns a private fund, it may include a controlled data-room index, legal documents, reports, calls and internal records. The system should state what was searched, what was unavailable and what was excluded.

3.3 Tables, XBRL and numerical reasoning

Financial documents are semi-structured. A table cell derives meaning from row label, column header, unit, period, note and surrounding disclosure. Flattening a table into arbitrary text chunks can destroy that relationship. TAT-QA and FinQA were constructed around the interaction of text, tables and numerical reasoning [3,4]. ConvFinQA adds dependencies across successive questions [5].

The SEC provides public RESTful APIs for submissions and extracted XBRL data. The submissions API exposes filing history; XBRL endpoints expose company concepts, company facts and frames [13]. SEC documentation states that the JSON structures are updated as submissions are disseminated and describes nightly bulk archives [13]. These interfaces create a useful official data channel. They do not remove the need to verify contexts, taxonomies, custom extensions, periods, units or the filed document.

The production pipeline should retain both representations:

- document-native evidence for narrative, footnotes and presentation; and
- structured facts for definitions, units, periods and deterministic computation.

Generated arithmetic should be replaced by a calculation service where a formula can be expressed. The output record should contain the formula, operands, units, period, source locations, code version and reconciliation result.

3.4 Permissions and confidential research

A correct result returned to an unauthorised user is a control failure. Retrieval must filter by user, role, legal entity, mandate, client, matter, confidentiality, data licence and permitted purpose before candidate context is assembled. Post-generation redaction cannot repair the disclosure of restricted information to the model or user.

Private-market research may contain material non-public information, personal data, commercial terms and documents subject to confidentiality restrictions. A source registry should identify restricted classes. The

orchestration layer should deny incompatible requests, select approved processing locations and record every access decision.

3.5 Long context does not remove retrieval risk

Larger context windows allow more material to be presented to a model, yet context capacity is distinct from reliable evidence use. Liu et al. found that performance on the models and tasks they studied changed with the position of relevant information; performance often declined when relevant information appeared in the middle of a long input [7]. This result supports testing context order, distractors and evidence placement. It does not establish the behaviour of every later model.

Long-context prompting and RAG can be complementary. Retrieval narrows the candidate set, metadata preserves boundaries, reranking prioritises evidence and context assembly organises the selected material. For small, stable documents, direct long-context review may be appropriate. For large, changing, permissioned or multi-source corpora, a registered retrieval layer provides stronger control over scope and reproducibility.

4. USE-CASE BOUNDARIES

4.1 RAG, search, fine-tuning and deterministic systems

These methods solve different problems.

Method	Primary role	Appropriate financial-research use	Main boundary
Keyword or semantic search	Locate candidate sources	Analyst discovery and navigation	Does not synthesise or prove support
Long-context generation	Analyse a bounded supplied set	One memorandum or small document bundle	Position, omission and cost require testing
RAG	Retrieve changing external evidence and draft grounded output	Multi-source research with citations	Retrieval and claim support require separate tests
Fine-tuning	Change behaviour, format or domain response patterns	Classification or controlled output style	Poor mechanism for current factual knowledge
Deterministic extraction	Map defined fields and structures	Dates, identifiers, XBRL facts and table cells	Requires schema and reconciliation
Calculation service	Execute formulas or code	Ratios, bridges, returns and covenant tests	Depends on correct definitions and inputs
Human analysis	Interpret materiality and exercise authority	Thesis, challenge, recommendation and release	Requires sufficient evidence and documented judgement

A robust workflow combines these components. Retrieval does not replace calculations. Fine-tuning does not replace a current source registry. A language model does not become the system of record.

4.2 Suitable initial use cases

Initial production use should be assistive, bounded and reviewable:

- retrieve and cite an approved policy, mandate or filing;
- compare specified disclosures across two versions;
- draft a source manifest and evidence index;
- extract defined fields with exact source locations;
- prepare a cited company, fund or manager profile;
- identify unresolved questions for diligence; and

- produce a first draft of a monitoring note from an approved packet.

These tasks have observable inputs, bounded outputs and clear review criteria.

4.3 Higher-risk uses

Uses receive stronger controls when the output informs an investment recommendation, valuation, allocation, trade, client statement, regulatory submission or external communication. The same summary may be low materiality when used for orientation and high materiality when quoted in an investment memorandum.

The following actions require named human authority:

- approving or rejecting an investment;
- setting valuation or financial assumptions;
- changing a portfolio, hedge or order;
- communicating a material statement externally;
- asserting regulatory or legal compliance; and
- disclosing confidential or client-specific information.

5. GOVERNED REFERENCE ARCHITECTURE

5.1 Layers

The proposed architecture has seven connected layers:

- **Research experience:** question intake, source review, evidence viewer, calculation workbench, draft editor and approval interface.
- **Workflow and policy:** identity, purpose, materiality, state, approvals, stops, exception handling and records.
- **Retrieval and reasoning:** query decomposition, hybrid retrieval, reranking, context assembly, generation and claim verification.
- **Deterministic services:** entity resolution, parsing, table extraction, XBRL, calculations and reconciliation.
- **Evidence and data:** registered documents, structured facts, metadata, permissions, versions and evidence graph.
- **Integration:** official APIs, data rooms, document systems, portfolio data and governed events.
- **Security and observability:** access, encryption, leakage controls, logs, traces, service, evaluation, cost and incidents.

