AI Business Review

Journal homepage: https://theaibr.com/index.php/aibr

Collaborative Intelligence in Accounting: A Human + AI Complementarity Framework for Professional Work

Cory Ng *

1 Department of Accounting & Information Systems, Villanova University, Villanova, PA, 19085, cory.ng@villanova.edu

* Correspondence: cory.ng@villanova.edu

Abstract:

Background: Public discussion of generative artificial intelligence (AI) in accounting often swings between the allure of full automation and job displacement anxiety, yet the most immediate reality in organizations is human + AI work: AI accelerates drafting, summarization, and pattern detection while professionals remain accountable for judgment, materiality, and defensibility in financial reporting and analysis.

Methods: This paper synthesizes research and practitioner guidance to develop a practical model for designing human AI collaboration, sometimes described as collaborative intelligence, in the financial reporting function (often referred to as controllership), including period end close, financial statement preparation, variance explanation, management reporting narratives, and accounting policy documentation.

Results: This paper develops the C³ Framework, Complementarity, Controls, and Competencies, which maps accounting tasks by task structure and judgment/materiality to recommend collaboration modes. It also specifies five mandatory control points for high judgment use cases (source grounding and traceability, independent verification and tie out, contradiction testing, escalation and approval, and audit trail logging), and proposes a role taxonomy that clarifies review responsibility, escalation thresholds, and evidence retention.

Conclusions: The C³ Framework provides implementable design patterns and testable propositions intended to help accounting leaders capture productivity gains from human + AI work while preserving accountability, consistency, and alignment with governance expectations in high stakes reporting contexts.

Keywords: Generative AI; large language models (LLMs); financial reporting; AI risk management; human–AI collaboration; AI governance

Academic Editor(s): Juan Dempere
Received: January 6, 2026
Revised: April 2026
Accepted: May 8, 2026
Published: date: May 8, 2026
Citation: Ng, C. (2026) Collaborative Intelligence in Accounting: A Human + AI Complementarity Framework for Professional Work.

AI Business Review, *2*(1)10-29.
https://doi.org/10.64044/vpm19y22

1. Introduction

Accounting and finance teams increasingly confront a paradox: generative AI can draft variance explanations, summarize reporting packages, and support elements of the month-end close. At the same time, these same capabilities raise new requirements for defensibility, including clear traceability to the enterprise resource planning (ERP) and general ledger data, verification that figures and claims match the ledger, adherence to company accounting policies, and explicit professional ownership of decisions. For controllership teams, this is not an abstract concern. Financial close and reporting processes rely on trust in numbers, narratives, and explanations that shape executive decisions, resource allocation, and stakeholder confidence.

Recent field evidence suggests these changes are already observable in practice. Choi and Xie (2025) combine survey evidence from 277 accountants with transaction-level field data from 79 small and mid-sized firms using AI-enabled accounting software, documenting both productivity gains and meaningful task reallocation. They find that AI use is associated with reallocating 8.5% of accountant time away from routine data entry toward higher-value activities and with measurable improvements in reporting outcomes, including a 12% increase in general ledger granularity and a 7.5-day reduction in monthly close time.

The same evidence also underscores why "human + AI" work cannot be treated as a simple speed upgrade. Accountants report concerns about AI-generated errors and accuracy, data security, and job displacement. Choi and Xie (2025) document situations in which users over-rely on inaccurate AI-suggested classifications, an error pathway that becomes more consequential as AI-generated outputs appear increasingly fluent and authoritative. In other words, AI is not only accelerating existing tasks; it is reshaping how accounting professionals interact with information and make judgments, increasing the importance of workflow design, verification behaviors, and clear lines of accountability.

This pattern aligns with broader empirical evidence that generative AI functions primarily as an augmentation tool rather than a full substitute for professionals. In a preregistered experiment on professional writing tasks, access to ChatGPT increased productivity and improved average output quality (Noy & Zhang, 2023). Complementing that evidence, a large-scale field study of more than 5,000 customer-support agents using a generative-AI assistant found average productivity gains, but also substantial variation across workers: less experienced employees benefited the most, while gains for highly experienced employees were smaller and sometimes accompanied by quality tradeoffs (Brynjolfsson, Li, & Raymond, 2025). Together, these studies help explain why "human + AI" performance is not automatic; the size and reliability of benefits depend on task design, user expertise, and workflow guardrails that preserve accountability.

This creates a practical challenge for accounting work that is both quantitative and narrative. Financial reporting and analysis increasingly blends numbers with explanations, monthly variance narratives, management reporting commentary, accounting policy memos for nonroutine transactions, and documentation that must withstand internal review and (in many settings) audit or regulator scrutiny. Generative AI can accelerate drafting and pattern detection, but it can also amplify risk when fluent outputs are mistaken for verified conclusions. What changes is not only speed. It is how judgment is produced and defended, how professionals gather evidence, test competing explanations, apply materiality thresholds, and establish clear decision ownership and accountability of the final conclusion.

While accounting and information systems research is rapidly expanding on these topics, the knowledge base remains fragmented. A recent scoping review of generative AI research in accounting and finance identifies growing work on applications, research methods, and professional implications, while emphasizing that many important questions remain open, especially around how organizations should implement Large Language Models (LLMs) in high-stakes workflows (Dong, Stratopoulos, & Wang, 2024). At the same time, governance standards increasingly emphasize accountability, transparency, and risk management across the AI lifecycle. These expectations push organizations toward clearer documentation, controls, and role definitions (e.g., National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF); International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 42001). The practical gap is that "human-in-the-loop" is often invoked as a general principle. Accounting leaders need more operational guidance: which tasks should be hybrid, what control points are nonnegotiable, and who is responsible for what when AI contributes to reporting outputs.

