

Evidence Traceability in AI-Assisted Accounting Summaries for Small Enterprises

Ijang Faisal¹ | Cucu Sugyati²

1. Universitas Muhammadiyah Bandung | ijangfaisal@umbandung.ac.id

2. PELITA NUSA BANDUNG BARAT; STIMIK Mardira Bandung | Sugyaticucu@gmail.com

Abstract

Generative artificial intelligence can produce fluent accounting narratives without demonstrating that every figure and explanation follows from valid records. This conceptual article develops an evidence-traceability framework for small-enterprise summaries. It separates authorised sources, deterministic calculation, narrative generation, and human approval, and links material claims to source versions, periods, and verification status. Selected risk-management and data-governance publications provide contextual foundations; operational controls and propositions are developed analytically. No software evaluation, experiment, company dataset, or accuracy estimate is reported. Hypothetical cases distinguish arithmetic errors from mismatched periods and unsupported causal explanations. The article proposes a claim-evidence register, a review sequence, change controls, and a synthetic-data experiment for prospective testing. Three propositions concern detection of source and period errors, the distinct effect of calculation separation on numerical accuracy, and the balance between review burden and information reliability. The contribution is a testable control architecture rather than a claim that AI improves financial reporting. Application requires validation with authorised data, explicit treatment of uncertainty, and protection of unnecessary personal and commercial information. The framework does not replace accounting standards, professional judgement, or the responsibility of the person approving the summary.

Keywords: generative AI; accounting; evidence traceability; small enterprises; human review; information controls.

1. Introduction

An accounting summary must be understandable and its material claims must be supportable. AI can assist language production, but professional-sounding text may conceal the use of an incorrect source, inconsistent period, or unsupported explanation. The article asks what evidence needs to be retained so review remains proportionate without sacrificing the ability to verify important claims.

2. Conceptual method and risk-management context

The analysis separates the work of selecting sources, calculating values, composing language, and approving output. The unit of review is a material claim linked to a defined reporting purpose. This structure identifies where errors can enter and what evidence an authorised reviewer needs to detect them.

NIST (2023) supplies a general AI risk-management foundation. Its voluntary framework motivates attention to governance and risk throughout an AI system's use. This article adapts that broad perspective to a narrow accounting-summary workflow. It does not claim certification, formal conformance, or empirical validation under the NIST framework.

World Bank (2021) supplies context for the value and potential harm of data use. OECD (2021) supplies context for differences in small-enterprise digital capabilities. The sources are selected purposively; no systematic review or comparative product assessment is conducted. Their general insights do not establish that the proposed controls reduce errors in practice.

The conceptual procedure defines material claims, identifies error classes, constructs hypothetical failure cases, and derives propositions and a prospective test design. No real company financial statements or customer records are processed. All examples below are synthetic illustrations and should not be interpreted as measured application performance.

Work layer	Principal output	Main risk addressed
Source selection	Authorised, dated record set	Wrong or stale information
Calculation	Reproducible totals and ratios	Arithmetic and formula errors
Narrative generation	Explained descriptive claims	Unsupported interpretation
Human approval	Reviewed summary and exception record	Unexamined release

The layers can be implemented in a simple worksheet-based process. Complexity is not an objective in itself. The research question is whether separating responsibilities and preserving evidence improves decision-useful reliability at a cost that a small enterprise can sustain.

3. Authorised sources, versions, and reporting periods

The source layer establishes which records are permitted for the stated summary, their reporting periods, and whether they are provisional or final. A file's presence in a folder is not sufficient evidence that it is the correct source. Similar filenames, revised exports, and different cut-off dates can lead to inconsistent inputs.

Each source should have an identifier, version date, responsible owner, and status. When a document is revised, the summary must either use the newer version consistently or preserve the earlier version for a stated reason. Combining a revised revenue figure with an old expense total may create an internally inconsistent report even if each number once appeared in a valid record.