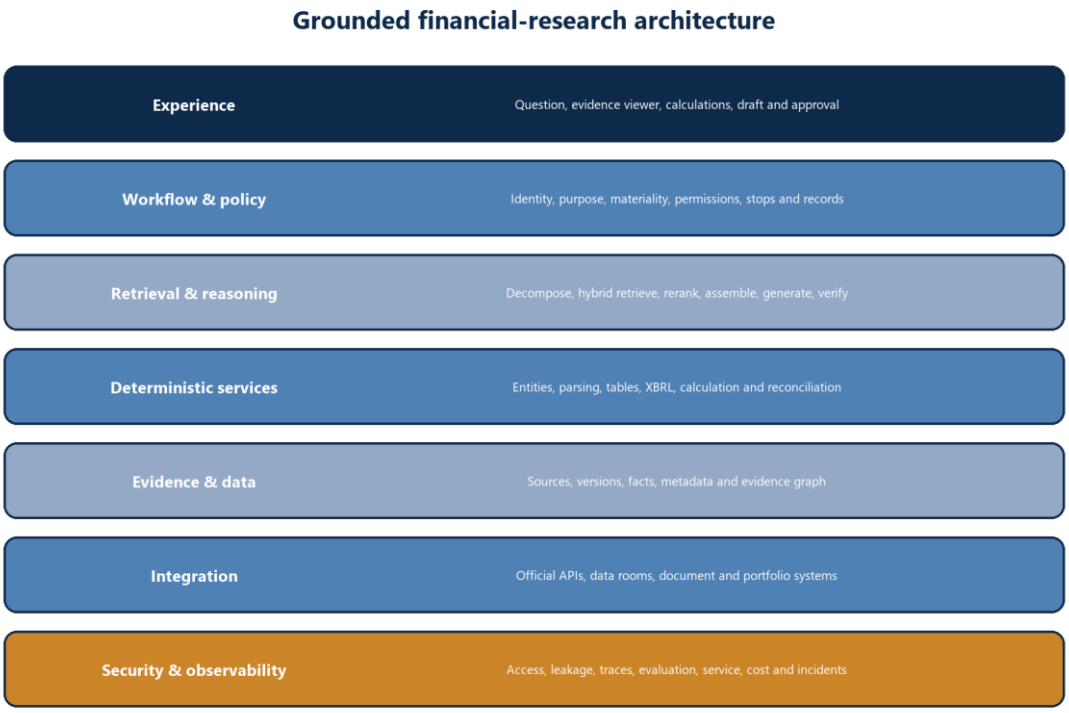

Figure 2. Governed retrieval-augmented architecture for financial research across experience, policy, retrieval, deterministic services, evidence, integration, security and observability

Source: Matchpoint architecture; technical and governance references [1,2,8,9,10,12,19,22].

The language model is one component inside the retrieval and reasoning layer. Policy, records and authority sit outside the model and remain enforceable when a model or provider changes.

5.2 Source registry and ingestion

Ingestion begins with source approval. A source owner defines the permissible use, users, update pattern, retention and quality checks. Connectors acquire material from official or authorised locations. The pipeline preserves original bytes, cryptographic hash, metadata and ingestion result.

Parsing should be document-aware. Headings, paragraphs, lists, tables, notes, page numbers and bounding boxes should survive ingestion. Optical character recognition needs confidence and human sampling. Financial tables require row-column relationships and merged-cell handling. Source pages remain available beside extracted text.

Version control should represent amendments, restatements and supersession. A later document should not silently overwrite historical evidence used for an earlier decision. The evidence graph links every extracted item to its immutable source version.

5.3 Chunking and metadata

Chunking defines the units presented to retrieval. Fixed token windows are simple and can split a definition from its qualification. Structure-aware chunks preserve section, table and footnote boundaries. Parent-child retrieval can search compact segments while returning the surrounding section. Table chunks should carry headers, units and notes with each row or cell group.

Each chunk should inherit and, where necessary, refine source metadata. Useful filters include entity, document type, filing form, period, section, geography, strategy, asset class, confidentiality, language and effective date. Metadata should support exact filters before semantic similarity is applied.

Chunking is an evaluated design choice. The test set should include footnotes, long tables, repeated boilerplate, similar period labels, amendments and cross-document questions.

5.4 Hybrid retrieval and reranking

Sparse retrieval is strong for exact terms, identifiers, accounting labels and citations. Dense retrieval can match semantic intent when wording differs [2]. Financial research benefits from a hybrid candidate set with explicit metadata filters.

A typical sequence is:

- classify the question and permitted source perimeter;
- resolve entities, instruments, dates and document types;
- decompose multi-part questions;
- retrieve sparse and dense candidates;
- merge and deduplicate by source segment;
- rerank against the specific question;
- expand parent context and linked table or footnote evidence;
- test authority, currentness and conflict; and
- assemble a bounded evidence pack.

Dense Passage Retrieval reported a 9 to 19 percentage-point absolute improvement in top-20 passage retrieval accuracy over the studied Lucene-BM25 system across the paper's open-domain datasets [2]. That result demonstrates the potential of learned retrieval within its experiment. Financial corpora require their own comparison of sparse, dense and hybrid retrieval.

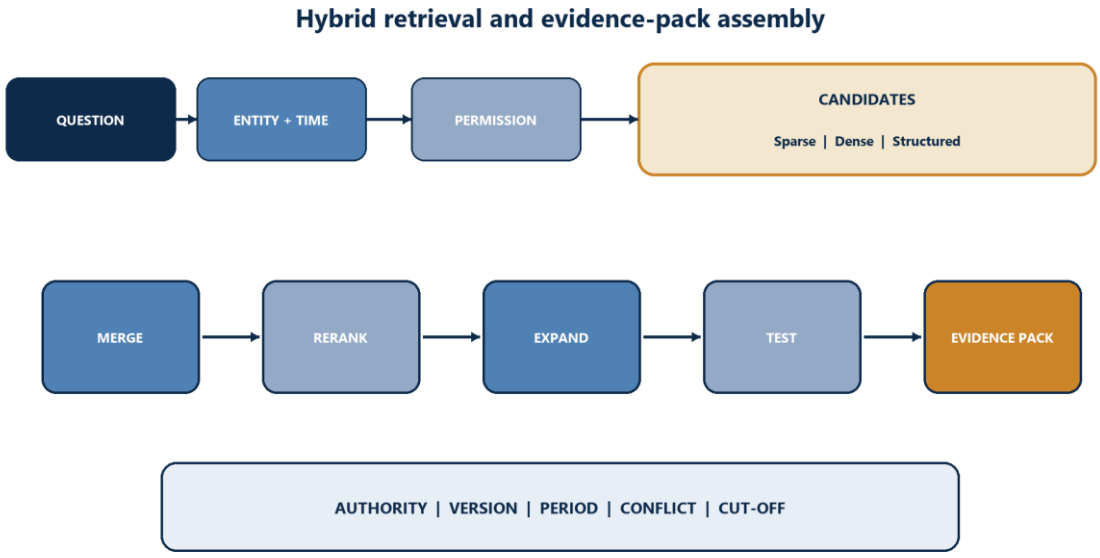

Figure 3. Hybrid retrieval, reranking and evidence-pack assembly with authority, version, period, conflict and cut-off controls

Source: Matchpoint framework; retrieval references [1,2,7,8].