This paper addresses that gap by developing a practical model for designing human–AI collaboration in financial reporting and analysis.This operating model is described as collaborative intelligence. Specifically, this paper develops the C³ Framework: Complementarity (how to allocate tasks between humans and AI based on structure and judgment/materiality), Controls (workflow gates that make AI-supported outputs defensible), and Competencies (the review behaviors, skills, and role clarity required for reliable adoption). The framework is accompanied by two implementation artifacts (Table 1 and Exhibit 1) and testable propositions intended to guide future empirical research.

2. Research Gap and Objective

Existing discussion about AI in accounting frequently emphasizes broad transformation narratives (e.g., automation of routine tasks, shifting talent needs) but is less explicit about when and why human–AI collaboration outperforms either humans-alone or AI-alone in specific accounting workflows, particularly within the accounting function across the month-end close, reconciliations, journal entries, variance analysis, and management reporting. The lack of a usable framework can leave organizations oscillating between two extremes: (1) restricting AI to low-stakes tasks only, or (2) adopting AI widely without sufficient controls, thereby increasing downstream risk.

2.2 Contributions

This paper contributes a practice-oriented conceptual model—the C³ Framework for Accounting Human–AI Collaboration—that:

1. classifies controllership tasks by their structure and judgment/materiality to recommend appropriate collaboration modes (Complementarity),

2. specifies five "mandatory control points" for high-judgment close and reporting use cases (Controls), and

3. identifies competencies and emerging roles required to make hybrid work reliable (Competencies).

This paper also provides two implementation artifacts (Table 1 and Exhibit 1) designed to enable near-term adoption, and articulates testable propositions to guide future empirical work.

2.3 Thesis

Human+AI outperforms humans-alone or AI-alone when the work is judgment- or language-intensive (e.g., policy narratives, variance explanations, and management reporting commentary), the workflow includes designed control points (verification, traceability, and escalation), and responsibilities are intentionally divided so that AI contributes scale, drafting, and pattern detection while professionals retain judgment, materiality assessment, critical evaluation, and accountability for the final decision.

3. Method

3.1 Review Approach

This paper performed a structured narrative review of sources published 1984-2025 using Google Scholar and publisher databases for academic literature, along with major professional bodies and practitioner outlets for applied guidance. The goal was synthesis rather than comprehensive meta-analysis: to integrate multidisciplinary insights into an accounting-operational framework suitable for financial reporting teams.

3.2 Inclusion and Exclusion Criteria

Sources were included if they addressed human–AI collaboration and complementarity (including augmentation/centaur approaches), generative AI use in accounting and finance professional work (with emphasis on close, reporting, and governance-related tasks), and/or enterprise AI governance frameworks and standards relevant to organizational adoption (including the NIST AI RMF and ISO/IEC 42001).

This paper excluded sources focused purely on model architecture without organizational implications and sources with no clear link to accounting/finance practice or governance.

3.3 Synthesis Procedure

This paper organized insights from the reviewed sources into three themes aligned with the framework:

(1) Task Complementarity: Where hybrid approaches add value and how tasks differ.

(2) Workflow Controls: What makes outputs defensible and accountable.

(3) Competencies and Roles: What professionals must do differently when AI contributes to reporting outputs.

These themes informed the development of the C^3 Framework, and the paper proposes testable propositions for future research.

4. Results: The C³ Framework for Human–AI Collaboration in Controllership, Close, and Reporting

This section presents the C³ Framework—a practice-oriented model for designing human–AI collaboration in accounting professional work, with a focus on controllership activities (record-to-report, close orchestration, management reporting, and financial narrative production). The framework integrates three layers that must align for hybrid performance ("human + AI") to exceed outcomes achieved by humans alone or AI alone. C¹ Complementarity focuses on task design and the selection of collaboration modes based on task structure and judgment/materiality. C² Controls specifies risk and accountability mechanisms through "mandatory control points" that make AI-supported work defensible and governable. C³ Competencies defines the human responsibilities, review behaviors, and emerging roles needed to operationalize collaborative intelligence in accounting and financial reporting workflows.

4.1 C¹ Complementarity: a 2×2 Task Typology for the Close and Reporting Cycle

Human–AI collaboration works best when teams deliberately allocate responsibilities (Jarrahi, 2018; Raisch & Krakowski, 2021; Leitner-Hanetseder et al., 2021), using machines for speed, drafting, summarizing, and pattern detection, while accounting professionals retain ownership of judgment, materiality assessment, context, and accountability (Saghafian & Idan, 2024; Hemmer et al., 2025). This paper operationalizes that allocation using a 2×2 typology based on task structure and judgment/materiality.

Table 1.

Task Structure × Judgment/Materiality Typology for Close & Reporting (with recommended collaboration modes). Materiality is used here in the practical sense of stakes: the potential consequence of error for reporting decisions, stakeholder reliance, or compliance/audit scrutiny.

