MatchPoint Research Working Paper No. 2026/T11

AI Copilots for Deal Teams: Measuring the Productivity Premium

A quality-adjusted measurement, architecture and adoption framework for family-office investment teams and GCC fund managers

Chennakeshav (CK) Adya
Independent Researcher
August 2026

ABSTRACT

Background. Generative-AI studies report substantial task gains in several knowledge-work settings and a slowdown in another, while financial firms identify data, cyber, model and third-party risks. **Objective.** This paper develops a governed method for measuring the productivity premium from AI copilots used by family-office investment teams and GCC fund managers. **Approach.** The analysis reviews 24 primary studies, official surveys, standards and regulatory publications available through 1 August 2026; task, population, jurisdiction and evidence boundaries are retained. **Findings.** The proposed design measures quality-adjusted accepted deal-work packets, includes producer, reviewer, remediation and failure time, and separates gross release, realised capacity and approved economic attribution. It combines a task frontier, source-grounded architecture, deterministic numerical checks, release authority and a baseline-shadow-pilot sequence. **Implications.** Teams can begin with evidence assembly, DDQ reuse, data-room readiness, monitoring and controlled memo drafting. Investment, valuation, conflict, legal and external-release authority remains with named professionals. Attributed revenue, cost reduction and loss reduction remain zero until approved observed evidence supports them.

Highlights

Unit. The accountable unit is a quality-adjusted accepted deal-work packet.

Evidence. External productivity percentages remain bounded to their tested tasks and populations.

Measurement. Producer, reviewer, remediation, failure and fallback time remain in the denominator.

Authority. Investment, valuation, conflicts, legal conclusions and external release remain human decisions.

Attribution. Revenue, cost reduction and loss reduction remain zero until approved observed evidence exists.

JEL Classification: G23, G24, G32, L84, L86, M15, O32, O33

Keywords: artificial intelligence, generative AI, AI copilot, deal team, productivity measurement, family office, fund manager, investment committee, due diligence, human oversight, model risk

1. INTRODUCTION

Deal teams produce decisions through a chain of evidence, judgement, calculation, challenge and approval. A family-office investment team may screen an opportunity, test portfolio fit, interrogate a manager, reconcile a model, write an investment-committee memorandum and monitor the position. A fund manager may originate an asset, manage a data room, coordinate diligence, prepare a valuation case, draft committee materials, answer limited-partner questions and preserve a record of the decision. Each output inherits the quality and permissions of the material that preceded it.

Generative artificial intelligence can assist with several parts of that chain. It can retrieve approved material, extract terms, compare documents, draft summaries, structure questions, prepare first-pass narrative and route exceptions. Those capabilities do not establish a productivity premium by themselves. A faster draft that requires more review, loses source traceability, exposes confidential information or introduces a material error can reduce economic value. The relevant unit is a **quality-adjusted, accepted deal-work packet**: a bounded output that a named reviewer accepts after required evidence, calculations, conflicts and permissions have passed their controls.

The external evidence supports a measured approach. In a preregistered experiment involving 453 college-educated professionals completing writing tasks, Noy and Zhang reported a 40% reduction in completion time and an 18% increase in output quality for the group given ChatGPT [1]. In a field study of 5,179 customer-support agents, Brynjolfsson, Li and Raymond reported a 14% average increase in issues resolved per hour, with larger gains among novice and lower-skilled workers [2]. In an experiment involving 758 consultants, AI-assisted participants completed more tasks, worked faster and produced higher-rated work for tasks inside the tested capability frontier; performance fell for a task outside that frontier [3]. These studies examine writing, service and consulting tasks. They do not measure investment returns, deal completion, fiduciary outcomes or realised revenue for the two target ICPs in this paper.

Evidence published since those early studies adds important boundaries. A field experiment across 66 organisations and 7,137 knowledge workers found that active users of an integrated generative-AI tool spent two fewer hours on email per week, while the researchers did not detect changes in the overall quantity or composition of tasks from individual access alone [5]. A randomised study of experienced open-source developers working on their own repositories found that early-2025 tools increased completion time by 19% in that setting [6]. METR's 2026 follow-up reported indications of modest later speed-ups, while explicitly describing selection and measurement problems that prevented a reliable estimate [7]. Productivity is therefore a property of a task, workflow, population, tool configuration and measurement period. It is not a transferable percentage attached to a product.

Financial-services adoption is already material. The Bank of England and Financial Conduct Authority received 118 responses in their 2024 survey; 75% of respondents said they were using AI and another 10% planned to use it within three years [8]. Foundation models represented 17% of reported AI use cases, operations and IT represented the largest business area, and 46% of respondents described only a partial understanding of the AI technologies they used [8]. The same survey identified data-related issues across four of the five highest perceived current risks and reported cybersecurity as the highest perceived systemic risk [8]. These results describe the survey population and respondents' perceptions. They do not establish an outcome for a specific family office or fund manager.

This paper develops an operating and measurement system for **Family-Office CIOs and Heads of Alternatives (A2)** and **GCC Fund Managers or GPs Raising Capital (B2)**. It answers five questions:

1. Which deal tasks are suitable for a copilot, and which decisions remain outside autonomous scope?
2. How should a firm establish a baseline and measure speed, quality, throughput, review burden and reallocation?
3. What architecture preserves source lineage, numerical control, confidentiality, recordkeeping and named authority?
4. How should the productivity result be translated into capacity value, cost, risk and potential revenue without unsupported attribution?
5. What staged adoption path can produce credible evidence before scale?

The central proposition is testable: a deal-team copilot creates economic value only when it raises accepted, quality-adjusted output or releases capacity that the organisation can verify and redeploy. The proposed scorecard records the full conversion from model interaction to accepted output, realised capacity and approved economic attribution.

At launch, **Attributed revenue: USD 0. Attributed cost reduction: USD 0. Quantified loss reduction: USD 0.** Those fields change only after a named authority accepts observed evidence and its attribution method.

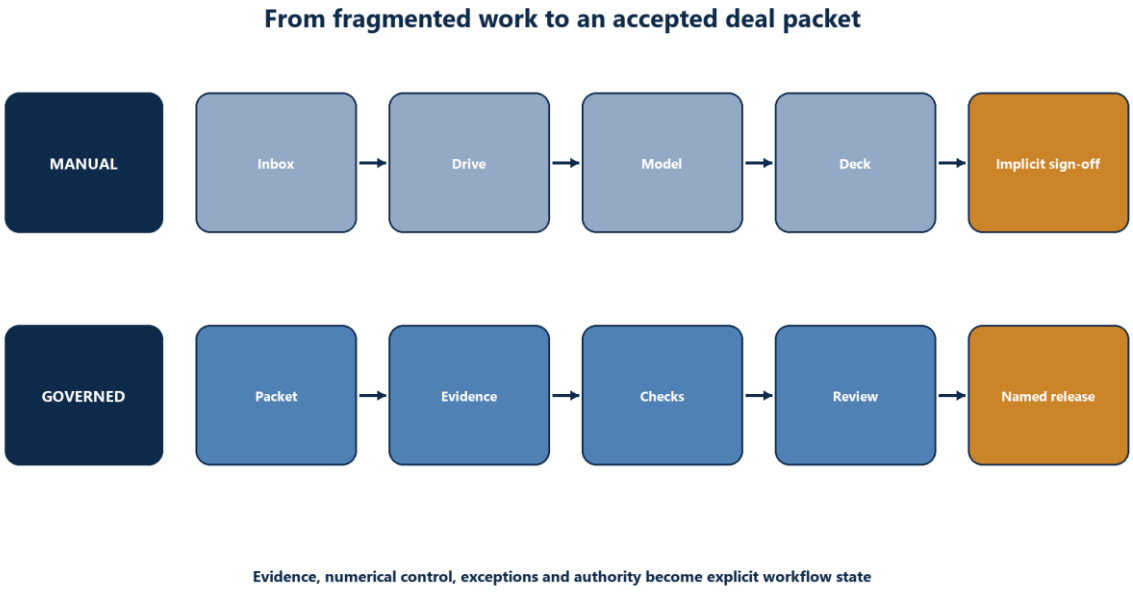

Figure 1. Deal workflow before and after a governed AI copilot
Source: Matchpoint framework.