5.5 Context assembly

Context assembly is a control surface. The system should present source identity, date, period, section and citation alongside each segment. It should separate authoritative evidence from background material, order evidence according to the research question, and expose contradictory or superseded sources.

The context builder should allocate explicit space for:

- question and definitions;

- source manifest and cut-off;
- primary evidence;
- structured facts and calculation results;
- conflicts and missing evidence;
- output instructions and authority limits; and
- required citation format.

Repeated and low-value passages consume attention. Reranking and diversity controls should reduce near-duplicates while retaining distinct viewpoints or periods. Tests should vary evidence position because long-context use can be position-sensitive [7].

5.6 Generation and claim-evidence mapping

The generator produces a draft with atomic material claims. Each claim is linked to one or more evidence records or to a calculation result. The system should distinguish direct quotation, paraphrase, numerical fact, calculation, assumption and analytical judgement.

An evidence link should contain:

- source and version identifier;
- page, section, table, row or data-field locator;
- supporting excerpt or structured fact;
- relation type: supports, qualifies, conflicts or background;
- retrieval and verification status; and
- reviewer disposition.

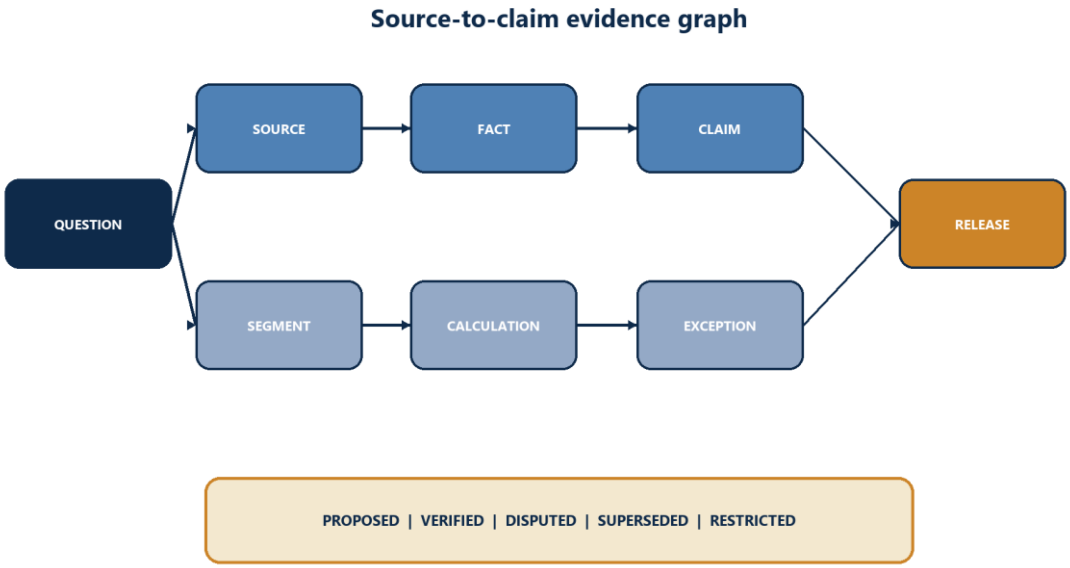

Figure 4. Source-to-claim evidence graph and research-packet states from question through human release
Source: Matchpoint framework.

The generator should abstain when required evidence is missing, stale, contradictory or outside permission. An abstention is a controlled result, not a model failure, when the evidence boundary is correctly enforced.

5.7 Evidence ledger and reproducibility

The evidence ledger records the complete run. It should be possible to reconstruct which source versions and system components produced a draft. At minimum, retain:

Record	Required fields
Request	User, purpose, question, time, mandate and materiality
Corpus	Source universe, versions, permissions and cut-off
Retrieval	Query variants, filters, candidates, scores and selected context
Processing	Parser, embeddings, reranker, model, prompt and tool versions
Facts	Value, unit, period, definition, location and verification status
Calculations	Formula, inputs, code version, output and reconciliation
Claims	Text, evidence links, confidence status and exception
Review	Reviewer, corrections, challenge, approval and release time
Operations	Latency, cost, failures, fallback and incident links

Reproduction may be limited by unavailable model versions or external sources. The ledger should record that limitation and preserve all controllable inputs and outputs.

6. EVALUATION

6.1 Evaluation is layered

An end-to-end answer score can hide the component that failed. The evaluation stack separates source coverage, retrieval, extraction, calculation, generation, workflow and human acceptance.

RAGAs proposed reference-free metrics for context relevance, faithfulness and answer relevance [9]. ARES proposed automated evaluation using synthetic training data, lightweight judges and a small human-annotated set with prediction-powered inference [10]. These approaches can accelerate iteration. Production acceptance should retain human-labelled task sets, deterministic checks and reviewer evidence because automated judges can introduce their own error.

Figure 5. Evaluation and control stack from corpus and retrieval through numerical verification, claims, workflow and human authority
Source: Matchpoint synthesis based on [3-11].