Judgment/Materiality ↓ \ Task Structure →	Structured (rules-based, repeatable)	Unstructured (narrative, ambiguous, context-dependent)
Low judgment / low materiality	Quadrant A: Automate with exception handling. Examples: transaction coding; routine reconciliations; standardized close checklist status reporting; basic roll-forwards. Hybrid mode: AI/automation executes; humans monitor exceptions and fix upstream data issues. Primary risks: silent input errors; over-automation in boundary cases. Minimum controls: exception thresholds; reasonableness tests; ownership of exception queues; periodic sampling review.	Quadrant C: AI drafts; humans edit. Examples: close meeting notes; SOP drafts; internal documentation narratives; first drafts of low-stakes internal reporting summaries. Hybrid mode: AI produces first draft; humans tailor to context and ensure clarity. Primary risks: inconsistency across periods; "polished but wrong" statements (usually low impact here). Minimum controls: standard templates; approved terminology; prompt library; light review checklist.
High judgment / high materiality	Quadrant B: AI assists; humans decide. Examples: structured revenue recognition triage with exceptions;	Quadrant D: Co-pilot mode (hybrid wins only with controls). Examples: flux analysis narratives for executive/board reporting; draft accounting policy

Judgment/Materiality ↓ \ Task Structure →	Structured (rules-based, repeatable)	Unstructured (narrative, ambiguous, context-dependent)
	impairment indicator screens; reserve/estimate support packages; tax position triage with nuanced facts. Hybrid mode: AI organizes evidence and flags anomalies; humans make judgment calls and sign off. Primary risks: overreliance on plausible rationales; missing contextual factors. Minimum controls: traceability header; tie-outs for key claims; contradiction-testing step; escalation thresholds.	memos for complex transactions; explanations for unusual entries or reclasses; high-stakes stakeholder narratives. Hybrid mode: AI generates drafts + alternative explanations; humans verify, apply materiality judgment, escalate, and approve. Primary risks: confident narrative errors; causal overreach; misinterpretation of policy; accountability diffusion. Minimum controls: all five mandatory control points (grounding/traceability; verification/tie-out; contradiction testing; escalation/approval; audit-trail logging).

As shown in Table 1, the two dimensions are defined as follows:

Task structure (Structured ↔ Unstructured): rules-based and repeatable versus narrative and context-dependent.

Judgment/materiality (Low ↔ High): routine, low-stakes outputs versus outputs with meaningful reporting, stakeholder, or compliance consequences.

Structured tasks generally follow stable rules or checklists, while unstructured tasks require explanation, interpretation, or narrative judgment (Gorry & Scott Morton, 1971; Autor et al, 2003; Bonner, 1994). The judgment/materiality dimension captures not only how much professional judgment is required, but also how costly it is to be wrong.

Quadrant A: Structured + Low Judgment (automation with exception handling)

This quadrant includes tasks such as routine account reconciliations with stable rules, transaction coding to a known chart of accounts, standard close checklist status reporting, and automated extraction of trial balance and roll-forward schedules. These are strong candidates for automation because the goal is consistency and speed, and the "right answer" is usually well-defined.

The human role is to manage exceptions, resolve upstream data issues, and validate boundary cases. Use reasonableness checks and an owned exception queue to prevent silent error propagation.

Quadrant B: Structured + High Judgment (AI assists; humans decide)

Quadrant B includes work that has structure—often in the form of checklists or decision trees—but still requires judgment because facts are nuanced or outcomes are material (Gorry & Scott Morton, 1971; Bonner, 1994; Messier et al., 2005). Examples include revenue recognition triage for unusual contracts, impairment indicator screening, support packages for reserves and estimates (allowance, warranty, returns), and tax position triage when rules exist but interpretation depends on the underlying facts.

Here, AI can add value as a screening and organizing layer. It can summarize relevant policies, surface anomalies, assemble support documentation, and propose hypotheses about drivers (Vasarhelyi et al, 2023; Li & Vasarhelyi, 2024; Fulcer et al., 2025). However, while AI is useful in the preparatory stage, the final evaluative step is associated with weaker human-AI performance and therefore requires careful human oversight (Vaccaro et al., 2024), especially when materiality judgments are consequential and errors are costly(Messier et al., 2005). For that reason, Quadrant B works best when teams treat AI as a preparer and classifier while requiring explicit tie-outs and a skepticism step before conclusions are finalized (see the control points in Section 3.2).

Quadrant C: Unstructured + Low Judgment (AI drafts; humans edit)

This quadrant covers narrative work that is useful but typically low-stakes, such as first drafts of close meeting notes and action summaries, internal process documentation, initial drafts of standard operating procedures, and rewriting or standardizing internal management report narratives intended for routine audiences.

This is often the highest-productivity zone for generative AI because it helps standardize language across teams. Noy and Zhang (2023) find that ChatGPT substantially increased productivity, reducing average completion time by 40% while improving output quality by 18%, and also reduced inequality between workers. Relatedly, Brynjolfsson et al. (2025) show that generative AI can raise productivity by diffusing the practices of stronger workers and disproportionately helping less-experienced workers, suggesting one mechanism by which standardized language and workflow templates may improve consistency across teams. At the same time, Vaccaro et al. (2024) find that human–AI combinations perform better in creation tasks than in decision tasks, which is consistent with using GenAI in this zone for drafting and synthesis while preserving stronger human review at the evaluative stage. The human contribution is editing for accuracy, relevance, tone, and consistency with internal terminology. The main risk here is output inconsistency over time. Chen et al. (2024) show that the behavior of the same LLM service can change substantially over a relatively short period, creating reproducibility and workflow-integration problems. Wang et al. (2024) further show that different prompt designs can produce materially different consistency and reliability outcomes. From a controls perspective, COSO (2026) notes that model drift, prompt-based manipulation, and frequent configuration changes can undermine the integrity of operations and reporting if they are not governed through robust internal controls. A practical response is to standardize prompts and templates and maintain an approved set of terminology and sources for recurring close and reporting narratives (see the competencies and roles in Section 3.3).

