

AI-Powered Invoice Processing in ERP Systems: A Reference Architecture for Confidence-Governed Touchless Accounts Payable Automation

Manish Kumar ¹, Manykandaprebou Vaitinadin ²

¹Independent Researcher, NC, USA

²Independent Researcher, CA, USA

Correspondence: Manish Kumar (rajpmanish@gmail.com)

Abstract

Manual invoice processing in enterprise Accounts Payable (AP) systems remains time-consuming, error-prone, and difficult to scale in high-volume environments, and a substantial body of recent research has responded with increasingly capable artificial intelligence (AI) methods for extracting structured data from unstructured invoice documents. This literature, however, is concentrated almost entirely on the extraction problem, treating integration with Enterprise Resource Planning (ERP) systems as a thin final step rather than an architectural concern in its own right. This paper addresses that gap by proposing a reference architecture that extends AI-based invoice extraction into a complete, ERP-integrated invoice-to-pay pipeline, comprising ingestion, extraction, validation, three-way matching, tax and general ledger (GL) determination, and idempotent ERP posting. The architecture's central contribution is a confidence-governance layer that aggregates per-stage confidence scores using a weakest-link function and routes each invoice to touchless posting, human review, or rejection according to calibrated thresholds, with human override outcomes feeding back to recalibrate those thresholds over time. The study is developed through Design Science Research, grounded in information quality theory and control theory, and illustrated through a worked scenario demonstrating how a single low-confidence dimension, such as an ambiguous tax determination, appropriately withholds full automation even when extraction confidence is high. Because no proprietary production system was implemented for this study, empirical findings are presented as theoretically derived propositions rather than validated results, explicitly framed as expected outcomes for future empirical work. The paper contributes a reference architecture addressing an underserved layer of the AI-invoice-processing literature, a formal confidence-governance mechanism applicable beyond invoice processing to other AI-augmented enterprise transaction workflows, and practical guidance for enterprises evaluating touchless AP automation.

Keywords: invoice processing; accounts payable automation; enterprise resource planning; touchless processing; confidence governance; intelligent document processing; design science research

1. Introduction

Accounts Payable (AP) departments in large enterprises process invoice volumes that have grown substantially with the expansion of digital commerce and multi-entity, multi-currency operations, while the underlying process, in many organizations, remains anchored in manual data entry: a clerk opens a PDF or scanned invoice, transcribes vendor, amount, and line-item data into an ERP screen, and manually determines the appropriate tax code and general ledger account. This process is slow, costly, and a documented source of posting errors, duplicate payments, and delayed period-close cycles. The practical case for automating it is well established; the academic and technical challenge has been building AI systems capable of reliably extracting structured data from invoices that vary enormously in layout, vendor, language, and format, with no shared template to exploit.

A substantial and rapidly maturing literature now addresses this extraction challenge. Early work applied recurrent neural networks to configuration-free field extraction (Palm, Winther, and Laws 2017), later extended to end-to-end document parsing without word-level labels (Palm, Laws, and Winther 2019). Subsequent research incorporated

document layout explicitly through graph-based representations (Liu et al. 2019; Krieger et al. 2021), and the most recent work applies large language models (LLMs) directly to invoice images and text, reporting substantial accuracy gains over classical optical character recognition (OCR) pipelines (Saout, Lardeux, and Saubion 2024). A nearly contemporaneous study proposes a layered architecture combining OCR, LLM-based analysis, and ERP integration for end-to-end invoice automation underscoring how active this specific combination of technologies has become in the current literature.

This concentration of research effort on extraction accuracy, however, has left a comparatively underdeveloped architectural problem largely unaddressed: what happens after a field is extracted, and with what confidence should an enterprise allow that extraction to trigger an automated financial posting. Existing architectures, including recent ones, typically describe ERP integration as a final, thin layer, an application programming interface (API) call or a structured data push, without formalizing the validation, three-way matching, tax and general ledger (GL) determination, and governance logic that determines whether an extracted invoice is safe to post without human review. This is not a peripheral detail. In enterprise financial control environments, an incorrectly posted invoice, whether due to a misread amount, an unmatched purchase order, or an incorrect tax determination, carries financial, compliance, and audit consequences that a extraction-accuracy metric alone does not capture. The central research problem this paper addresses is therefore not whether AI can extract data from invoices accurately, a question the reviewed literature has substantially answered, but how an enterprise should architect the decision of when extracted, validated, and matched invoice data is trustworthy enough to post without human intervention, commonly termed touchless processing in AP automation practice.