2. SCOPE, DEFINITIONS AND EVIDENCE BOUNDARIES

2.1 What this paper means by an AI copilot

An AI copilot is an assistive system operating inside a human-owned workflow. It can combine a language model with retrieval, document parsing, deterministic calculations, workflow tools and policy controls. It does not possess delegated investment authority under this framework. It may propose, organise or test work; the accountable professional owns the decision, the communication and the release.

The term **deal team** covers the professionals and approved specialists producing an investment, financing, fund-raising or transaction decision. It includes investment professionals, capital-formation professionals, finance, legal, tax, compliance, operations and external advisers where their work is part of the controlled process. The proposed measurements apply to work packets rather than job titles, because the same professional may perform high-volume drafting and high-stakes judgement within one day.

The **productivity premium** is the measured difference between a governed copilot workflow and an approved comparator after speed, quality, acceptance, review burden, failure cost and realised reallocation have been considered. It is expressed separately as time, capacity, quality, throughput and economics. Combining them prematurely can hide a trade-off.

Term	Operational definition	Required evidence
Attempted packet	A bounded task submitted to the copilot workflow	Packet identifier, owner, task class, start time
Completed packet	An output returned with the required fields	Output, version, elapsed time, tool record
Accepted packet	A completed packet approved by the named reviewer	Review result, exceptions, approval time
First-pass acceptance	Acceptance without material correction	Reviewer classification and correction log
Material defect	An error that could alter a decision, disclosure, recipient action or compliance outcome	Defect record, severity, affected source or calculation

Term	Operational definition	Required evidence
Gross time saved	Baseline median time less treatment median time for comparable accepted packets	Matched or randomised observation
Realised capacity	Verified time released and demonstrably redeployed or removed	Time record plus approved reallocation evidence
Productivity premium	Quality-adjusted accepted output per paid hour relative to the comparator	Baseline, treatment and quality denominator
Economic attribution	Approved connection between measured change and revenue, cost or avoided loss	Attribution method, finance approval, period and caveats

2.2 Evidence hierarchy

This paper separates four kinds of evidence. **Tier 1** is a randomised or preregistered study with observable task outcomes. **Tier 2** is a large field study, official survey or peer-reviewed empirical paper. **Tier 3** is an official regulatory, standards or supervisory publication. **Tier 4** is a vendor study, working paper, management estimate or proposed Matchpoint framework. Each tier can inform a decision, but it cannot answer questions outside its design.

The evidence cut-off is 1 August 2026. Working papers are labelled. Vendor research is not treated as independent validation. Regulatory statements are reported for their stated jurisdiction and scope. The paper does not provide legal advice and does not determine whether a particular system, firm or use case is subject to a specific rule.

2.3 What the external studies do and do not show

Source and setting	Reported result	Boundary for deal teams
Noy and Zhang; 453 professionals, incentivised writing tasks [1]	Average time decreased 40%; quality increased 18%	Useful benchmark for bounded professional drafting; not a deal, valuation or revenue study
Brynjolfsson, Li and Raymond; 5,179 support agents [2]	Productivity increased 14% on average; 34% for novice and lower-skilled workers	Shows heterogeneity and possible knowledge diffusion; work was customer support
Dell'Acqua et al.; 758 consultants [3]	More tasks, higher speed and higher quality inside the tested AI frontier; lower correctness outside it	Closest task structure to analytical deal work; capability boundary and task design remain decisive
Dell'Acqua et al.; 776 product professionals [4]	Individuals using AI matched the performance of teams without AI on the tested product-innovation challenge	Supports testing collaboration structures; not evidence that AI replaces transaction teams
Dillon et al.; 7,137 workers across 66 firms [5]	Active users spent two fewer hours on email per week; no detected change in total task quantity or composition	Time saved can remain unconverted unless workflows and reallocation change
METR; 16 experienced developers and 246 repository tasks [6]	AI use increased completion time by 19% in the tested early-2025 setting	Demonstrates that expertise, context and tool overhead can reverse the expected effect
METR 2026 update [7]	Later results suggested small speed-ups but were described as unreliable because of selection and measurement effects	Measurement design can become invalid when users select out or run parallel agents
GitHub; 95 developers on a controlled coding task [23]	Vendor study reported 55% faster completion for the Copilot group	Relevant as product-sponsored task evidence; not an independent finance benchmark

The product-innovation teamwork experiment adds evidence on human-AI collaboration [4]. The GitHub result adds a vendor-conducted coding benchmark and is labelled accordingly [23]. The studies collectively support three design choices. First, results must be reported by task and experience cohort because averages conceal heterogeneity. Second, a credible pilot requires an outcome that includes quality and reviewer effort. Third, workflow redesign and reallocation must be measured after the individual task effect; an organisation cannot book a capacity benefit merely because a draft was generated faster.

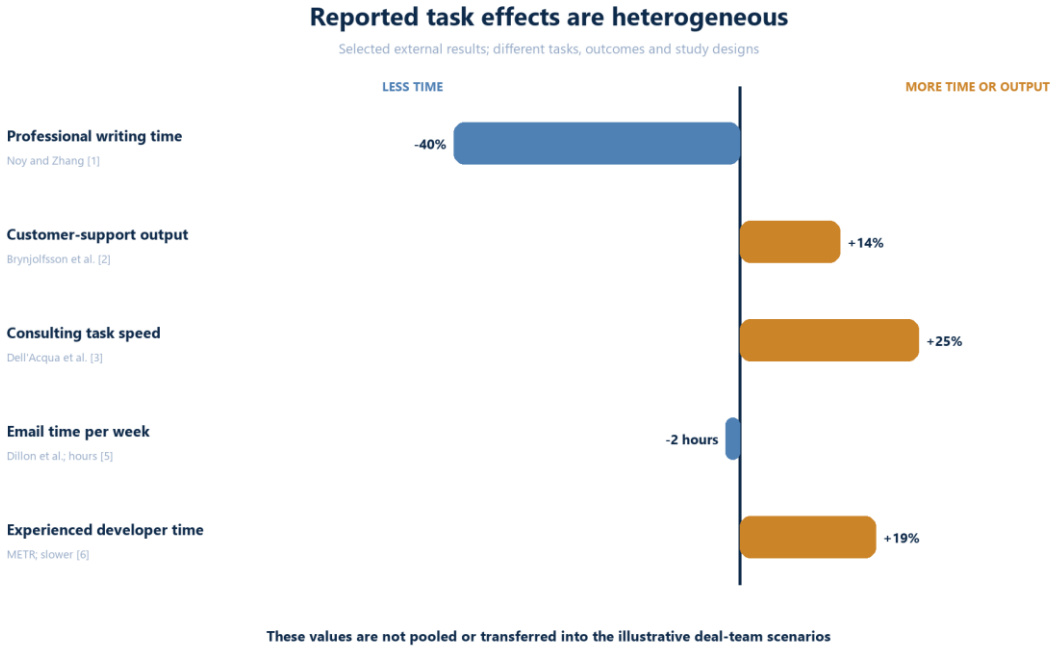

Figure 2. External productivity evidence and its transfer limits

Source: Matchpoint synthesis of [1-3,5-7]. Results use different outcomes and are not pooled.

2.4 Financial-services context

FINRA states that its rules and the federal securities laws continue to apply when member firms use generative AI [9]. Its 2020 report describes AI uses across customer communication, investment processes and operations, while highlighting governance, model risk, data, privacy, cybersecurity, outsourcing and supervisory considerations [10]. IOSCO's 2025 consultation describes capital-markets use of large language models in information and process management, internal knowledge search, document standardisation, meeting materials and code-related work; it also examines risks to investor protection, market integrity and financial stability [11]. ESMA's 2024 statement identifies potential benefits and risks when firms use AI in investment services and says management bodies remain responsible for decisions [12].

The SEC's 2024 settlements with two investment advisers concerned false or misleading statements about how they used AI [13]. Separate SEC recordkeeping actions against financial firms reinforce that electronic communications and required records remain subject to preservation obligations [21]. These publications do not prohibit controlled internal assistance. They establish a strong reason to connect every released claim, client communication and decision record to evidence, supervision and policy.

For UAE-based organisations, Federal Decree-Law No. 45 of 2021 provides the federal personal-data protection framework described by the UAE Government portal [16]. Other regimes may apply depending on establishment, data subjects, free-zone status, clients and processing. The EU AI Act applies according to its stated scope, including certain providers and deployers in or connected to the Union [17]. A firm should obtain qualified advice for its exact facts before relying on a jurisdictional conclusion.