The period definition includes the start and end dates, aggregation level, and any comparability qualification. Monthly amounts, year-to-date totals, and rolling averages answer different questions. Labels should remain visible when data are transferred to a calculation sheet or narrative prompt.

Access should be limited to the information needed for the task. A descriptive summary of aggregate sales need not include customer identification or unrelated private documents. Restricting inputs reduces unnecessary exposure and can also make the evidential basis easier to inspect.

Source field	Function	Example of prevented confusion
Record identifier	Locate the origin of a value	Multiple similar exports
Version and status	Establish which revision is used	Mixing provisional and final figures
Reporting period	Define the measurement window	Monthly versus year-to-date totals
Responsible owner	Route questions and corrections	Unclear update responsibility
Permitted use	Bound data access	Including unrelated personal records

A missing source should be recorded as missing, not silently replaced with a plausible number. The narrative may explain that the summary is incomplete, or the workflow may stop until material data are available. The choice depends on reporting purpose and materiality, but must remain visible to the reviewer.

4. Deterministic calculations and reproducibility

Totals, percentage changes, and ratios should have inspectable formulas and recorded inputs. A language model need not be the sole calculation engine. A worksheet or deterministic program can produce the values used by the narrative, with the formula version retained for later review.

For a change calculation, the numerator, denominator, and comparison periods must be explicit. A correct formula applied to incomparable periods still produces a misleading statement. Reproducibility therefore includes definitions and labels, not merely the ability to recompute an identical number.

Calculations need appropriate treatment of zero and negative denominators. A percentage change from zero is not defined by the usual formula. Replacing the denominator with an arbitrary small value creates an impressive but unsupported percentage. The summary should instead use an appropriate absolute comparison or explain the limitation.

Rounding rules should be consistent. Small differences can arise when rounded components are added rather than rounding the total once. Such differences should be documented where material to the report. Arithmetic agreement at the displayed precision does not validate the underlying source.

Calculation control	Evidence retained	Review question
Formula definition	Formula and version	Does it answer the intended question?
Input mapping	Source identifiers and periods	Are the compared values compatible?
Exception handling	Zero, missing, or negative cases	Is the reported result meaningful?
Rounding policy	Units and displayed precision	Can differences be explained?
Reproduction	Saved inputs and outputs	Can an authorised reviewer recompute it?

The calculation layer reduces one class of uncertainty while leaving others unresolved. It cannot establish whether the source record is complete or whether a causal explanation is justified. Research should therefore measure arithmetic accuracy separately from source selection, period alignment, and interpretation quality, rather than combine all errors into a single unqualified success rate.

5. Narrative claims, interpretation, and recommendations

The narrative layer receives verified calculations and instructions about the claims it may express. Descriptive facts, interpretations, and recommendations should be distinguishable. A statement that receivables increased can follow from comparable balances. A statement that collection effort weakened requires additional evidence. A recommendation to intensify collection adds a judgement about action and consequences.

An evidence link for a number does not automatically support an explanation attached to it. Rising sales and rising advertising expenditure may occur together without demonstrating that advertising caused the sales change. The summary should preserve this distinction rather than convert timing into causality through fluent wording.

Materiality depends on reporting purpose. A claim may be important because of its monetary size, its effect on a decision, or the sensitivity of the subject. A small amount can still be consequential if it implies non-compliance or a significant operational problem. The review scope should therefore be stated before narrative generation.

The claim-evidence register records each material assertion, its category, supporting source, calculation where relevant, period, and verification status. Unsupported statements can be removed, qualified, or sent for additional evidence. The reviewer may approve a numerical fact while rejecting its causal interpretation.