6.2 Retrieval tests

Retrieval evaluation requires a labelled corpus of realistic questions with required evidence. Metrics can include:

- recall at k: whether all required evidence appears in the top k;
- precision at k: proportion of retrieved items judged relevant;
- mean reciprocal rank for first relevant evidence;
- normalised discounted cumulative gain for graded relevance;
- source-authority precision;
- version and period accuracy;
- permission-filter accuracy; and
- conflict and missing-source detection.

A financially complete question may require several pieces of evidence. Set recall should therefore supplement single-passage metrics. Tests should include near-duplicate issuers, amended filings, fiscal-year changes, repeated accounting labels and adversarial distractors.

6.3 Extraction and numerical tests

Extraction evaluation should require exact agreement on value, unit, currency, period, sign, perimeter and source location. A numerically correct value attached to the wrong period fails.

Calculation tests should rerun expected formulas on verified operands. They should cover percentage change, margins, leverage, returns, currency conversion, weighted averages, bridge analysis and covenant definitions. Financial question-answering benchmarks show why text and numerical reasoning need separate treatment [3-5]. Srivastava et al. evaluated mathematical reasoning across FinQA, ConvFinQA, TAT-QA and MultiHiertt and reported sensitivity to table complexity and the number of arithmetic steps in their experiments [11].

6.4 Claim and citation tests

For every material claim, reviewers can assess:

- citation entailment: does the evidence support the claim;
- citation completeness: are all material parts supported;
- citation precision: are cited items actually used;
- qualification retention: are caveats, scope and uncertainty preserved;
- temporal accuracy: was the evidence current at the cut-off;
- numerical consistency: do text, table and calculations agree; and
- omission severity: is a material counterpoint missing.

Claim-level support should be evaluated separately from writing quality. A polished unsupported sentence remains a failed claim.

6.5 Workflow and authority tests

The workflow should be tested for denied-source access, prohibited model routing, required abstention, approval bypass, incorrect recipient, record failure, provider outage and rollback. An output passes only when content and process meet the approved use-case standard.

Layer	Failure mode	Test	Release evidence
Corpus	Missing, stale or unauthorised source	Manifest and access tests	Approved source perimeter
Retrieval	Required evidence absent	Labelled recall and set-coverage tests	Evidence pack pass
Extraction	Wrong value, unit or period	Exact-match and reconciliation	Verified fact set
Calculation	Formula or mapping error	Independent recomputation	Calculation pass
Generation	Unsupported or incomplete claim	Claim-citation and omission review	Accepted cited draft
Workflow	Wrong state, tool or recipient	Stop, permission and approval tests	Release record
Operations	Outage or degraded service	Scenario and fallback tests	Service evidence

6.6 Acceptance thresholds

Thresholds should be set by use case and materiality. A policy-search assistant and an investment-committee draft require different evidence coverage and review. Thresholds should specify both quality and confidence interval where sample size permits. A release decision should identify the evaluation set, date, corpus version, component versions and known exclusions.

No single benchmark establishes production readiness. FinanceBench, FinQA, TAT-QA and ConvFinQA offer valuable financial tasks [3-6]. A firm still needs a private test set drawn from its documents, questions, definitions, errors and authority boundaries.

7. GOVERNANCE, RECORDS AND REPRESENTATIONS

7.1 Governance lifecycle

NIST's Generative AI Profile is a voluntary cross-sector companion to AI RMF 1.0 [12]. It organises risk-management considerations around the AI lifecycle. For a financial-research service, governance should cover use-case approval, data, development, evaluation, release, monitoring, incident response and retirement.

The FSB's June 2026 consultation proposes 12 sound practices for financial institutions across organisation-wide governance and AI lifecycle stages [19]. It is a consultation and is not binding. IOSCO's February 2025 consultation describes AI use cases, risks and challenges in capital markets, including investment research [18]. It is also consultation material. These sources indicate the direction of current supervisory attention while retaining their status.

7.2 Accountability

Named roles should include:

- business owner for the research outcome;
- source and data owner;
- technology and model owner;
- security and privacy owner;
- independent validation or risk challenge;
- records and compliance owner where applicable; and
- human authority for release or investment use.

The Bank of England and FCA's 2024 survey received 118 responses across financial-services sectors. Seventy-five per cent of respondents reported current AI use; a third of reported current AI use cases were third-party implementations; and 34% of firms using or planning AI reported complete understanding of the AI technologies they used [20]. These are respondent-reported, cross-sector findings. They support explicit third-party and understanding controls without establishing the state of any individual family office or allocator.

7.3 Records

Applicable recordkeeping depends on entity and jurisdiction. SEC Rule 204-2 requires registered investment advisers to make and keep specified books and records and includes records relating to communications, advertisements, performance and advisory activity [15]. A firm should obtain legal analysis of which prompts, sources, generated drafts, calculations, approvals and external outputs form required records.

The SEC Division of Examinations' fiscal year 2025 priorities stated that examinations would cover the use of emerging technologies and controls intended to protect investor information, records and assets [21]. The priorities described examination focus and did not create a separate rule.

Operationally, the system should retain sufficient evidence to reconstruct a material output and the human decision. Retention, immutability, legal hold, access and deletion policies should be defined before production.

7.4 External representations

Marketing statements about AI should match implemented capability and evidence. In March 2024, the SEC announced settled charges against two investment advisers for false and misleading statements about their purported AI use; the firms agreed to USD 400,000 in total civil penalties [14]. The enforcement facts are specific to those matters. The practical control is a substantiation register linking every external capability claim to deployed systems, approved use cases and current evidence.

A research page should distinguish observed results, benchmark findings, internal pilot measures and illustrative scenarios. Statements about productivity, revenue, risk or performance should identify their basis and population.

8. SECURITY, PRIVACY AND DATA RIGHTS

8.1 Identity and least privilege

Every retrieval request should carry verified identity and purpose. Access is evaluated at source and segment level before retrieval results reach context assembly. Service accounts receive bounded permissions. Administrative activity and changes to permissions are recorded.