3. THE DEAL-TEAM FRONTIER

3.1 The task, not the job, is the unit of adoption

A single investment professional may complete tasks with very different risk profiles. Extracting a maturity date from a signed agreement is bounded and verifiable. Reconciling EBITDA adjustments requires accounting context and deterministic calculations. Judging the credibility of management, assessing a conflict, setting a valuation range or approving a commitment involves material professional authority. A useful adoption plan decomposes the workflow into packets with explicit evidence, tests and owners.

Three dimensions determine initial suitability:

- **Verifiability:** Can the output be tested against authoritative sources or deterministic rules?
- **Materiality:** Could an error alter capital allocation, a client communication, a legal right or a regulatory outcome?
- **Context stability:** Is the task repeatable with stable definitions, or does it depend on tacit knowledge and a changing transaction narrative?

The resulting frontier is deliberately conservative. Green tasks can enter a controlled pilot after data and permission approval. Amber tasks require stronger review, deterministic checks and restricted release. Red tasks remain human decisions; the copilot may assemble evidence or questions without making the decision.

Deal task	Initial zone	Copilot role	Human control
Data-room indexing and document classification	Green	Tag approved files, detect duplicates, identify missing classes	Data-room owner confirms population and permissions
Contract-term extraction	Green	Extract named terms with page and clause citations	Reviewer verifies against executed version
Meeting-note structuring	Green	Convert approved transcript into actions and open questions	Meeting owner edits, approves and controls distribution
DDQ and RFI first draft	Green	Retrieve approved prior answers and draft cited response	Function owner verifies currency, scope and recipient
Comparable-company evidence pack	Amber	Assemble dated sources and normalise fields	Analyst validates source date, units, adjustments and population
Financial-model commentary	Amber	Explain approved model outputs and surface inconsistencies	Model owner controls formulas, assumptions and conclusions
Investment-memo first draft	Amber	Structure evidence, alternatives, risks and unresolved questions	Deal lead owns thesis, challenge, recommendation and sign-off
LP or client communication	Amber	Draft from approved facts and disclosure library	Authorised person approves accuracy, balance and distribution
Valuation conclusion	Red	Assemble evidence and sensitivity questions	Qualified professional determines and approves value
Investment recommendation or commitment	Red	Prepare record and challenge prompts	Investment committee retains decision authority
Conflict waiver or legal interpretation	Red	Retrieve approved policy and route issue	Compliance or qualified counsel decides
Autonomous external negotiation or release	Red	No autonomous action under this framework	Named authorised professional communicates or executes

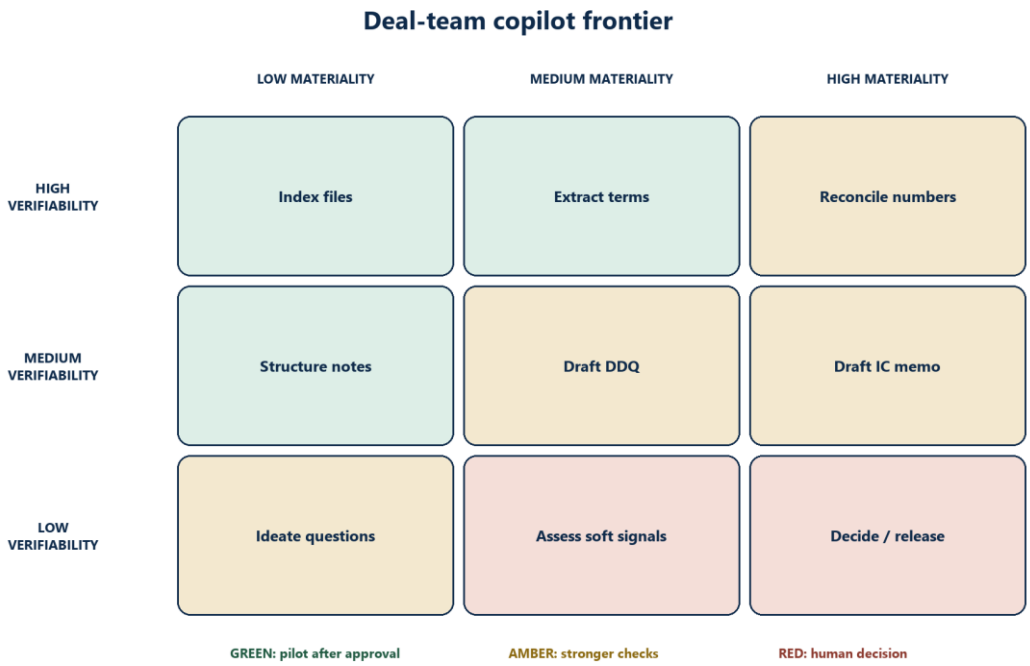

Figure 3. The deal-team copilot frontier by verifiability and materiality
Source: Matchpoint framework; governance references [9-15,18-22].

3.2 Before-and-after workflow

The manual workflow often moves through email, shared drives, spreadsheets, chat, decks and memory. Provenance fragments across versions. Reviewers spend time rebuilding the population and locating support. A governed copilot can create a packet with a stable identity, source manifest, draft, exceptions and decision record. The objective is a shorter, more auditable route to acceptance.

Stage	Common manual state	Governed copilot state	Acceptance evidence
Intake	Request arrives without a stable identifier	Packet receives deal, entity, task, owner, deadline and materiality fields	Complete intake record
Retrieval	Analyst searches several repositories	Permission-aware retrieval returns approved sources and exclusions	Source manifest and population check
Production	Facts, calculations and narrative are mixed	Extracted facts, deterministic calculations, model inferences and draft narrative remain separate	Field-level lineage and calculation outputs
Challenge	Review occurs through comments and memory	Contradictions, missing evidence and policy triggers are explicit exceptions	Exception register and resolution
Approval	Approval may be implicit in circulation	Named authority accepts, rejects or returns the packet	Signed decision state
Release	Output is copied to email, deck or portal	Approved version and recipients pass a release gate	Release log and retained record

3.3 Failure modes that erase the premium

The most common measurement error is to count model speed while ignoring the downstream work it creates. Hallucinated citations, stale comparables, unit errors, missing pages, duplicated entities, uncontrolled assumptions and prose that overstates evidence all increase review time. Retrieval can fail silently when permissions, indexing or document versions are wrong. A fluent answer can also encourage automation bias, causing reviewers to inspect less carefully.

The University of Chicago page for a withdrawn working paper on large-language-model financial-statement analysis states that co-authors identified inconsistencies in underlying data and analyses and withdrew the paper while reviewing the findings [24]. That event does not establish a general conclusion about LLM financial analysis. It illustrates why reproducibility, source retention and independent checks matter when claims are financially consequential.

NIST's AI Risk Management Framework organises risk work around Govern, Map, Measure and Manage [14]. Its Generative AI Profile identifies risks that include confabulation, data privacy, information integrity, information security, intellectual property and value-chain or component integration, and proposes actions across the same functions [15]. The OECD AI Principles and NIST Cybersecurity Framework 2.0 provide additional cross-sector governance and cybersecurity reference points [20,22]. The framework in this paper translates those principles into the packet, metric and release controls used by deal teams.

4. MEASURING THE PRODUCTIVITY PREMIUM

4.1 The measurement chain

The measurement chain has six gates:

1. **Eligible task:** The task belongs to an approved use-case class and data boundary.
2. **Completed output:** The system returns the required fields within the measurement window.
3. **Accepted output:** A named reviewer accepts the output under a defined rubric.
4. **Quality adjustment:** Material defects, rework, reviewer time and delayed exceptions are included.
5. **Realised capacity:** The organisation verifies how released time was redeployed or removed.
6. **Economic attribution:** Finance or management approves the link to capacity value, cost, revenue or avoided loss.

Each gate has a denominator. Reporting only successful attempts creates survivorship bias. The dashboard therefore includes all eligible packets, withdrawals, system failures, rejected outputs and manual fallbacks.

The primary productivity measure is:

Quality-adjusted accepted output per paid hour = accepted packets x quality weight / (producer time + reviewer time + remediation time).

The relative productivity premium is:

Premium = treatment quality-adjusted output per paid hour / comparator quality-adjusted output per paid hour - 1.