Quadrant D: Unstructured + High Judgment (co-pilot mode; hybrid wins only with controls)

Quadrant D can still offer substantial upside from human + AI collaboration because much of the work involves open-ended drafting, synthesis, and explanation. This is the type of task context in which human–AI combinations tend to perform better than in decision tasks (Vaccaro et al., 2024), and Noy and Zhang (2023) show that generative AI can substantially improve productivity and quality in drafting-oriented work. But the cost of error is also high in this quadrant because these outputs often bear on materiality judgments and decision-useful reporting; Messier et al. (2005) show that materiality is consequential

because the significance of omission or misstatement depends on its effect on users' judgments.

Hybrid performance depends on a clear division of labor. Prior research suggests that human–AI performance gains depend on deliberate task partitioning rather than undifferentiated substitution, with humans retaining responsibility for higher-order interpretation and oversight (Raisch & Krakowski, 2021; Leitner-Hanetseder et al., 2021). AI drafts and proposes alternative explanations; professionals verify, apply judgment/ materiality thresholds, escalate issues when needed, and take responsibility for the final position. This is consistent with prior work on human-AI complementarity (Saghafian & Idan, 2024; Hemmer et al., 2025). The critical risk in

Quadrant D is confident narrative error—numbers, causal stories, or policy interpretations that read well but are wrong. This risk is well documented in the LLM literature: Ji et al. (2023) show that generated text can be fluent and coherent while still being factually unsupported, and Farquhar et al. (2024) identify a class of hallucinations they describe as confabulations—outputs that are plausible-sounding but arbitrary and incorrect. From a controls perspective, COSO (2026) warns that GenAI can be confidently wrong, making validation especially important in reporting workflows. That is why Quadrant D requires the mandatory control points in Section 3.2 to make the workflow defensible.

4.2 C² Controls: Five Mandatory Control Points for High-Judgment Close and Reporting Work

A central contribution of the C³ Framework is translating "human-in-the-loop" from a general aspiration into a set of concrete workflow gates that can be embedded into close and reporting. These control points are mandatory for Quadrant D work (unstructured + high judgment) and strongly recommended for Quadrant B work (structured + high judgment), where errors can affect reporting outcomes, stakeholder reliance, or compliance obligations. The intent is to make AI-supported outputs defensible by design—not by relying on individual vigilance under time pressure. This logic is consistent with Reason's (2000) argument that robust systems should reduce dependence on individual vigilance by building layered defenses against error; in the close context, Janvrin and Mascha (2014) describe a compressed, high-stakes reporting process in which time pressure heightens the importance of workflow design; and COSO (2026) extends that principle to GenAI by emphasizing built-in control activities, monitoring, and audit-ready evidence rather than informal reviewer effort alone. The control set is also designed to fit within enterprise AI governance approaches, including the NIST AI Risk Management Framework and ISO/IEC 42001's emphasis on accountability and documentation (ISO/IEC 42001, 2023; NIST, 2023).

Control Point 1: Source Grounding and Traceability. AI-assisted narratives and analyses should be constrained to approved sources, and the workflow should make those sources visible. This is consistent with Lewis et al. (2020), who show that retrieval-augmented generation improves knowledge-intensive outputs by grounding generation in retrieved source material rather than relying only on parametric model memory. In close and reporting work, many failures are caused by bad inputs (wrong period, an incomplete data pull, inconsistent KPIs, outdated policies). This reflects a broader data-quality problem: Wang and Strong (1996) identify completeness, timeliness, and consistency as core dimensions of data quality; Bai et al. (2012) show that poor data quality propagates through accounting information systems and requires explicit control strategies; and Ge and McVay (2005) document that internal-control weaknesses in reporting environments often

involve reconciliation problems, policy deficiencies, and related process failures. Requiring source grounding and traceability forces teams to specify source-of-truth inputs and include a short "source header" that documents the reporting period, extraction timestamp, and source system/version.

Control Point 2: Independent Verification and Tie-out. Any AI-supported output that contains numbers, or makes factual claims tied to the numbers, should pass a verification gate before it is finalized. This requirement is grounded in both controls and evidence standards: COSO (2026) recommends treating AI-generated content as an assertion that requires corroborating evidence before inclusion in official records; NIST AI 600-1 calls for fact-checking GenAI outputs against ground truth or upstream source data; and PCAOB AS 1105 provides the audit-evidence basis for requiring sufficient appropriate support before quantitative or factual assertions are finalized. Quantitative statements should tie to authoritative records, and reviewers should confirm that cited drivers reconcile to the variance. A "numbers-first gate" is helpful: no narrative is approved until tie-outs are complete.

Control Point 3: Contradiction Testing. The workflow should include a structured step that deliberately surfaces counterevidence, alternative explanations, missing assumptions, and uncertainty. This step is consistent with classic judgment and auditing research: Lord et al. (1984) show that explicitly "considering the opposite" can reduce confirmation bias and improve judgment quality; Nelson (2009) defines professional skepticism as a greater tendency to doubt the validity of an assertion and to require more persuasive evidence before concluding that it is correct; and Lyell and Coiera (2017) show that automation bias can reduce spontaneous verification and counterevidence search unless these checks are deliberately structured into the workflow. In practice, require the preparer/reviewer to identify plausible alternative explanations, key assumptions that could change the conclusion, and what additional data would validate the explanation.