8.2 Prompt injection and source manipulation

Retrieved documents are untrusted content. A document can contain instructions designed to alter model behaviour or exfiltrate information. The orchestration layer should separate data from instructions, restrict tools, sanitise active content, detect anomalous instructions and require approval for material actions.

Source integrity controls include approved connectors, hashes, signature or provenance checks where available, version monitoring and quarantine. A changed source should trigger re-indexing and, where material, re-evaluation.

8.3 Data leakage

Controls should address model-provider retention, training use, processing region, encryption, logging, support access, backups and subcontractors. Sensitive prompts and outputs should be classified. The system should block unauthorised export and avoid embedding confidential content into services whose terms or architecture do not support the intended use.

8.4 Data licences

Retrieval rights do not automatically include generation, redistribution or persistent embedding. Market data, research, transcripts and data-room documents may have contractual restrictions. The source registry should encode allowed users, purpose, derived output, retention and deletion. Citations in an output should respect redistribution limits.

9. OPERATIONAL RESILIENCE AND THIRD PARTIES

9.1 Service mapping

The research service depends on connectors, object storage, parsers, embedding models, search indexes, rerankers, language models, calculation tools, identity, logs and human review. Each dependency should have an owner, service objective, failure mode, fallback and exit path.

The FCA states that firms within its operational-resilience rules were required by 31 March 2025 to be able to operate important business services within impact tolerances [16]. The scope is defined by the FCA and does not cover every family office or allocator. The underlying mapping discipline remains useful for any material research workflow.

DORA applies from 17 January 2025 and establishes digital-operational-resilience requirements for in-scope EU financial entities and ICT third-party arrangements [17]. Entity-specific applicability requires legal analysis.

9.2 Degraded modes

A production design should support:

- model or provider substitution tested against the approved evaluation set;
- a search-only mode that returns evidence without generation;
- deterministic extraction and calculation when language generation is unavailable;
- a manually assembled research packet for material deadlines;
- immutable copies of required source evidence; and
- visible service-status and incident communication.

Fallback should preserve permissions, records and authority. A faster alternative that bypasses controls is not an acceptable degraded mode.

9.3 Change and regression

Model, embedding, reranker, parser and prompt changes can alter output. Each change should run the labelled evaluation suite and compare quality, latency, cost and incidents. Material regressions block release. Older decision records retain the versions used at the time.

10. FAMILY-OFFICE APPLICATION

10.1 Principal and portfolio research

A family office may combine public markets, private funds, direct investments, real assets and operating-company interests. Evidence resides across managers, advisers, data rooms, custodians and internal staff. A common retrieval layer can give authorised users one research interface while preserving portfolio, vehicle, family branch and matter permissions.

The source graph should distinguish legal ownership, beneficial exposure, manager, vehicle, asset, counterparty and document. This reduces the risk that a similarly named entity or fund series is treated as the target investment.

10.2 Manager and fund diligence

A manager-diligence packet can include the private placement memorandum, limited partnership agreement, side letters, DDQ, track record, audited statements, valuation policy, risk reports, reference calls and internal notes. The system can retrieve clause and evidence candidates, compare versions and draft an issue list.

Human reviewers own interpretation of legal terms, track-record comparability, organisational quality and investment judgement. The packet should show missing documents, unresolved conflicts and items requiring counsel, tax or specialist review.

10.3 Monitoring

Monitoring can retrieve new reports, filings, notices and approved data; compare them with prior evidence; calculate defined thresholds; and draft a cited alert. Each alert records the trigger, source, calculation and owner. Materiality and escalation rules remain explicit.

10.4 Family confidentiality

Permissions may need to reflect family branch, generation, trust, entity, adviser and project. The corpus should avoid broad shared indexes when isolation requirements cannot be enforced at retrieval time. Sensitive personal or governance questions should use tightly controlled sources and processing paths.

11. INSTITUTIONAL-ALLOCATOR APPLICATION

11.1 Manager research at scale

Institutional allocators review large quantities of manager, market and portfolio material. A governed RAG service can assemble comparable evidence across approved templates while retaining source-specific nuance. It can identify where one manager reports gross returns and another net returns, where periods differ, or where a definition is absent.

The comparison table should link every cell to evidence and mark management-normalised values separately from manager-reported values.

11.2 Investment committee

An investment-committee packet may include mandate fit, portfolio role, exposures, liquidity, fees, terms, organisation, track record, scenario analysis, risks, conflicts and open conditions. RAG can prepare source manifests, evidence excerpts and cited drafts. Deterministic services should calculate exposure, liquidity and fee scenarios.

The committee receives the evidence boundary, assumptions, exceptions and reviewer challenge. The committee's authority, dissent and conditions are recorded outside the model.

11.3 Policy and mandate compliance

Retrieval can map a proposed investment to investment-policy statements, delegated authorities and side-letter obligations. Policy clauses require version and effective-date control. A generated compliance conclusion should be treated as a proposal for authorised review.

12. PRODUCTIVITY AND ECONOMIC CASE

12.1 Measurement unit

The economic unit is an **accepted research packet**, not a generated answer. A packet is accepted when it meets the defined evidence, numerical, citation, review and authority criteria without material correction after release.

Measure:

- attempted and accepted packets;
- analyst and reviewer active hours;
- elapsed service time;
- first-pass acceptance;
- material correction and unsupported-claim rates;
- source coverage and calculation reconciliation;
- model, data, infrastructure and control cost;
- incidents and downtime; and
- authorised downstream use.

12.2 Value equation

For a defined period:

Verified capacity value = accepted packets x (baseline active hours - production active hours) x approved loaded hourly cost.

Net measured operating value = verified capacity value + approved attributable value - technology and data cost - validation and control cost - incident and remediation cost.

Approved attributable value remains zero until management has defensible evidence linking the service to an outcome. Research quality or speed does not by itself establish incremental fees, asset growth, investment performance or loss avoidance.

12.3 Unverified illustrative scenario

All inputs below are unverified illustrative management assumptions. They are not Matchpoint operating results, forecasts or recommendations.