This definition prevents a faster producer from appearing productive when reviewer or remediation time rises. The quality weight should be fixed before the pilot and based on observable fields. A simple scheme can assign 1.00 to an accepted packet with no material defect, 0.75 to an accepted packet with material correction before use, and 0 to a rejected or post-release materially defective packet. A firm may choose another rubric and should retain it unchanged during a measurement period.

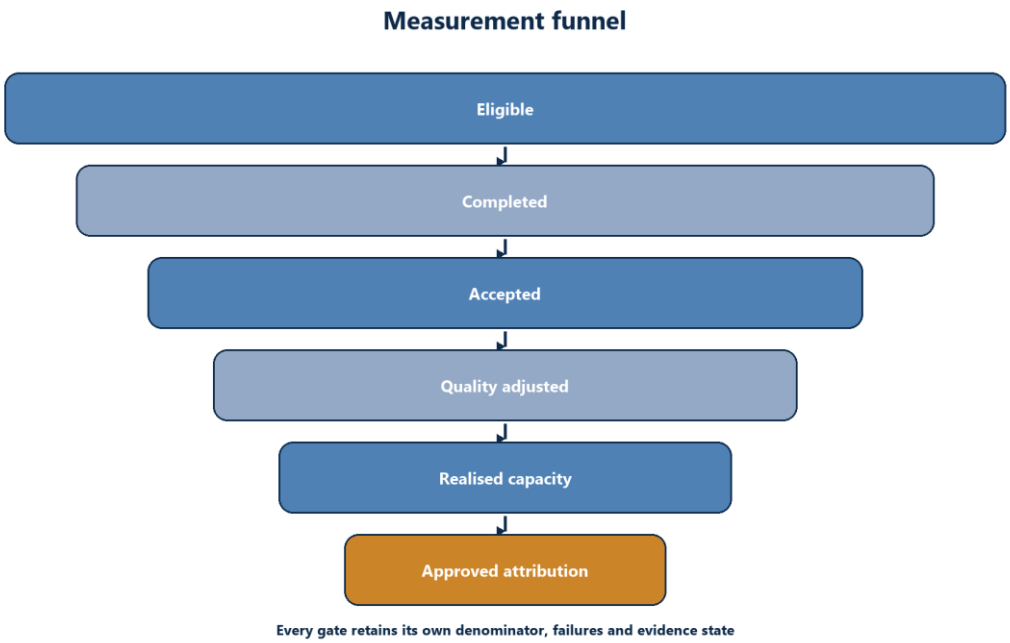

Figure 4. Measurement funnel from attempted packet to approved economic attribution
Source: Matchpoint measurement framework.

4.2 Baseline and treatment design

A credible pilot needs comparable tasks and enough observations to distinguish signal from transaction mix. Random assignment at packet level is strongest when contamination can be controlled. A matched crossover can be used where the same professional performs repeated work and withholding the tool is operationally acceptable. Team-level assignment may be necessary when collaboration makes packet-level contamination unavoidable.

Design element	Minimum specification	Reason
Population	Named teams, roles, locations and experience bands	Heterogeneous effects are expected [2,3]
Task classes	Predefined packet taxonomy and materiality	Prevents easy tasks entering only the treatment group
Comparator	Current approved workflow, versioned before pilot	Establishes what the tool is compared with
Assignment	Random, matched or staggered with documented method	Reduces selection bias
Start and stop	System timestamps plus active-time convention	Makes time definitions reproducible
Quality rubric	Pre-registered fields and defect severity	Prevents standards changing after results appear
Reviewer allocation	Balanced or independently sampled reviewers	Avoids reviewer leniency driving the result
Contamination	Record outside AI use and cross-team sharing	Protects comparator integrity
Failure capture	Include time-outs, fallbacks and rejected packets	Prevents survivorship bias
Analysis	Report median, distribution, confidence interval and cohort	Avoids an average masking dispersion

For a bounded first programme, the paper proposes a four-week baseline, a two-week shadow period and an eight-week controlled pilot. The duration is a proposed design, not a universal requirement. Baseline data should cover at least one complete cycle for the selected tasks. The shadow period runs the copilot without releasing its output,

which helps identify missing sources and control failures before production. The controlled pilot releases only accepted packets within the approved scope.

4.3 Metric dictionary

Metric	Numerator	Denominator	Decision use
Completion rate	Completed packets	Eligible attempted packets	Tool and workflow reliability
Acceptance rate	Accepted packets	Completed packets	Practical usefulness
First-pass acceptance	Accepted without material correction	Completed packets	Draft quality and review burden
Median cycle time	Median end-to-end elapsed time	Accepted packet	Service speed
Active producer time	Logged or sampled producer minutes	Accepted packet	Direct labour input
Reviewer time	Reviewer minutes	Accepted packet	Hidden transfer of effort
Material defect rate	Packets with material defect	Released packets	Decision and compliance risk
Citation validity	Claims supported by correct cited source	Claims requiring support	Evidence quality
Numeric consistency	Numeric fields passing deterministic checks	Numeric fields tested	Model and unit integrity
Exception latency	Time from exception creation to resolution	Resolved exception	Workflow friction
Realisation rate	Verified redeployed or removed hours	Gross measured hours released	Conversion to organisational value
Adoption rate	Active eligible users	Eligible users	Behavioural reach
Economic attribution rate	Approved attributed value	Modelled gross value	Discipline of value claims

Speed and quality should be reported side by side. Throughput is useful when demand exists, yet more output can include unnecessary work. A firm should define the service-level or decision need before treating volume as valuable.

4.4 From time to capacity and economics

Gross time release for task class i is the observed difference between comparator and treatment producer time for accepted, comparable packets, multiplied by accepted treatment volume. Reviewer and remediation time are then deducted. Realised capacity applies a verified reallocation factor:

Realised capacity hours = net measured hours released x verified reallocation rate.

Capacity value can be modelled using an approved loaded hourly cost or an opportunity-value method. It is not cost reduction unless payroll, contractor spend or another cost line actually changes and finance approves the attribution. Revenue attribution requires a defined causal mechanism, such as additional qualified opportunities worked, faster accepted response, increased win rate or more fee-bearing capacity. The model should retain the revenue field at zero until observed conversion and attribution data exist.

Net measured economic value = approved capacity value + approved incremental contribution + approved avoided external cost + approved loss reduction - licences - integration - governance - training - incremental review - incident and remediation cost.

Every term should show its evidence state: observed, approved model, unverified assumption or zero pending evidence. Mixing these states in one headline ROI figure is prohibited by this framework.

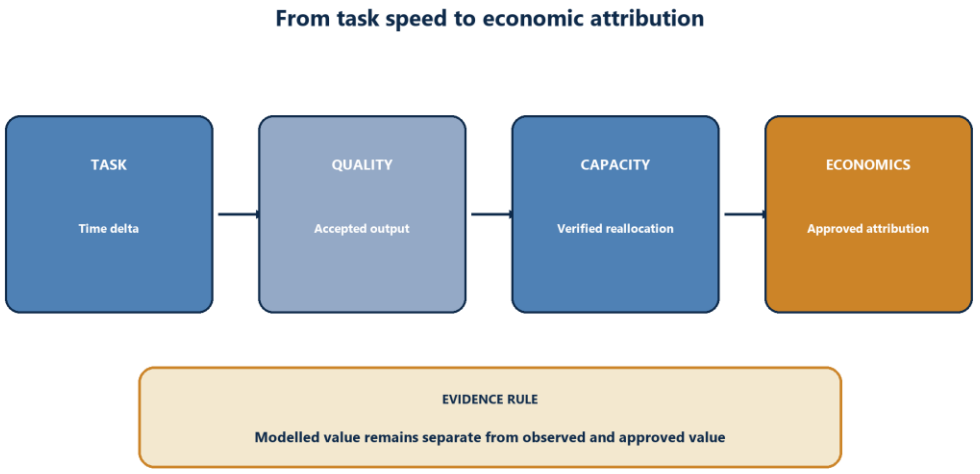

Figure 5. Productivity-to-value bridge with evidence gates
Source: Matchpoint measurement framework.

5. CONTROLLED ARCHITECTURE AND TOOL STACK

5.1 Design principle

The architecture starts with transaction identity, permissions and approved sources. The model is one component inside the workflow. Retrieval, calculations, policy checks, review and release remain separately observable. A user should be able to reconstruct which source versions, prompts, tools and validations produced an accepted packet.