For estimates and reserves, contradiction testing can target sensitivity (e.g., "What would cause this estimate to be overstated or understated?"). The point is to make "professional skepticism" a procedure, not merely an attitude.

Control Point 4: Escalation and Approval Thresholds. Hybrid workflows break down when accountability is unclear. Close and reporting teams should define explicit triggers that require escalation to a designated approver (e.g., controller, accounting policy lead). Common triggers include variances above a materiality threshold, novel transaction types or unusual contract terms, narratives that will be shared externally, and AI-supported outputs that imply a new or changed accounting policy position. Approval rules should specify who approves, what evidence must accompany the approval, and how long documentation is retained.

Control Point 5: Audit-Trail Logging and Version Discipline. Finally, organizations should retain an auditable record of what the AI produced, what the human changed, and what sources were used. A minimal "AI workpaper log" can be lightweight but should be consistent: the tool/model and version, the prompt (or prompt template identifier), any retrieval sources used, output versions, reviewer identity and notes, final approval, and the reporting period/date/time. This log turns an AI-assisted narrative from an informal draft into a defensible reporting artifact.

Putting the controls together: a close narrative co-pilot workflow. Consider a monthly flux analysis narrative prepared for executive reporting. The workflow starts by locking the inputs, approved trial balance and KPI extracts with consistent definitions and a documented reporting period. AI can then draft a narrative explanation and list candidate drivers, explicitly labeled as hypotheses. Next, a reviewer verifies each number and driver against ledger/KPI extracts and removes unsupported claims. The team then runs a contradiction-testing step to surface alternative drivers and missing data before deciding whether additional analysis is needed. Narratives above threshold are escalated for controller approval (and to the policy lead if accounting interpretation is implicated). Finally, the narrative, tie-out sheet, prompt template identifier, and review notes are archived in the close binder/workpaper repository. This co-pilot design preserves speed while improving defensibility by making verification, skepticism, and accountability explicit steps rather than optional best practices.

4.3 C³ Competencies: Skills and Roles for Collaborative Intelligence in the Controller's Organization

The final layer of the C³ Framework focuses on what people need to do differently for human + AI workflows to be reliable in close and reporting. The shift is not primarily technical. It is operational: accounting teams move from simply producing outputs to designing the workflow, verifying the evidence behind outputs, and taking responsibility for the final position.

In practice, scaling AI use in close and reporting typically requires five competencies, which synthesize recurring themes across the reviewed sources (Choi & Xie, 2025; CPA.com, 2025; Dong et al., 2024; KPMG, 2024; NIST, 2023):

1. Data and definition discipline (KPI definitions, period control, source-of-truth ownership).
2. Narrative verification (tie-outs for claims, not just numbers).;
3. Judgment/materiality discipline (what can be standardized vs what must be escalated)
4. Prompt and workflow standardization (templates, guardrails, reusable patterns).
5. Governance fluency (tool approval, data restrictions, documentation and retention expectations).

Together, these sources emphasize data discipline, verification practices, human judgment and escalation, workflow standardization, and governance expectations as practical requirements for reliable human–AI collaboration.

Emerging Roles that Make Hybrid Work Operational

Many accounting teams do not have dedicated AI teams, and they do not need them to get started. What they do need is clear ownership for a small number of responsibilities that otherwise fall between the cracks. The following "lightweight" roles can be assigned fractionally in smaller organizations (one person may wear multiple hats), but the responsibilities should be explicit.

1. **Finance AI Output Reviewer (Narrative Quality Assurance (QA) Lead).** This role owns the review step for AI-supported narratives and ensures key statements are tied to authoritative records. In practical terms, the reviewer maintains a simple checklist for narrative tie-outs, documents the contradiction-testing step when required, and confirms that a named approver has taken responsibility for the final version. The value is consistency: review becomes a repeatable close activity rather than an ad hoc scramble at the end.

2. **Finance Model-Risk / Governance Liaison (Finance–IT/GRC Liaison).** This role connects the finance team to information technology (IT) and governance, risk, and compliance (GRC) stakeholders so that close and reporting use cases use approved tools, approved data sources, and appropriate retention and access controls. Typical artifacts include a short use-case intake form, a lightweight risk rating, an approved tool list, and clear escalation triggers. The value is speed with safety: finance teams can adopt AI without repeatedly negotiating exceptions or discovering restrictions late in the process.

3. **Prompt and Policy Librarian (Close Enablement Lead).** This role maintains standardized prompt templates, recurring narrative patterns, and approved source lists for close tasks. The goal is not to "optimize prompts" in the abstract; it is to reduce variability across periods and across preparers. Practical artifacts include a prompt library organized by close use case, a style guide for management commentary, and a short "do-not-use" list covering sensitive data and prohibited behaviors (e.g., invented numbers or unsupported policy conclusions).

4. **Evidence and Traceability Steward.** This role ensures traceability, logging, and version discipline are implemented consistently so AI-assisted work can be defended, reviewed, and retained like any other close workpaper. Practical artifacts include a standard AI workpaper log template, retention mappings to existing close binders/workpaper repositories, and workflow checklists that ensure required fields (sources, period, reviewer, approvals) are captured.

The underlying principle is not optional: if responsibilities are left implicit, hybrid workflows tend to devolve into "AI drafts plus vague human oversight," which weakens accountability and increases risk.