Input	Pilot	Team deployment	Multi-team service	Status
Attempted packets per year	240	900	2,400	Unverified illustrative management assumption
Accepted-output rate	60%	78%	86%	Unverified illustrative management assumption
Baseline active hours per accepted packet	10	12	14	Unverified illustrative management assumption
Production active hours per accepted packet	8.5	9	10	Unverified illustrative management assumption
Approved loaded hourly cost	USD 120	USD 160	USD 200	Unverified illustrative management assumption

Input	Pilot	Team deployment	Multi-team service	Status
Technology and data cost	USD 90,000	USD 300,000	USD 800,000	Unverified illustrative management assumption
Validation and control cost	USD 70,000	USD 220,000	USD 550,000	Unverified illustrative management assumption
Approved attributable revenue or investment value	USD 0	USD 0	USD 0	No observed basis assumed

The scenario exposes the break-even logic. It does not establish that any deployment produces a positive return. Management should replace every input with an observed baseline, controlled pilot result, contracted cost or approved attribution.

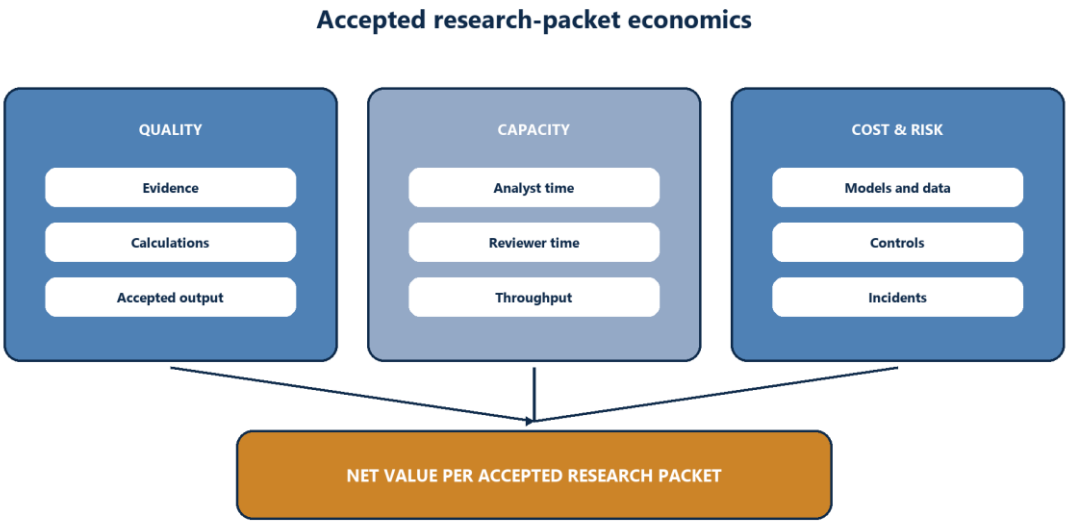

Figure 6. Productivity, quality and economic tree measured per accepted research packet

Source: Matchpoint scenario analysis. All operating inputs require observed baseline and pilot evidence.

12.4 Productivity claims

The Bank of England and FCA survey reported that respondents expected benefits related to operational efficiency, productivity and cost to increase over three years [20]. The survey measured respondents' perceptions, not realised RAG productivity. Any public Matchpoint claim should identify whether evidence comes from a benchmark, a controlled internal pilot, a client-approved case or an illustrative scenario.

13. ADOPTION ROADMAP

13.1 Stage 1; define the decision and evidence boundary

Select one research packet with a named owner, user population, source universe, materiality and acceptance standard. Record the current workflow, active hours, quality defects and service time.

13.2 Stage 2; register and secure sources

Create the source registry, entity model, version relationships, permissions and ingestion controls. Preserve original files and locations. Establish source-owner approvals and data-rights records.

13.3 Stage 3; build the retrieval baseline

Create a labelled question-evidence set. Compare sparse, dense and hybrid retrieval. Test metadata filters, amendments, tables, footnotes and permission denial. Record recall, precision and failure taxonomy.

13.4 Stage 4; add deterministic facts and calculations

Connect approved structured sources, table extraction and calculation services. Require exact value, unit, period and source-location tests. Reconcile calculations independently.

13.5 Stage 5; generate in shadow mode

Produce claim-level cited drafts without changing the existing research process. Review every packet. Measure unsupported claims, omissions, corrections, reviewer time, latency and cost.

13.6 Stage 6; controlled production

Release the service to a bounded user group after approved thresholds are met. Enforce human approval for material outputs. Monitor quality, service, permissions, cost and incidents.

13.7 Stage 7; expand by evidence

Add document types, asset classes and workflows only after regression tests and control mapping. Revalidate providers, permissions and impact tolerances. Preserve the ability to fall back to search-only or manual research.

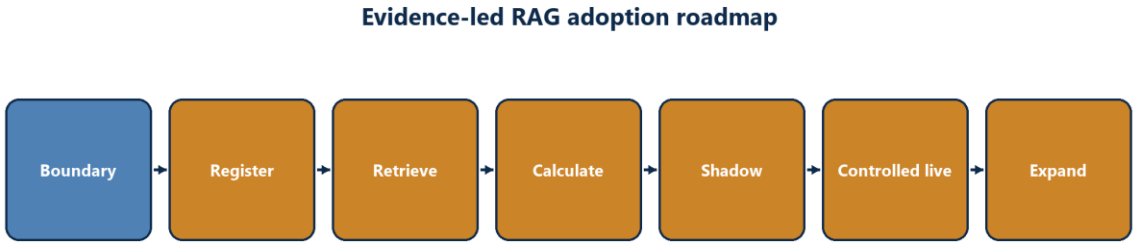

Figure 7. Adoption roadmap from evidence boundary and registered sources through shadow evaluation, controlled production and expansion

Source: Matchpoint framework.