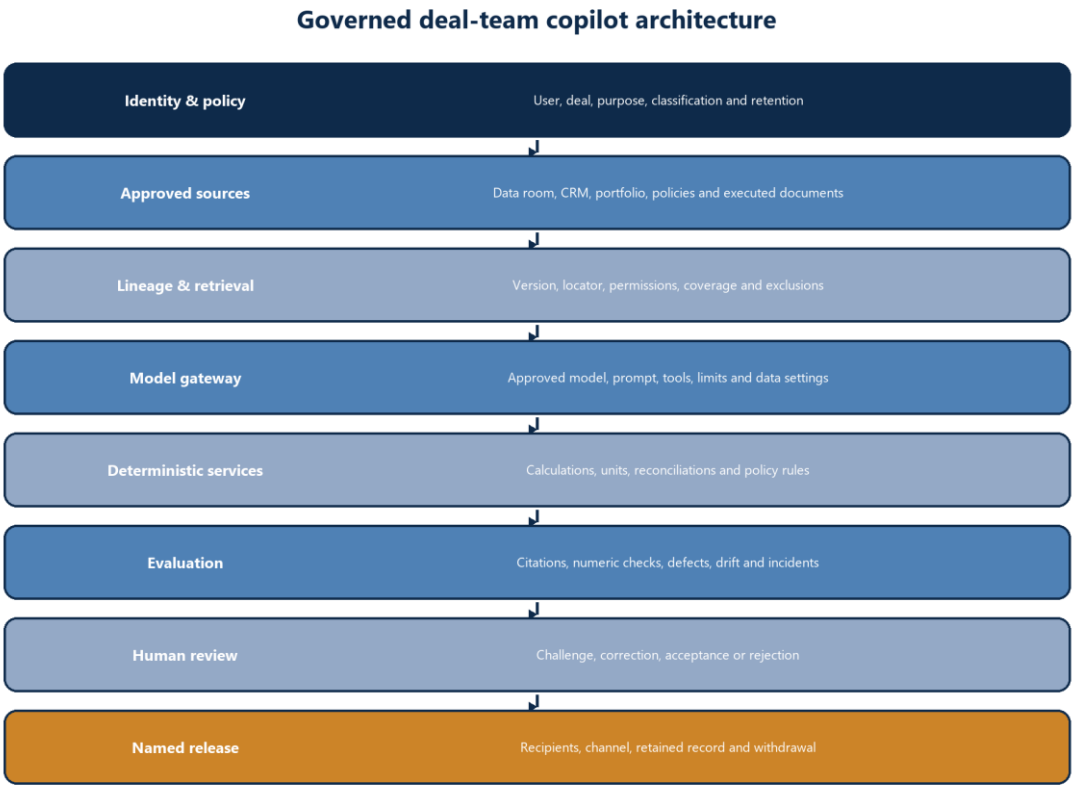

Figure 6. Governed copilot architecture for deal teams

Source: Matchpoint architecture; governance and risk references [8-12,14-22].

The proposed stack has eight layers:

1. **Identity and policy:** user, role, deal, client, entity, jurisdiction, data classification, permitted purpose and retention policy.
2. **Source layer:** approved virtual data room, CRM, portfolio system, research library, policies, executed agreements and financial models.
3. **Ingestion and lineage:** file hash, version, page, table, date, source owner, permission and supersession state.
4. **Retrieval:** permission-aware search that returns passages and exclusions, with a population or coverage test where completeness matters.
5. **Model gateway:** approved model and configuration, prompt template, token and tool limits, data-use settings and version pinning where available.
6. **Deterministic services:** calculations, unit checks, reconciliations, entity resolution, date logic and policy rules outside free-form generation.
7. **Evaluation and observability:** citations, field-level checks, exceptions, latency, cost, model changes, reviewer decisions and incidents.
8. **Human authority and release:** named reviewer, approval state, recipients, channel, retained record and rollback or withdrawal action.

Architecture layer	Failure exposure	Required control evidence
Identity and policy	Wrong user, deal or purpose	Authentication, role, deal membership, purpose and classification
Source layer	Unapproved, stale or incomplete evidence	Source register, owner, version, inclusion and exclusion record

Architecture layer	Failure exposure	Required control evidence
Ingestion and lineage	Lost page, table or version context	File hash, locator, extraction test and supersession rule
Retrieval	Missing or unauthorised context	Permission test, retrieval evaluation, population reconciliation
Model gateway	Model change, uncontrolled data use or tool access	Approved provider, terms, settings, version and allow-list
Deterministic services	Arithmetic, unit or entity error	Test cases, reconciliation, locked formula and exception result
Evaluation	Fluent error passes unseen	Rubric, sampled review, red-team case and drift threshold
Release	Wrong recipient or unapproved claim	Named authority, release gate, distribution record and retention

5.2 Source-grounded drafting

Source-grounded drafting requires more than attaching citations. Each factual proposition should link to the exact authoritative source version and locator used. The system should distinguish a source fact from a calculation, a model inference, a professional judgement and an unverified management assumption. A citation is invalid when the source exists but does not support the claim.

For completeness-sensitive work, retrieval must be tested as a population problem. If a task asks for every change-of-control clause across 120 contracts, finding plausible examples is inadequate. The packet should reconcile the expected contract population, processed population, unreadable files, excluded versions and unresolved exceptions. If completeness cannot be established, the output states that boundary before any summary.

The same rule applies to comparable companies, portfolio exposures, DDQ responses and regulatory obligations. Search ranking is a discovery mechanism. It is not proof of completeness.

5.3 Numerical control

The language model may explain an approved calculation. Material arithmetic should be performed by a deterministic spreadsheet, code service or controlled financial model with versioned inputs. The packet records units, currencies, dates, signs, scaling, source cells and reconciliation. Any number copied into narrative should be checked against the approved output.

The minimum numerical gates are:

- required input population reconciled;
- currency and unit explicit;
- formula version identified;
- balance, ownership or cash-flow checks passed where applicable;
- sensitivity assumptions separated from observed inputs;
- narrative numbers reconciled to tables and model outputs;
- reviewer confirms material adjustments and judgement.

This division keeps the copilot useful for explanation and exception routing while preserving the calculation engine as the numerical authority.

5.4 Confidentiality, privacy and security

Deal information can include personal data, material non-public information, commercial secrets, bank details, beneficial ownership and privileged advice. Each use case should therefore have an approved data boundary before

prompts or retrieval are enabled. Public consumer tools should not receive deal data unless the organisation has approved the exact service, contract, configuration, location, retention and use terms for that information.

NIST's GenAI Profile recommends governance and measurement across the AI lifecycle and discusses data privacy, information security, confabulation, information integrity, human-AI configuration and value-chain risks [15]. The IMF's 2023 note identifies privacy, bias, opacity, robustness, cybersecurity and potential systemic effects in finance [18]. Its 2026 work on AI and financial-sector cybersecurity emphasises third-party concentration, shared digital infrastructure and the potential for faster propagation of vulnerabilities [19]. These publications support controls for least privilege, data segregation, provider assessment, monitoring, recovery and exit.

Prompt injection and hostile document content require a specific boundary. Retrieved documents are untrusted data, not instructions. Tool calls should use typed allow-listed functions with scoped permissions. A document cannot authorise a payment, external message, model change, file deletion or recipient change. High-impact actions remain outside autonomous scope and require a separate human approval state.

5.5 Model and vendor change

A copilot can change without the workflow owner changing code. Provider model updates, retrieval changes, prompt edits, embedding changes and source-index rebuilds can affect results. The operating model therefore treats each significant component version as part of the packet evidence.

A release candidate should pass a fixed evaluation set containing ordinary cases, edge cases and known failures. Regression thresholds should cover citation validity, numeric consistency, material-defect rate, refusal or escalation, latency and cost. A model update enters production only after the owner accepts the evaluation and records any changed limitation. Provider concentration and exit plans should be reviewed in proportion to materiality.

6. ICP PLAYBOOKS

6.1 A2: Family-Office CIOs and Heads of Alternatives

The family-office use case is decision support across a private, often concentrated portfolio. The information set can span fund documents, direct investments, private banks, custodians, managers, operating businesses and family governance. Value comes from better evidence assembly, faster comparison and more consistent monitoring. The highest-risk failure is a fluent answer that obscures a missing source, stale valuation, conflict or concentration.