Section 3: Summary and Implementation Artifacts

Across the three layers, the C³ Framework provides a controllership-oriented model that is both actionable and researchable. It combines a task typology (where hybrid adds value and risk), a set of mandatory workflow gates for defensibility, and a role/competency map that assigns clear ownership.

To help teams put the C³ Framework into action, the paper includes two practical tools. Table 1 maps common close and reporting tasks to collaboration modes (automate, assist, draft-and-edit, or co-pilot). Exhibit 1 provides a one-page workflow blueprint covering grounding/traceability, verification and tie-out, contradiction testing, escalation and approval, and audit-trail logging. Together, the tools show how teams can achieve faster

outputs while preserving defensibility and accountability in high-judgment close and reporting work.

4.4 Illustrative Mini-Case: Applying C^3 to a Reserve/Estimate Support Memo (Hypothetical)

Background: A company's accounting team prepares a monthly (or quarterly) support memo for a significant estimate, such as the allowance for credit losses, inventory reserve, or warranty reserve. The memo is used internally for management review and, in many organizations, becomes part of the close binder that supports audit review. The team wants to use generative AI to speed up memo drafting and standardize structure across business units, but leadership is concerned about unsupported narratives, incorrect calculations, and inconsistent use of accounting policy language.

C^1 Complementarity (task classification and collaboration mode). The reserve memo is structured + high judgment/materiality (Quadrant B) when the workflow is checklist-driven (inputs, roll-forward, required disclosures) but still requires judgment about assumptions, overlays, and reasonableness. It can shift into Quadrant D when uncertainty or scrutiny increases (e.g., macro shocks, major customer defaults, material overlays, or methodology changes). In these situations, the memo becomes more interpretive and the consequences of error rise. The team uses AI assist; humans decide for Quadrant B work and switches to co-pilot mode for Quadrant D conditions.

C^2 Controls (workflow gates for defensibility). The team embeds the control points as a standard close memo workflow:

1. Grounding/traceability: AI is restricted to approved sources (ERP/GL extracts, aging reports, historical loss-rate schedules, reserve roll-forward, approved KPI definitions, and the internal accounting policy library). A short source header documents period, report timestamps, and versions.

2. Verification/tie-out: all quantitative statements (reserve balance, roll-forward, key ratios, overlays) are tied back to authoritative schedules. Any number included in the narrative must link to a tie-out field.

3. Contradiction testing: the workflow requires a "challenge pass" before finalizing, such as: "What assumptions would most likely cause this reserve to be understated?" "Which segments drive the change?" and "What evidence would argue for a different reserve level?" The reviewer documents why alternatives were rejected or what additional analysis was performed.

4. Escalation/approval: explicit thresholds trigger escalation (e.g., reserve change above a defined dollar/percentage threshold; new overlay; change in methodology; qualitative factor adjustments). Controller approval is required for material changes, and the accounting policy lead reviews any changes in policy framing.

5. Audit-trail logging: the memo retains the AI prompt/template identifier, input schedule list, output versions, reviewer notes, and approvals in the close binder so the evidence trail is clear.

C^3 Competencies (roles and responsibilities). The Finance AI Output Reviewer checks tie-outs, ensures the narrative matches the schedules, and confirms the contradiction-testing step is documented. The Prompt & Policy Librarian maintains a standard reserve memo template and approved policy language so the memo remains consistent period to period. The Evidence and Traceability Steward ensures the memo includes the source header and AI workpaper log and that all supporting schedules are attached or linked. If the organization restricts tools or data pathways, the Finance–IT/GRC Liaison ensures the approved tool list and retention rules are followed.

Outcome (what improves). The immediate benefit is faster drafting and more consistent memo structure, but the bigger gain is defensibility: key assumptions are surfaced explicitly, numbers are tied to schedules, alternative explanations are documented, and approval thresholds are clear. As a result, review cycles become more efficient because reviewers spend less time reconstructing how the memo was produced and more time evaluating the reasonableness of assumptions and conclusions.

5. Discussion

5.1 Interpreting the Thesis: why "Human + AI" Can Be Superior

The controllership close and reporting cycle is an archetypal environment for collaborative intelligence. It combines structured, repeatable routines such as reconciliations, roll-forwards, and checklist management with high-judgment decisions involving estimates, policy interpretation, and material close adjustments, while also requiring narrative work that translates numbers into explanations for decision-makers through flux narratives, executive decks, and stakeholder responses. This mix is consistent with prior work on the close process and accounting communication: Janvrin and Mascha (2014) describe the close as a coordinated, recurring process that combines standardized routines with judgment-intensive reporting decisions; Kokina et al. (2021) show that many accounting and finance tasks are suitable for automation because bots execute structured workflows, while accountants' roles shift toward explanation, sustaining, and review; and Loughran and McDonald (2016) survey a growing literature showing that textual disclosures such as SEC filings, earnings conference calls, and related business documents are an important and well-studied channel through which accounting and financial information is communicated to users. In such environments, the comparative advantage of human+AI is not simply speed. It is the ability to produce multiple candidate narratives and hypotheses rapidly while preserving human accountability for judgment, materiality, and verification. Hybrid workflows can outperform humans-alone when they reduce blank-page time and broaden the hypothesis set (AI as drafter, critic, or simulator). Hybrid workflows can outperform AI-alone when organizations preserve the "professional core": humans own the decision, apply context, and verify outputs against authoritative sources.