14. IMPLEMENTATION SCORECARD

Dimension	Measure	Definition
Corpus	Required-source coverage	Required authoritative sources present and current
Permissions	Denial accuracy	Restricted sources excluded for unauthorised requests
Retrieval	Evidence-set recall	Complete required evidence retrieved within approved k
Retrieval	Authority precision	Selected sources approved for the stated purpose
Extraction	Field exact match	Value, unit, period, sign and location all correct
Calculation	Reconciliation pass	Independent result agrees within approved tolerance
Claims	Citation entailment	Evidence supports the complete material claim
Claims	Unsupported-claim rate	Material claims without adequate evidence
Completeness	Material omission rate	Required qualifications or counterevidence absent

Dimension	Measure	Definition
Quality	First-pass acceptance	Packets accepted without material correction
Productivity	Reviewer active time	Median reviewer hours per accepted packet
Service	Elapsed time	Intake to accepted and released packet
Authority	Approval violation	Material release without required human approval
Operations	Availability and degraded-mode pass	Service and fallback within approved objective
Risk	Classified incident rate	Incidents per production packet or defined period
Cost	Total cost per accepted packet	Data, models, infrastructure, people and controls

Thresholds are management decisions tied to use-case materiality. The scorecard should report sample size and confidence where appropriate. Averages should not conceal severe errors in high-materiality questions.

15. LIMITATIONS AND RESEARCH AGENDA

This paper is a design synthesis rather than an observed deployment study. It reports no Matchpoint, family-office or institutional-allocator pilot results. The economic inputs are unverified illustrative management assumptions.

The technical literature evolves quickly. Reported benchmark results depend on the model, corpus, prompting, tools and evaluation used at the time. Public financial benchmarks may underrepresent private documents, complex legal terms, multimodal evidence, permission boundaries and organisation-specific definitions.

Automated evaluation methods such as RAGAs and ARES can support iteration [9,10]. Their metrics and judges should themselves be validated against human review for the target task. A high aggregate score can coexist with rare material failures.

The architecture cannot determine legal applicability. Data protection, confidentiality, intellectual property, market-data rights, records, communications, fiduciary duties and operational-resilience requirements depend on entity, jurisdiction and use.

Research priorities include:

- permission-aware retrieval benchmarks over realistic multi-tenant corpora;
- financial table, footnote and amendment retrieval with complete evidence sets;
- claim-level evaluation that captures qualifications and omissions;
- reproducible numerical reasoning using deterministic tools;
- adversarial testing for prompt injection and manipulated source documents;
- calibration and abstention for missing or conflicting evidence;
- human-review studies measuring accepted packets and material corrections;
- provider concentration and degraded-mode testing; and
- longitudinal evidence linking research-service quality to approved business outcomes.

16. CONCLUSION

Retrieval-augmented generation can provide a controlled route from a changing evidence corpus to a cited research draft. Financial research requires a fuller system: authoritative sources, permissions, versions, structure-aware parsing, hybrid retrieval, deterministic facts and calculations, claim-level evidence, layered evaluation, records and named human authority.

The practical sequence is evidence first. A firm defines the question and source perimeter, builds a labelled retrieval baseline, adds deterministic services, evaluates cited drafts in shadow mode and enters controlled production only after approved quality, authority, service, cost and incident thresholds are met.

The economic case should be measured per accepted research packet. Reviewer time and throughput matter only when evidence, calculations, citations and authority satisfy the release standard. Revenue, investment-performance and loss-avoidance claims require separate observed evidence and approved attribution.

For family offices and institutional allocators, the central design objective is a reproducible evidence-to-decision chain. RAG is valuable when it makes that chain easier to assemble, inspect, challenge and govern.

DECLARATIONS

Author: Chennakeshav (CK) Adya, Independent Researcher.

Date: July 2026.

Funding: No external funding was reported for this paper.

Conflicts of interest: The author is associated with Matchpoint Partners. This paper presents a Matchpoint framework and identifies that status where relevant.

Data and code availability: No proprietary dataset or production system was used. Public source locations are listed in the references. The figures and scenario framework are original Matchpoint synthesis.

Use of generative AI: Generative AI assisted drafting and production. The author remains responsible for source selection, interpretation and final content. No statement should be treated as independently verified solely because it was generated or cited.

Disclaimer: Research paper only. This document is not legal, tax, accounting, regulatory, cybersecurity or investment advice. Regulatory applicability and production controls require entity-specific professional analysis. The scenario does not report observed productivity, revenue, investment performance or risk reduction.

REFERENCES

[1] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33. https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[2] Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of EMNLP 2020, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[3] Chen, Z., Chen, W., Smiley, C., et al. (2021). FinQA: A Dataset of Numerical Reasoning over Financial Data. Proceedings of EMNLP 2021. https://doi.org/10.18653/v1/2021.emnlp-main.300

[4] Zhu, F., Lei, W., Huang, Y., et al. (2021). TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. Proceedings of ACL-IJCNLP 2021, 3277-3287. https://doi.org/10.18653/v1/2021.acl-long.254

[5] Chen, Z., Li, S., Smiley, C., et al. (2022). ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. Proceedings of EMNLP 2022, 6279-6292. https://doi.org/10.18653/v1/2022.emnlp-main.421

[6] Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N. and Vidgen, B. (2023). FinanceBench: A New Benchmark for Financial Question Answering. arXiv:2311.11944. https://arxiv.org/abs/2311.11944

[7] Liu, N. F., Lin, K., Hewitt, J., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, 157-173. https://doi.org/10.1162/tacl_a_00638

[8] Asai, A., Wu, Z., Wang, Y., Sil, A. and Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024. https://openreview.net/forum?id=hSyW5go0v8

[9] Es, S., James, J., Espinosa Anke, L. and Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of EACL 2024 System Demonstrations, 150-158. https://doi.org/10.18653/v1/2024.eacl-demo.16