A2 workflow	Bounded copilot use	Core measure	Authority retained by people
Opportunity triage	Build cited summary, mandate fit and open-question list	Time to accepted screen; source coverage	CIO or delegate decides whether to proceed
Manager due diligence	Reuse approved DDQ library, compare versions, surface inconsistencies	First-pass acceptance; unresolved exceptions	Investment team evaluates manager and recommendation
Investment committee memo	Structure thesis, alternatives, risks, portfolio fit and evidence	Reviewer time; citation validity; material defects	Deal lead and committee own recommendation and decision
Portfolio monitoring	Extract notices, covenants and reported changes; route exceptions	Exception latency; coverage; false-negative sampling	Asset owner determines action and valuation implications
Board or family reporting	Draft from approved portfolio and decision records	Preparation time; numeric reconciliation	Authorised family-office leadership approves distribution

The family office should begin with repeatable packets where evidence is already controlled. A manager-update comparison or investment-memo source manifest is safer than autonomous portfolio advice. Portfolio construction, liquidity planning, related-party conflicts, tax, legal and succession conclusions require their relevant qualified authorities.

6.2 B2: GCC fund managers or GPs raising capital

The fund-manager use case combines investment activity with capital formation, operational due diligence and ongoing LP communication. The copilot can reduce duplication across DDQs, data rooms, quarterly letters, investment-committee materials and pipeline management. It can also increase risk if a stale answer, inconsistent metric or unsupported performance statement reaches an investor.

B2 workflow	Bounded copilot use	Core measure	Authority retained by people
DDQ and RFP response	Retrieve approved answer blocks, flag expiry and draft cited response	Cycle time; reuse; owner corrections	Functional owners approve each response
Data-room readiness	Classify documents, reconcile required index and flag missing items	Population coverage; exception closure	CFO, COO, counsel or authorised owner approves room
Fund-raising materials	Check consistency across strategy, team, track record and terms	Numeric consistency; claim support	Authorised principals and counsel approve marketing content
LP meeting preparation	Assemble relationship history, open questions and approved updates	Preparation time; source freshness	Relationship owner determines messaging
Investment committee	Produce evidence manifest, first draft and challenge list	Accepted-packet throughput; reviewer time	Investment professionals and committee decide
Quarterly reporting	Draft narrative from approved performance and portfolio records	Reconciliation; review time; late corrections	CFO, fund administrator and authorised signatory approve

FINRA's technology-neutral reminder and the SEC's AI-washing actions are US-specific publications, yet their control logic is useful more broadly: existing obligations and truthfulness remain relevant when AI is used [9,13]. A GCC manager should map the exact rules, fund documents, investor jurisdictions and marketing permissions with qualified advisers.

6.3 Shared service design

Both ICPs benefit from a shared packet schema:

- deal or fund identity;
- task class and materiality;
- requester, producer, reviewer and decision authority;
- approved source population and exclusions;
- facts, calculations, inferences and judgements in separate fields;
- open questions and contradictions;
- required legal, compliance, tax or specialist review;
- outcome, corrections and release state;
- timing, model, tool and cost telemetry;
- retained evidence and disposal rule.

The schema allows a small team to accumulate reusable knowledge without turning prior text into an ungoverned source. Reuse occurs through approved answer blocks, policies, templates and structured facts with owners and expiry dates.

7. ILLUSTRATIVE ECONOMICS

7.1 Scenario status

[Unverified illustrative management assumptions] The scenarios in this section are worked examples. They are not observed Matchpoint, family-office, fund-manager or market data. They do not represent a forecast, promise, benchmark or quoted vendor price. Each organisation should replace every input with approved observed evidence.

The illustrative team has 12 professionals, 220 working days per year and 8 paid hours per day, producing 21,120 annual paid hours. The loaded hourly value is assumed at AED 650. The model separates task eligibility, direct task-time reduction, active adoption, quality acceptance and verified reallocation. Annual licences, integration, governance, training and incremental review are grouped as programme cost. No revenue, cost reduction or loss reduction is attributed.

Assumption	Conservative	Base	Upside	Evidence state
Annual paid hours	21,120	21,120	21,120	Unverified illustrative assumption
Share of time in eligible packets	25%	35%	45%	Unverified illustrative assumption
Direct task-time reduction	10%	20%	30%	Unverified illustrative assumption
Active adoption or coverage	60%	75%	85%	Unverified illustrative assumption
Quality acceptance factor	80%	90%	95%	Unverified illustrative assumption
Verified reallocation factor	50%	65%	75%	Unverified illustrative assumption
Loaded value per realised hour	AED 650	AED 650	AED 650	Unverified illustrative assumption
Annual programme cost	AED 250,000	AED 310,000	AED 390,000	Unverified illustrative assumption

7.2 Worked outputs

Modelled output	Conservative	Base	Upside
Eligible annual hours	5,280	7,392	9,504
Direct gross hours released	528	1,478	2,851
Hours after adoption and quality	253	998	2,302
Realised capacity hours	127	649	1,727
Modelled realised capacity value	AED 82,368	AED 421,621	AED 1,122,388
Less annual programme cost	AED 250,000	AED 310,000	AED 390,000
Modelled net capacity value	AED (167,632)	AED 111,621	AED 732,388
Attributed revenue	USD 0	USD 0	USD 0
Attributed cost reduction	USD 0	USD 0	USD 0
Quantified loss reduction	USD 0	USD 0	USD 0

Rounding explains small differences between displayed intermediate values and calculated totals. The model shows why a percentage speed gain is insufficient. Eligibility, adoption, quality and reallocation compound. In the conservative case, the programme has negative modelled net capacity value. The base case becomes positive only after a 65% reallocation assumption. The upside case depends on broad eligibility and strong adoption that require evidence. No scenario should be presented externally as an achieved ROI.

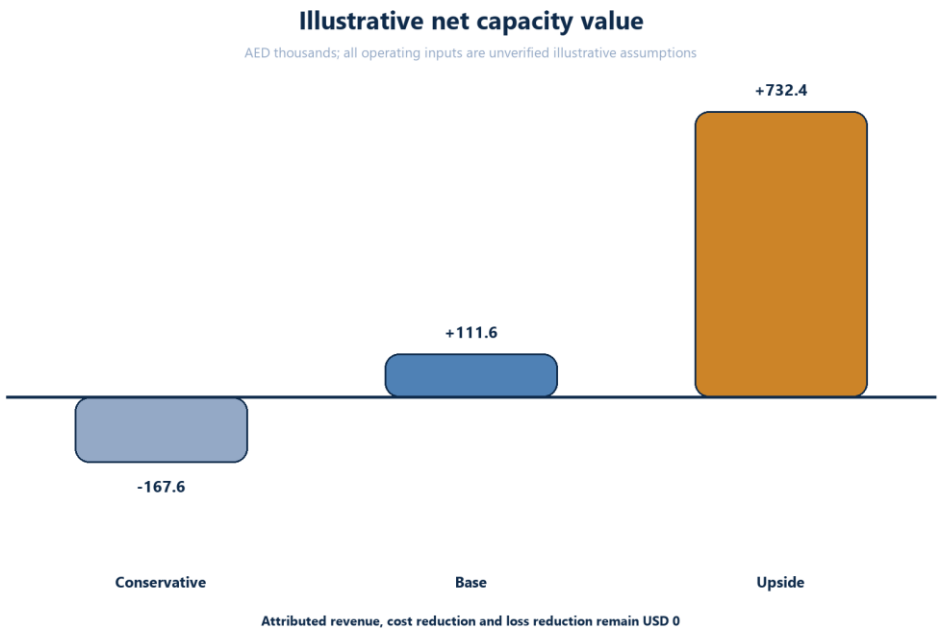

Figure 7. Illustrative productivity and economics under three scenarios

Source: Matchpoint scenario analysis. All operating inputs are unverified illustrative management assumptions.

7.3 Revenue channel

The Topic Tracker hook asks how the technology can multiply productivity and revenue. The proposed revenue logic is a measurement route:

Released qualified capacity -> additional or faster qualified work -> observable client or investment outcome -> approved contribution attribution.

For a fund manager, observable intermediate measures could include complete qualified LP follow-ups, approved DDQ response time, accepted data-room readiness and additional substantive meetings handled without a fall in quality. For a family office, useful measures could include additional screened opportunities, faster completion of accepted diligence packets and earlier closure of material exceptions. These are operational measures. Revenue or investment performance requires a further causal link and should remain unclaimed until it is demonstrated.