Claim type	Hypothetical statement	Evidence required
Descriptive fact	Monthly revenue increased	Comparable verified revenue records
Interpretation	Collection performance weakened	Collection history and relevant context
Causal explanation	Advertising caused the increase	Appropriate causal evidence
Recommendation	Change the collection process	Justified problem and assessed action

The register need not duplicate the entire report. Its purpose is to make consequential claims inspectable. For a short summary, a compact table may suffice. For complex reports, claim identifiers can connect paragraphs to supporting records without exposing unnecessary underlying data to every reader.

6. Hypothetical error cases

Suppose a synthetic dataset records September revenue of IDR 120 million and August revenue of IDR 100 million. The comparable monthly change is 20 percent. If September is instead compared with a January-to-August total of IDR 800 million, the formula can be arithmetically correct while answering a different question.

A second synthetic case uses the correct monthly values but attributes the increase entirely to a promotion because advertising spending also rose. The figures establish co-occurrence, not causal attribution. Additional information about prices, seasonal demand, product availability, and comparison conditions would be needed before making the stated explanation.

Synthetic error	Why it can look credible	Appropriate check
Wrong source version	Figure existed in an earlier export	Compare approved version identifiers
Wrong reporting period	Arithmetic is internally correct	Inspect period labels and aggregation
Wrong calculation	Narrative is fluent	Recompute formula from verified inputs
Unsupported cause	Explanation fits intuition	Identify evidence beyond timing
Premature finality	Provisional total is plausible	Check source status and exceptions

A third case treats a provisional expense total as final. The number may later be revised after missing invoices are recorded. A traceable summary should state provisional status and identify the outstanding completeness issue rather than offer an unqualified conclusion about profitability.

These cases show why review must address several error classes. Formula verification alone would miss period mismatch and causal overstatement. Source verification alone would miss a calculation error. Professional-looking prose does not resolve either. The framework therefore proposes separate checks and a recorded exception status. It makes no claim that these checks eliminate every error or that a particular AI product behaves in the illustrated way.

7. Human approval and the correction record

Human approval should examine material claims against evidence, rather than merely read the narrative for plausibility. The proposed sequence begins with source and period checks, proceeds to calculations, and then assesses interpretations and recommendations. This order is a researchable control proposal, not a proven method for eliminating reviewer bias.

The reviewer records approved, qualified, rejected, and unresolved claims. A summary can be released with clearly communicated limitations only when its reporting purpose permits that approach. Unresolved material errors should not be disguised through softer wording that preserves the same misleading implication.

Corrections should retain the original claim, the reason for change, the supporting evidence, and the approved replacement. This record distinguishes source failures from calculation or narrative failures. Repeated errors can then be addressed at the appropriate layer rather than corrected sentence by sentence indefinitely.

The approving person remains responsible for the released summary within the relevant organisational arrangements. AI assistance should not obscure who authorised its use or who approved the final content. Access to correction histories should also be proportionate, especially if the underlying records contain confidential information.

Review outcome	Meaning	Recorded response
Approved	Evidence supports the material claim	Reviewer and date
Qualified	Claim valid within a stated limitation	Limitation included in summary
Rejected	Claim unsupported or incorrect	Removal or evidenced replacement
Unresolved	Material evidence still unavailable	Hold or permitted incomplete reporting

Review effort must be evaluated. Checking every trivial sentence may remove the efficiency benefit, while checking too little risks consequential error. The framework proposes a material-claim focus with random or targeted checks of less material statements where justified. Future studies should report the errors that remain, not merely the number corrected before release.

8. Change control and proportionate implementation

The workflow should retain the source set, calculation version, model or service identifier where available, generation date, instructions, output, and approval record. This information supports reconstruction when results change. A newer output is not automatically better simply because it sounds more polished.

A source correction should trigger review of dependent calculations and claims. Otherwise, changing a figure in one sheet may leave the narrative unchanged. Claim identifiers can help locate affected sentences. The system should also distinguish a routine wording edit from a material change in financial interpretation.

Small enterprises can start with descriptive summaries drawn from structured aggregate records. Expansion to explanations or recommendations should follow demonstrated review capability. A simple process with clearly assigned responsibility may be preferable to a complex system whose evidence cannot be inspected by its users.