This paper is guided by three research questions.

- RQ1: What architectural components are required to extend AI-based invoice data extraction into a complete, ERP-integrated invoice-to-pay pipeline, including validation, three-way matching, tax determination, and general ledger posting?
- RQ2: How can extraction and downstream validation confidence be formalized into a governance mechanism that determines touchless versus human-in-the-loop processing paths, and what are its failure modes?
- RQ3: What practical and technological requirements does such an architecture impose when embedded within real enterprise ERP platforms?

The paper makes the following contributions.

- C1. A reference architecture extending AI-based invoice extraction into a complete invoice-to-pay pipeline, addressing the ERP-side validation, matching, tax and GL determination, and posting logic that the current extraction-focused literature treats as a thin final step.
- C2. A formal confidence-governance mechanism, grounded in information quality theory, that aggregates multi-stage pipeline confidence via a weakest-link function and routes invoices to touchless posting, human review, or rejection according to calibrated, recalibratable thresholds.
- C3. A control-theoretic account of threshold recalibration, treating human override outcomes as a feedback signal that adjusts governance thresholds over time, analogous to closed-loop control in operational systems.
- C4. An illustrative worked scenario demonstrating the architecture's central design principle, that a single low-confidence pipeline stage should withhold full automation even when other stages, including extraction itself, are highly confident.

The remainder of this paper proceeds as follows. Section 2 reviews the literature thematically. Section 3 develops the theoretical background. Section 4 presents the conceptual framework. Section 5 describes the research methodology. Section 6 presents the proposed reference architecture. Section 7 works through an illustrative scenario. Section 8 discusses theoretical, practical, managerial, technological, and policy implications. Section 9 states expected findings

as propositions. Section 10 details the paper's contributions. Section 11 discusses limitations. Section 12 outlines future research directions. Section 13 concludes.

2. Literature Review

2.1 The Optical Character Recognition and Recurrent Neural Network Era

Early automated invoice processing relied on template-based OCR, which required manually configured extraction rules per vendor layout and broke down when new or unfamiliar layouts appeared. [Palm, Winther, and Laws \(2017\)](#) addressed this directly with CloudScan, a configuration-free system using recurrent neural networks trained across a large, heterogeneous invoice dataset, reporting strong field-level extraction accuracy on both previously seen and entirely unseen invoice layouts without per-vendor configuration. [Palm, Laws, and Winther \(2019\)](#) extended this line of work with an end-to-end architecture trained directly on document-level data, removing the need for word-level annotation and further improving extraction robustness across layout variety. This body of work established the now-standard evaluation framing for invoice extraction research, namely accuracy on seen versus unseen layouts, and remains foundational to the extraction methods surveyed below.

2.2 Layout-Aware and Graph-Based Extraction

A second research stream recognized that invoice fields are informative not only through their text content but through their spatial and structural relationships on the page, motivating graph-based representations of document layout. [Liu et al. \(2019\)](#) proposed a graph convolution approach combining textual and visual information for multimodal extraction from visually rich documents. [Krieger et al. \(2021\)](#) applied a graph neural network specifically to invoice datasets exhibiting high layout variety, demonstrating improved robustness relative to sequence-based models when vendor layouts differ substantially. This layout-aware line of work is directly complementary to the architecture proposed in this paper, since layout and structural confidence signals are a natural input to the extraction-confidence dimension formalized in Section 6.