5.2 The Central Managerial Lesson: Controls Are What Make Hybrid Performance Real

A recurring failure mode in early GenAI adoption is confusing "draft quality" with "decision quality." Ji et al. (2023) show that generated text can be fluent and natural even when it is unfaithful or nonsensical, and Farquhar et al. (2024) show that LLMs can produce confabulations that are plausible yet arbitrary and incorrect. Lyell and Coiera (2017) further show that automation bias can reduce vigilant information seeking and verification when checking automated output is cognitively demanding. Fluent narratives can create an illusion of correctness, and teams may become prone to approving outputs when verification is cognitively demanding or under time pressure. The C^3 Framework

addresses this by embedding mandatory control points as workflow gates for Quadrant D tasks and recommended gates for Quadrant B tasks. These gates translate governance principles into operational practice consistent with standards-based expectations around risk management, accountability, and documentation (ISO/IEC 42001, 2023; NIST, 2023).

Implication for controllers: The goal is not to prohibit AI in close narratives; it is to design "co-pilot" workflows where AI's contribution is bounded and auditable. A well-designed workflow makes it easier for professionals to do the right thing under time pressure—because verification, contradiction testing, and logging are required steps rather than optional best practices.

5.3 Implications for the Accounting Profession: from Production to Verification and Ownership

The competencies and roles proposed in C^3 imply a shift in professional identity: accountants increasingly move from being "producers of drafts" to being "owners of decisions and verifiers of evidence." This aligns with emerging professional narratives about AI increasing productivity while elevating the importance of human judgment and communication (CPA.com, 2025). In close and reporting, this shift is particularly salient because many deliverables are narratives that appear authoritative and travel quickly through the organization.

A practical consequence is that talent development should emphasize verification literacy for narratives (not just numeric tie-outs), judgment under uncertainty, and governance fluency around what data may be used, what must be logged, and what requires escalation. Talent development should also include standardized prompt and template development so that recurring AI-assisted reporting tasks operate with consistent inputs, review steps, and documentation expectations.

5.4 Research Agenda: Testable Propositions

This conceptual work sets up several propositions suitable for future empirical testing. Researchers can examine outcomes such as accuracy, time-to-close, quality of variance explanations, escalation frequency, and consistency across periods. Studies could use archival close deliverables, controlled experiments with professionals, or simulations with standardized datasets and evaluation rubrics. Potential research designs include within-subject comparisons across collaboration modes (human-only vs AI-only vs hybrid) and field-embedded process analyses where controls are introduced as interventions.

5.5 Limitations

This paper is conceptual and synthesizes research and practice guidance rather than presenting primary empirical evidence. Accordingly, the propositions and control architecture are intended as a structured starting point that invites empirical testing and refinement. Additionally, regulatory expectations and governance standards may evolve; the control points are designed to be robust to such change by focusing on traceability, verification, accountability, and documentation.

6. Conclusion

This paper proposes the C^3 Framework as an actionable model for deploying collaborative intelligence in controllership, close, and reporting. It argues that human+AI outperforms humans-alone or AI-alone most reliably when (i) tasks are judgment- or language-

intensive, (ii) organizations embed mandatory control points that make outputs defensible, and (iii) responsibilities are assigned to align machine strengths with human accountability.

Key takeaways for practice

- The highest return on investment (ROI) for GenAI in controllership is often narrative work (flux explanations, executive commentary), but this is also where risk is highest.
- Hybrid superiority is not automatic; it depends on workflow design.
- Five control points—grounding, verification, contradiction testing, escalation, and logging—convert "human-in-the-loop" from principle to practice.
- Controllers should invest in explicit roles (review, model-risk liaison, prompt/policy stewardship, evidence engineering) to prevent accountability diffusion.
- Table 1 and Exhibit 1 provide ready-to-adopt artifacts for implementing the framework in close and reporting workflows.

Data Availability Statement: No primary dataset was generated or analyzed for this conceptual paper; data availability is not applicable.

Ethics Statement: Not applicable. This study did not involve human subjects research.

Conflict of Interest Statement: The author declares no conflicts of interest.

Funding Statement: No external funding was received for this work.

Disclosure of AI tool use. The author used generative AI tools for ideation and drafting support. The author determined the scope, selected and reviewed all sources, made all decisions regarding inclusion, framing, and interpretation, and takes full responsibility for the final content.

References

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Bai, X., Nunez, M., & Kalagnanam, J. R. (2012). Managing data quality risk in accounting information systems. *Information Systems Research, 23*(2), 453–473. https://doi.org/10.1287/isre.1110.0371

Bonner, S. E. (1994). A model of the effects of audit task complexity. *Accounting, Organizations and Society, 19*(3), 213–234. https://doi.org/10.1016/0361-3682(94)90033-7

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Chen, L., Zaharia, M., & Zou, J. (2024). How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review, 6*(2). https://doi.org/10.1162/99608f92.5317da47

Choi, J. H., & Xie, C. (2025). Human + AI in Accounting: Early Evidence from the Field. *Working Papers (Faculty) - Stanford Graduate School of Business,* 1-101. http://dx.doi.org/10.2139/ssrn.5240924

Committee of Sponsoring Organizations of the Treadway Commission. (2026). *Achieving effective internal control over generative AI (GenAI).* https://www.coso.org/generative-ai

CPA.com. (2025). *CPA.com 2025 AI in Accounting Report*. https://www.cpa.com/sites/cpa/files/2025-06/2025_AI_in_Accounting_Report.pdf

Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems, 55,*100715. https://doi.org/10.1016/j.accinf.2024.100715