An attribution design can compare contemporaneous matched cohorts, staggered rollout or pre-specified changes in eligible throughput. It should control for market conditions, team seniority, transaction mix, fund-raising stage, seasonality and parallel process changes. Management approval of an estimate does not convert it into observed fact; the dashboard should retain its evidence label.

8. RISK, GOVERNANCE AND ACCEPTANCE

8.1 Risk register

Risk	Leading indicator	Gate or response	Accountable owner
Unsupported factual claim	Citation-validity failure	Block release; correct source or remove claim	Packet reviewer
Numeric inconsistency	Reconciliation or unit failure	Re-run deterministic check; block narrative	Model owner
Missing document population	Expected and processed counts differ	State incompleteness; route exception	Data-room owner

Risk	Leading indicator	Gate or response	Accountable owner
Confidential-data exposure	Policy or data-loss-prevention alert	Stop processing; contain and investigate	Information-security owner
Prompt injection or tool abuse	Untrusted instruction or unexpected tool request	Isolate content; deny action; review incident	AI platform owner
Automation bias	Falling reviewer time with rising correction or incident rate	Increase independent sample; retrain or pause	Business owner
Model drift	Evaluation threshold breached after component change	Roll back or reapprove	Model owner
Stale reusable content	Owner or expiry missing	Remove from retrieval until renewed	Knowledge owner
Misleading AI or performance claim	Unsupported external statement	Block release; legal or compliance review	Communications owner
Vendor concentration or outage	Service-level or exit trigger	Use tested fallback; review continuity	Technology and procurement owners
Recordkeeping gap	Missing prompt, output, approval or distribution record	Block or reconstruct before release	Compliance or records owner
Unproved economic value	Modelled value presented as achieved	Correct label; keep attribution at zero	Finance or management authority

8.2 Acceptance gates

A use case enters controlled production when its owner accepts all applicable gates:

1. **Purpose:** the business purpose, user population and prohibited uses are explicit.
2. **Data:** sources, permissions, personal data, confidentiality, locations and retention are approved.
3. **Evidence:** retrieval and population tests meet the threshold for the task.
4. **Quality:** citation, numeric and material-defect thresholds pass on the fixed evaluation set.
5. **Security:** prompt-injection, data-exfiltration, access and tool-abuse tests pass.
6. **Human authority:** reviewers, decision owners and escalation routes are named.
7. **Operations:** latency, availability, cost, fallback, incident and rollback procedures are tested.
8. **Measurement:** comparator, assignment, metrics and analysis are pre-specified.
9. **Release:** recipient, channel, recordkeeping and disclosure requirements are enforced.
10. **Economics:** assumptions and observed values remain separate; attribution authority is named.

The owner may accept a residual risk within delegated authority. Material residual risk outside that authority is escalated. An expired or failed gate returns the use case to shadow mode or manual fallback.

8.3 Governance cadence

Daily operations review failed packets, security alerts and blocked releases. Weekly product review examines exceptions, task-level performance, user feedback and evaluation drift. Monthly business review examines quality-adjusted productivity, realised capacity, cost and incidents by task and cohort. Quarterly governance review confirms use-case inventory, provider dependencies, legal and regulatory change, data boundaries, control effectiveness and continued authority.

The Bank/FCA survey reported that 84% of responding firms had an accountable person for their AI framework and that accountability was often distributed across several persons or bodies [8]. The proposed cadence reflects that operating reality while requiring a single named owner for each use case and packet.

9. ADOPTION ROADMAP

9.1 Days 0 to 30: define and baseline

- appoint business, technology, data, security and compliance owners;
- select two or three repeatable green or amber packet classes;
- define prohibited data, actions and releases;
- version the current manual workflow and quality rubric;
- collect baseline producer, reviewer, quality and failure data;
- build the approved source and reusable-content register;
- create the fixed evaluation set and incident taxonomy.

The output is an approved pilot charter, baseline dataset and go or no-go decision for shadow operation.

9.2 Days 31 to 60: build and shadow

- connect only approved sources using least privilege;
- implement packet identity, lineage, deterministic checks and reviewer workflow;
- run retrieval, completeness, prompt-injection and numerical tests;
- operate in shadow mode with no external release;
- compare copilot packets with the manual output;
- repair source, workflow and evaluation failures;
- approve the controlled-pilot design and user training.

The output is an evaluation report, control record and explicit production-scope decision.

9.3 Days 61 to 90: controlled pilot

- assign eligible packets according to the pre-specified design;
- capture every attempt, failure, fallback, review and correction;
- hold daily incident and weekly quality reviews;
- keep external and investment authority with named professionals;
- measure task and cohort distributions, not only averages;
- publish an internal result with confidence intervals and limitations;
- accept, amend, pause or retire each use case.

The output is an observed task-level productivity and quality result. Economic attribution remains separate.

9.4 Months 4 to 12: scale evidence, not access alone

Scaling adds use cases only after the existing workflow remains within quality and security thresholds. Reusable knowledge receives owners and expiry dates. Model and retrieval changes pass regression tests. Realised-capacity evidence is reviewed with team planning, and finance approves any cost or revenue attribution. The firm maintains a manual fallback and tested exit path for material provider dependency.

Phase	Decision	Required evidence	Stop condition
Define	Is the packet suitable and valuable?	Purpose, baseline, task frontier, data boundary	No owner, no comparator or prohibited data
Shadow	Can the system produce a reviewable packet?	Evaluation results, lineage, controls, fallback	Material defect or security threshold breached

Phase	Decision	Required evidence	Stop condition
Pilot	Does quality-adjusted productivity improve?	Assigned observations and full denominators	Reviewer burden or failure cost erases benefit
Scale	Can the benefit survive broader use?	Cohort results, realised capacity, operations	Drift, concentration, incident or control failure
Attribute	Did economics change for the stated reason?	Finance-approved causal and accounting evidence	Model estimate presented as observed value

10. LIMITATIONS AND RESEARCH AGENDA

The external productivity studies use different tasks, populations, tools, time periods and outcomes. Their reported percentages are not directly comparable and should not be pooled into a deal-team forecast. Tool capabilities and user behaviour also change quickly. Model updates can invalidate a prior benchmark, while learning can alter both comparator and treatment performance.

Deal work has additional measurement problems. Transaction complexity varies; difficult cases may select into human-only treatment; team members share knowledge; quality can become observable only after closing; and reviewers may change effort when they know AI was used. Revenue and investment outcomes are affected by market conditions, client relationships, capital availability and judgement. A short pilot can measure packet quality and time more credibly than long-horizon investment performance.

The illustrative economics use unverified assumptions and exclude taxes, financing, depreciation, opportunity costs outside the stated capacity method and tail losses. A negative or positive modelled result is sensitive to the eligible-time, quality and reallocation assumptions. Actual decision-makers should perform sensitivity analysis with local data.

The framework also requires legal, regulatory, tax, data-protection, employment, intellectual-property and contractual analysis for the specific firm and use case. Official publications cited in this paper do not constitute permission, approval or a complete obligation map.

Future research should publish task-level experiments for investment and capital-formation workflows, including independent quality ratings, reviewer time, error severity and longitudinal reallocation. Studies should compare source-grounded copilots, deterministic tools and general chat interfaces; report negative results; preserve evaluation sets; and examine whether gains persist after task novelty and training effects fade.

11. CONCLUSION

An AI copilot can accelerate bounded deal work when the task is verifiable, the source set is approved, deterministic calculations remain authoritative and a named professional controls release. External studies show material gains in several knowledge-work settings and material losses in another. That dispersion makes measurement a governance requirement.

For family-office investment teams, the strongest starting points are evidence assembly, manager-diligence comparison, monitoring exceptions and investment-memo support. For GCC fund managers, the strongest starting points are DDQ reuse, data-room readiness, consistency checking, LP preparation and controlled committee drafting. Valuation, investment decisions, conflicts, legal conclusions and autonomous external actions remain with authorised professionals.

The proposed productivity premium is quality-adjusted accepted output per paid hour. Its measurement includes producer time, reviewer time, remediation, defects, failures and fallbacks. Released time becomes capacity value only after verified reallocation. Revenue, cost reduction and loss reduction become attributed only after an approved causal and accounting basis exists.