2.3 Large Language Model-Based Extraction

The most recent research phase applies large language models directly to invoice understanding, either as a replacement for or a complement to traditional OCR. [Saout, Lardeux, and Saubion \(2024\)](#) provide a comprehensive overview of data extraction methods from invoices, documenting the field's progression toward LLM-based approaches and identifying accuracy, robustness to layout variety, and integration with downstream business systems as the field's current priorities. This is, to the author's knowledge, the closest existing published architecture to the one proposed here, and the comparison is instructive: that architecture's integration layer is described functionally, as facilitating data transfer and error reporting, without formalizing the validation, matching, tax determination, or confidence-governance logic that determines what happens between extraction and ERP posting. This paper's contribution is precisely to develop that underspecified segment of the pipeline.

2.4 Enterprise Resource Planning Integration and Accounts Payable Automation Practice

A separate, largely practitioner-oriented literature addresses AP automation more broadly, encompassing robotic process automation (RPA) for repetitive AP tasks, three-way matching automation, and touchless processing as an operational target. This practitioner literature consistently identifies touchless processing rate, the percentage of invoices posted without human intervention, as a key performance indicator for AP automation maturity, but rarely formalizes the decision logic determining when an invoice is eligible for touchless treatment beyond simple rule-based thresholds on individual fields. Existing enterprise tax engine integration practice, such as external tax determination services used in ERP tax configuration, similarly formalizes tax and rate determination logic without an explicit, unified confidence framework connecting that determination to upstream extraction and matching confidence.

2.5 Research Gap

Table 1 summarizes the literature reviewed above against the dimensions relevant to this paper's contribution. The extraction problem, across OCR, graph-based, and LLM-based methods, is well addressed by prior work, including a near-contemporaneous architecture combining these methods with ERP integration. What remains comparatively unaddressed is the formal architecture and governance logic connecting a confident extraction to a trustworthy automated posting: three-way matching, tax and GL determination, and, centrally, a principled mechanism for deciding when the accumulated evidence across all pipeline stages justifies touchless processing rather than human review.

Study	Extraction method	Formalizes ERP-side matching/tax logic	Formalizes confidence governance	Addresses touchless routing
Palm, Winther, and Laws (2017)	RNN	X	X	X
Palm, Laws, and Winther (2019)	End-to-end neural	X	X	X
Liu et al. (2019)	Graph convolution	X	X	X
Krieger et al. (2021)	Graph neural network	X	X	X
Saout, Lardeux, and Saubion (2024)	Survey (OCR to LLM)	Partial (survey)	X	Partial (survey)
This paper	Extraction-agnostic (input layer)	✓	✓	✓

Table 1. Comparison of prior invoice-processing studies against the dimensions relevant to this paper's contribution.

3. Theoretical Background

3.1 Information Quality Theory

Wang and Strong's (1996) information quality framework distinguishes intrinsic, contextual, representational, and accessibility dimensions of data quality, arguing that data quality cannot be reduced to accuracy alone but must be assessed relative to the context in which the data will be used. This framing is directly applicable to invoice processing: an extracted field may be intrinsically accurate, a correctly transcribed amount, while remaining contextually invalid for automated posting, for instance because it does not correspond to any open purchase order. This paper adopts information quality theory as its primary theoretical grounding for treating confidence as inherently multi-dimensional, extraction accuracy, validation against business rules, matching against transactional context, and tax or GL determination correctness, rather than collapsing it into a single extraction-accuracy metric, which is the implicit framing of most of the literature reviewed in Section 2.

3.2 Control Theory as a Second Kernel

Information quality theory motivates why confidence should be multi-dimensional but does not, on its own, specify how a governance threshold should be set or adjusted over time. This paper adopts control theory as a second theoretical kernel, following Ivanov et al.'s (2018) treatment of operational systems as control-theoretic objects subject to continuous disturbance. Under this framing, the governance thresholds separating touchless posting, human review, and rejection function as a control boundary, and human override outcomes, cases where a human reviewer corrects,

confirms, or rejects a routed invoice, function as an error signal indicating whether the current threshold is miscalibrated relative to the true reliability of the pipeline's confidence estimates. This connects the paper's architecture directly to its central design claim: governance thresholds are not fixed configuration parameters set once at implementation but a continuously recalibrated control boundary, formalized in Section 6.