The framework does not require uploading all financial records to an external service. Input selection should follow the reporting purpose and permitted access. Where a narrative can be generated from approved aggregate values, unrelated documents and personal details should remain outside the input set.

Change event	Required review target	Reason
Source amount corrected	Dependent figures and factual claims	Prevent stale outputs
Reporting period changed	Comparison labels and ratios	Maintain comparability
Generation instructions changed	Narrative scope and claim types	Detect expanded interpretation
Tool version changed	Representative output checks	Identify altered behaviour
Approval responsibility changed	Release and exception records	Preserve accountability

Implementation should measure time spent selecting inputs, reconciling data, checking claims, and correcting outputs. A claimed time saving that excludes preparation and review overstates the benefit. The relevant outcome is the total work needed to produce a usable and adequately supported summary, including the cost of maintaining evidence over repeated reporting cycles.

9. Propositions and prospective experiment

Proposition 1 states that explicit evidence links for material claims will improve detection of source and period errors. A synthetic-data experiment can compare reviewers with and without a claim-evidence register, using otherwise comparable summaries. Detection should be scored against pre-specified planted errors, not an evaluator's impression of report quality.

Proposition 2 states that separating deterministic calculations from narrative generation will reduce numerical errors without necessarily reducing interpretation errors. The experiment should therefore report arithmetic, source, period, and reasoning outcomes separately. Improvement in one category cannot justify a claim about total reporting reliability.

Proposition 3 states that proportionate claim review can improve reliability without imposing excessive total workload. The relevant trade-off includes preparation, verification, corrections, and residual errors. A shorter reading time alone does not demonstrate efficient control.

Experimental measure	Proposed observation	Interpretation limit
Error detection	Correctly identified planted errors	Synthetic tasks may differ from practice
Residual material error	Errors surviving approval	Define materiality before testing
False correction	Supported claims wrongly changed	Avoid rewarding intervention alone
Total workload	Preparation plus review and correction time	Include evidence-maintenance effort
Reviewer experience	Accounting and AI familiarity	Test heterogeneous effects

Random assignment, comparable task difficulty, and pre-specified scoring can strengthen the design. Repeated tasks require attention to learning effects. Synthetic or properly authorised de-identified data should be used, and participant privacy preserved. Laboratory results would not directly establish safety or effectiveness for real financing decisions. No experimental results or accuracy percentages are reported in this article.

10. Conclusion, limitations, and research contribution

AI-assisted accounting summaries should preserve a verifiable connection between material claims, approved sources, reporting periods, and calculations. Separating source selection, deterministic calculation, narrative generation, and human approval makes the control object more explicit. The synthetic illustrations explain why arithmetic agreement alone cannot validate a summary.

The conceptual contribution is a claim-level model of evidence traceability with distinct error categories. The operational contribution consists of a source register, claim-evidence map, approval outcomes, and change controls. These proposals remain unvalidated and should not be presented as an assurance that AI-generated reports are accurate.

The framework does not replace accounting standards, professional judgement, or sector-specific duties. It has not assessed any commercial AI system, processed a company dataset, or established a performance gain. Literature selection is limited, and practical usefulness will depend on staff capability, source quality, task complexity, and administrative cost.

Further research should test error detection and false corrections, measure total reporting workload, and examine whether users retain responsibility as automation expands. Validation with authorised real-world records may follow controlled synthetic studies. The objective is an adequately supported summary whose efficiency and reliability can both be examined, rather than an unqualified promise of automated accounting quality.

References

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>

OECD. (2021). *The Digital Transformation of SMEs*. OECD Publishing. <https://doi.org/10.1787/bdb9256a-en>

World Bank. (2021). *World Development Report 2021: Data for Better Lives*. World Bank. <https://wdr2021.worldbank.org/>