The practical sequence is clear: define the task frontier, establish the baseline, build a source-grounded controlled stack, run shadow evaluation, conduct a bounded pilot, publish the full result and scale only the use cases whose evidence survives. That process can turn a technology claim into an operating fact.

DECLARATIONS

Authorship. Chennakeshav (CK) Adya is identified as the author. Matchpoint Partners is the publisher of this working paper.

Author scenarios. No unverified personal anecdote or client result is presented as an authored experience. The economic scenarios are labelled unverified illustrative management assumptions.

Conflicts of interest. No external financial sponsorship for this paper has been identified in the materials provided for production. This statement has not been independently audited.

Data availability. The paper uses public sources and illustrative calculations. No confidential client dataset or observed Matchpoint productivity dataset is used.

Use of generative AI. Generative AI assisted research organisation, drafting, editing, figure production and quality checks. The named author retains responsibility for review, approval and publication.

Advice boundary. This paper is general research for professional audiences. It is not investment, legal, regulatory, tax, accounting, valuation, cybersecurity, data-protection, employment or technology-procurement advice. It is not an offer, solicitation, recommendation or promise of results.

REFERENCES

[1] Noy, S. and Zhang, W. (2023). "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654), 187-192. https://doi.org/10.1126/science.adh2586

[2] Brynjolfsson, E., Li, D. and Raymond, L. R. (2023, revised 2023). "Generative AI at Work." NBER Working Paper 31161. https://doi.org/10.3386/w31161

[3] Dell'Acqua, F. et al. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013. https://www.hbs.edu/ris/download.aspx?name=24-013.pdf

[4] Dell'Acqua, F. et al. (2025; published version 2026). "The Cybernetic Teammate: A Field Experiment on Generative AI and Teamwork." NBER Working Paper 33641; *Organization Science*, online June 2026. https://doi.org/10.3386/w33641

[5] Dillon, E. W., Jaffe, S., Immorlica, N. and Stanton, C. T. (2025, revised 2025). "Shifting Work Patterns with Generative AI." NBER Working Paper 33795. https://doi.org/10.3386/w33795

[6] Becker, J., Rush, N., Barnes, B. and Rein, D. (2025). "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." METR. https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf

[7] Becker, J., Rush, N., Cunningham, T., Rein, D. and Mahamud, K. (2026). "We are Changing our Developer Productivity Experiment Design." METR, 24 February 2026. https://metr.org/blog/2026-02-24-uplift-update/

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

[9] 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

[10] Financial Industry Regulatory Authority (2020). "Artificial Intelligence (AI) in the Securities Industry." https://www.finra.org/rules-guidance/key-topics/fintech/report/artificial-intelligence-in-the-securities-industry

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

[12] European Securities and Markets Authority (2024). "Public Statement on the use of Artificial Intelligence in the provision of retail investment services." ESMA35-335435667-5924. https://www.esma.europa.eu/sites/default/files/2024-05/ESMA35-335435667-5924__Public_Statement_on_AI_and_investment_services.pdf

[13] 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

[14] 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

[15] 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

[16] UAE Government (2021, portal page current at research cut-off). "Data protection laws." Federal Decree-Law No. 45 of 2021 Regarding the Protection of Personal Data. https://u.ae/en/about-the-uae/digital-uae/data/data-protection-laws

[17] European Union (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/eng

[18] Shabsigh, G. and Boukherouaa, E. B. (2023). *Generative Artificial Intelligence in Finance: Risk Considerations.* IMF FinTech Note 2023/006. https://www.imf.org/-/media/files/publications/ftn063/2023/english/ftnea2023006.pdf

[19] Adrian, T., Gaidosch, T. and Ravikumar, R. (2026). *Artificial Intelligence and Cybersecurity in the Financial Sector.* IMF Note 2026/005. https://www.elibrary.imf.org/view/journals/068/2026/005/article-A001-en.xml

[20] Organisation for Economic Co-operation and Development (2024). "OECD AI Principles." Updated May 2024. https://oecd.ai/en/ai-principles

[21] 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

[22] National Institute of Standards and Technology (2024). *Cybersecurity Framework 2.0.* NIST CSWP 29. https://doi.org/10.6028/NIST.CSWP.29

[23] GitHub (2022, updated 2024). "Research: quantifying GitHub Copilot's impact on developer productivity and happiness." Vendor-conducted study. https://github.blog/news-insights/research/research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

[24] Nikolaev, V. (page current at research cut-off). "Ongoing Research Projects." University of Chicago Booth School of Business; includes notice concerning withdrawal of "Financial Statement Analysis with Large Language Models." https://faculty.chicagobooth.edu/valeri-nikolaev/ongoing-research-projects

APPENDIX A. PILOT CHARTER

Field	Required entry
Business outcome	Specific service, quality or capacity decision to test
Use-case owner	Named accountable executive
Packet class	Exact repeated task and materiality
Eligible users	Named roles, teams and experience bands
Comparator	Versioned approved current workflow
Data boundary	Approved sources, personal data, confidentiality and prohibited content
Authority boundary	Decisions and actions the copilot cannot take
Assignment	Random, matched, crossover or staggered method
Primary metric	Quality-adjusted accepted output per paid hour
Secondary metrics	Speed, review, defects, failures, adoption, cost and reallocation
Quality rubric	Pre-specified fields, weights and material-defect definition
Evaluation set	Ordinary, edge, adversarial and known-failure cases
Stop conditions	Security, quality, availability, legal or cost threshold
Review cadence	Daily, weekly, monthly and final decision forums
Economic authority	Named approver for capacity, cost, revenue and loss attribution

APPENDIX B. MINIMUM PACKET RECORD

1. Packet identifier, deal or fund identifier and task class.
2. Requester, producer, reviewer, decision owner and release authority.
3. Start, completion, review and release timestamps.
4. Model, prompt, retrieval index and tool versions.
5. Approved source manifest, exact locators, exclusions and unresolved population gaps.
6. Extracted facts, deterministic calculations, model inferences, professional judgements and assumptions in distinct fields.
7. Citation checks, numerical reconciliations, policy rules and security results.
8. Reviewer corrections, defect severity and acceptance state.
9. Release version, recipients, channel and retained record.
10. Producer, reviewer, remediation and fallback time.
11. Model, retrieval, storage and external-service cost.
12. Incident, withdrawal, correction and post-release outcome where applicable.

APPENDIX C. PILOT ACCEPTANCE CHECKLIST

- Purpose, users, data and prohibited uses are approved.
- Manual comparator and baseline period are versioned.
- Task classes and assignment method are fixed before treatment.
- Required source population and retrieval tests pass.
- Facts, calculations, inferences and judgements remain distinct.
- Material calculations use deterministic services and reconcile.

- Citation-validity and material-defect thresholds pass.
- Prompt-injection, data-exfiltration and tool-abuse tests pass.
- Named reviewers and decision authorities are active.
- Failures, fallbacks and rejected packets remain in the denominator.
- Producer, reviewer and remediation time are captured.
- Realised-capacity evidence is separate from gross time saved.
- Revenue, cost and loss attribution remain zero until approved evidence exists.
- Manual fallback, rollback, incident and provider-exit procedures are tested.
- Final results report task, cohort, distribution, confidence interval and limitations.

APPENDIX D. GLOSSARY

Accepted packet. A bounded output approved by the named reviewer after required controls.
AI copilot. An assistive AI system operating inside a human-owned workflow without delegated investment authority under this framework.
Comparator. The approved workflow against which the treatment is measured.
Confabulation. Generated content that is false or unsupported while appearing plausible; NIST uses the term within its GenAI risk profile [15].
Deterministic service. A calculation or rule engine expected to return the same result for the same versioned inputs.
Eligible packet. A task within the approved pilot taxonomy and data boundary.
First-pass acceptance. Acceptance without material correction.
Material defect. An error capable of altering a decision, disclosure, recipient action or compliance outcome.
Modelled value. A calculated scenario output based partly or wholly on assumptions; it is not an observed or attributed result.
Productivity premium. Relative improvement in quality-adjusted accepted output per paid hour against the approved comparator.
Realised capacity. Verified net time released and demonstrably redeployed or removed.
Source manifest. The versioned record of sources included, excluded and unresolved for a packet.
Task frontier. The changing boundary between tasks that a configured AI workflow can perform acceptably and tasks that require a different method or direct human work.
Treatment. The governed copilot workflow tested against the comparator.