[10] Saad-Falcon, J., Khattab, O., Potts, C. and Zaharia, M. (2024). ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. Proceedings of NAACL 2024, 338-354. https://doi.org/10.18653/v1/2024.naacl-long.20

[11] Srivastava, P., Malik, M., Gupta, V., Ganu, T. and Roth, D. (2024). Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering. Findings of ACL 2024. https://doi.org/10.18653/v1/2024.findings-acl.231

[12] Autio, C., Schwartz, R., Dunietz, J., et al. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[13] U.S. Securities and Exchange Commission (2025). EDGAR Application Programming Interfaces. Last reviewed or updated 8 April 2025. https://www.sec.gov/search-filings/edgar-application-programming-interfaces

[14] 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, 18 March 2024. https://www.sec.gov/newsroom/press-releases/2024-36

[15] U.S. Securities and Exchange Commission (2024). 17 CFR 275.204-2, Books and records to be maintained by investment advisers; reverted rule text following vacatur notice. https://www.sec.gov/files/investment/pfa-vacatur-reverted-rule-text.pdf

[16] Financial Conduct Authority (2026). Operational resilience. Last updated 14 July 2026. https://www.fca.org.uk/firms/operational-resilience

[17] European Parliament and Council (2022). Regulation (EU) 2022/2554 on digital operational resilience for the financial sector. Official Journal L 333, 1-79; applicable from 17 January 2025. https://eur-lex.europa.eu/eli/reg/2022/2554/oj

[18] International Organization of Securities Commissions (2025). Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges. Consultation Report, CR01/2025. https://www.iosco.org/library/pubdocs/pdf/IOSCOPD788.pdf

[19] Financial Stability Board (2026). Sound Practices for Responsible Adoption of Artificial Intelligence: Consultation Report. 10 June 2026. https://www.fsb.org/2026/06/sound-practices-for-responsible-adoption-of-artificial-intelligence-ai-consultation-report/

[20] Bank of England and Financial Conduct Authority (2024). Artificial intelligence in UK financial services - 2024. 21 November 2024. https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024

[21] U.S. Securities and Exchange Commission, Division of Examinations (2024). Fiscal Year 2025 Examination Priorities. https://www.sec.gov/files/2025-exam-priorities.pdf

[22] National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

APPENDIX A. RESEARCH-PACKET SCHEMA

Field	Requirement
Packet ID	Stable identifier, use case and materiality
Question	Exact question, definitions and required output
Requester	User, role, entity, mandate and purpose
Cut-off	Date and time through which evidence was checked
Source universe	Required and optional source classes
Source manifest	Source IDs, versions, permissions and ingestion status
Retrieval record	Queries, filters, candidates, scores and selected context
Facts	Value or statement, unit, period, perimeter and location
Calculations	Formula, operands, code version, output and reconciliation
Assumptions	Value, range, basis, owner and scenario
Claims	Atomic claim, evidence links and support status
Conflicts	Contradictory evidence and reviewer disposition
Missing evidence	Unavailable required sources and effect on conclusion
Draft	Generated and edited analysis with citations
Challenge	Reviewer questions, corrections and residual issues
Approval	Named authority, date, conditions and release state
Operations	Component versions, latency, cost, errors and incident links

APPENDIX B. SOURCE REGISTRY

Field	Example control
Publisher	SEC, issuer, manager, counsel or internal system owner
Source type	Filing, agreement, report, transcript, dataset or model
Entity	Verified legal and internal identifiers
Period	Reporting, effective and publication dates
Version	Original, amendment, restatement or superseded version
Permission	User, role, entity, mandate and purpose
Licence	Ingestion, embedding, generation, redistribution and retention rights
Integrity	Original bytes, hash and connector provenance
Parser	Parser, OCR and table-extraction versions
Quality	Verified, disputed, stale, restricted or quarantined
Owner	Business owner and technical custodian

APPENDIX C. PILOT TEST SET

The pilot test set should include ordinary and deliberately difficult cases:

- exact filing and policy lookups;
- similar entity and instrument names;
- amended or superseded documents;
- tables with multi-level headers and footnotes;
- unit, currency and sign ambiguity;
- fiscal and calendar-period mismatch;
- questions requiring several sources;
- missing authoritative evidence;
- conflicting sources;
- restricted sources requested by an unauthorised user;
- prompt injection embedded in a document;
- calculations requiring deterministic tools;
- material qualifications placed in the middle of long context; and
- outage and degraded-mode scenarios.

Each case should specify required evidence, forbidden evidence, expected facts, formula where applicable, acceptable conclusion boundary, required abstention and approval state.

APPENDIX D. AUTHORITY MATRIX

Activity	System role	Human reviewer	Named authority
Register source	Prepare metadata and checks	Source owner verifies	Data authority approves source class
Retrieve evidence	Apply filters and rank candidates	Analyst inspects	Within approved use case
Extract fact	Propose value and location	Analyst verifies	Research lead accepts fact set
Calculate	Execute approved deterministic service	Independent check	Analyst owns interpretation
Draft claim	Generate cited text	Senior reviewer challenges	Research authority releases
Prepare investment memorandum	Assemble packet	Deal or manager team owns conclusion	Committee decides
Prepare external communication	Draft from approved evidence	Compliance or legal where required	Named approver releases
Change component	Build and test release	Validation, security and business review	Change authority approves

APPENDIX E. GLOSSARY

Abstention: A controlled output indicating that the system lacks permitted, current or sufficient evidence.
Chunk: A retrievable unit derived from a source while retaining metadata and provenance.
Context assembly: Selection and organisation of retrieved evidence supplied to a model.
Dense retrieval: Retrieval using learned vector representations of queries and passages.
Evidence graph: Linked representation of source, segment, fact, calculation, claim and decision.
Hybrid retrieval: Combination of sparse, dense and metadata-constrained retrieval.
Reranker: A component that reorders retrieved candidates against a specific question.
Sparse retrieval: Term-based retrieval such as BM25 that uses lexical matching.
XBRL: Structured reporting format used for tagged financial facts.