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0

Fulcer, K., Gu, H., Hu, H., Huang, Q., Kogan, A., Vasarhelyi, M. A., Wei, D., & Young, J. (2025). *Application of outlier detection methods in audit analytics. Accounting Horizons, 39*(3), 143–157. https://doi.org/10.2308/HORIZONS-2023-071

Ge, W., & McVay, S. (2005). The disclosure of material weaknesses in internal control after the Sarbanes-Oxley Act. *Accounting Horizons, 19*(3), 137–158. https://doi.org/10.2308/acch.2005.19.3.137

Gorry, G. A., & Scott Morton, M. S. (1971). A framework for management information systems. *Sloan Management Review, 13*(1), 55–70. https://dspace.mit.edu/handle/1721.1/47936

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2025). Complementarity in human-AI collaboration: Concept, sources, and evidence. *European Journal of Information Systems, 34*(6), 979-1002. https://doi.org/10.1080/0960085X.2025.2475962

International Organization for Standardization & International Electrotechnical Commission. (2023). *ISO/IEC 42001:2023—Artificial intelligence management systems—Requirements*. ISO https://www.iso.org/standard/42001

Janvrin, D., & Mascha, M. F. (2014). The financial close process: Implications for future research. *International Journal of Accounting Information Systems, 15*(4), 381–399. https://doi.org/10.1016/j.accinf.2014.05.007

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kokina, J., Gilleran, R., Blanchette, S., & Stoddard, D. (2021). Accountant as digital innovator: Roles and competencies in the age of automation. *Accounting Horizons, 35*(1), 153–184. https://doi.org/10.2308/HORIZONS-19-145

Leitner-Hanetseder, S., Lehner, O. M., Eisl, C., & Forstenlechner, C. (2021). A profession in transition: Actors, tasks and roles in AI-based accounting. *Journal of Applied Accounting Research, 22*(3), 539–555. https://doi.org/10.1108/JAAR-10-2020-0201

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Li, H., & Vasarhelyi, M. A. (2024). Applying large language models in accounting: A comparative analysis of different methodologies and off-the-shelf examples. *Journal of Emerging Technologies in Accounting, 21*(2), 133–152. https://doi.org/10.2308/JETA-2023-065

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*(6), 1231–1243. https://doi.org/10.1037/0022-3514.47.6.1231

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association, 24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Messier, W. F., Jr., Martinov-Bennie, N., & Eilifsen, A. (2005). A review and integration of empirical research on materiality: Two decades later. *Auditing: A Journal of Practice & Theory, 24*(2), 153–187. https://doi.org/10.2308/aud.2005.24.2.153

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

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

Nelson, M. W. (2009). A model and literature review of professional skepticism in auditing. *Auditing: A Journal of Practice & Theory, 28*(2), 1–34. https://doi.org/10.2308/aud.2009.28.2.1

Noy, S., & 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

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Reason, J. (2000). Human error: Models and management. *BMJ, 320*(7237), 768–770. https://doi.org/10.1136/bmj.320.7237.768

Saghafian, S., & Idan, L. (2024). *Effective generative AI: The human-algorithm centaur*. Harvard Data Science Review. https://doi.org/10.1162/99608f92.19d78478

Vasarhelyi, M. A., Moffitt, K. C., Stewart, T., & Sunderland, D. (2023). Large language models: An emerging technology in accounting. *Journal of Emerging Technologies in Accounting, 20*(2), 1–10. https://doi.org/10.2308/JETA-2023-047

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*, 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine, 7*, 41. https://doi.org/10.1038/s41746-024-01029-4

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Exhibit 1.

Hybrid Workflow Blueprint for Close Narratives and Reporting (1-page checklist)

Use this checklist for Quadrant B (structured + high judgment) and require it for Quadrant D (unstructured + high judgment).

A. Use-case intake and risk rating (before using GenAI)

- Use case (e.g., monthly flux narrative; reserve memo; executive commentary; policy position summary)
- Task quadrant (A/B/C/D)
- Audience (internal ops / executive / board / external-facing)
- Materiality/stakes (low/medium/high)
- Decision rights: who owns final content (name + role)

Approved tools and data

- Approved GenAI tool/model
- Approved data sources only (system-of-record extracts, KPI definitions, prior approved memos, policy library)
- Data extraction timestamp and period
- Sensitive data restrictions (personally identifiable information (PII) / confidential terms)

B. Drafting protocol (what the AI is allowed to do)

- Allowed: draft narrative; summarize approved documents; propose candidate drivers (labeled hypotheses); generate alternatives and questions
- Disallowed: invent numbers; cite non-approved sources as authoritative; make policy conclusions without references and sign-off
- Prompt template identifier; style guide/standard terminology applied

C. Mandatory control points (defensible workflow gates)

1. Grounding/Traceability gate: traceability header included
2. Verification/Tie-out gate: numbers tied; drivers reconcile; claims cross-checked
3. Contradiction-testing gate: alternatives reviewed; missing data identified; uncertainty documented
4. Escalation & approval gate: thresholds, novel transactions, policy issues, board/external use

5. Audit-trail logging gate: prompts/template identifiers, outputs, versions, reviewer notes, tie-out links, approver, timestamp

D. Finalization and retention

- Final human owner
- Retention location (close binder/workpaper repository)
- Retention period/policy
- Disclosure note (if relevant)

Governance alignment note: This blueprint operationalizes human-in-the-loop through traceability, verification, accountability, and documentation—capabilities emphasized in NIST AI RMF and ISO/IEC 42001 (ISO/IEC 42001, 2023; NIST, 2023).