3.3 Design Science Framing

Consistent with the Design Science Research (DSR) paradigm (Hevner et al. 2004) and operationalized through the process model of Peffers et al. (2007), this paper treats the proposed architecture as a design artifact combining a model, the formal confidence-aggregation and routing logic in Section 6, and a method, the eight-layer reference architecture and its constituent components. As a conceptual contribution, this paper corresponds to the problem identification, objective definition, and design activities of the DSRM process; the demonstration and evaluation activities, implementing and empirically validating the architecture within a real ERP environment, are identified in Section 12 as the necessary next stage.

4. Conceptual Framework

Figure 1 presents the conceptual framework linking the theoretical grounding in Section 3 to the architecture developed in Section 6. Four pipeline-stage confidence dimensions, extraction, validation, matching, and tax and GL determination, are proposed, directly operationalizing the multi-dimensional information quality framing of Section 3.1 rather than treating confidence as a single extraction-accuracy score.

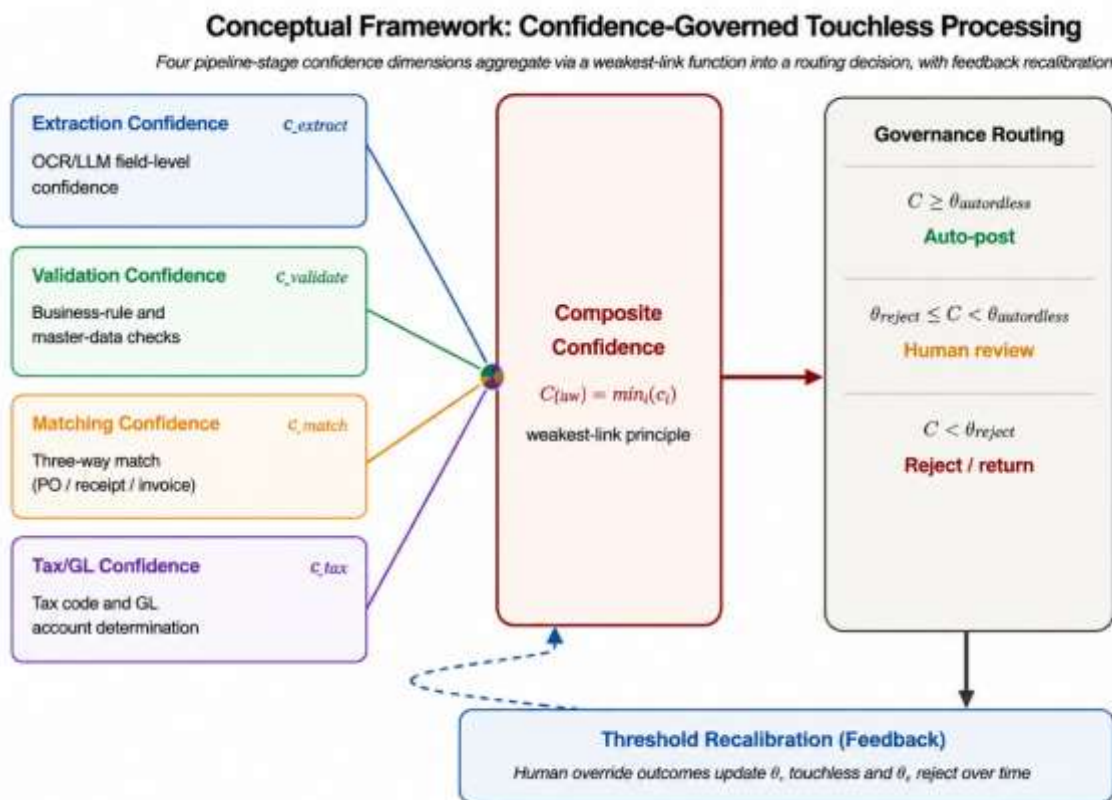


Figure 1. Conceptual framework for confidence-governed touchless invoice processing, linking four pipeline-stage confidence dimensions through a weakest-link composite function to governance routing, with a control-theoretic feedback loop recalibrating thresholds from human override outcomes.

The framework's central design decision, developed formally in Section 6.6, is that these four dimensions are aggregated via a weakest-link function rather than an averaged or weighted-sum score. This reflects a specific, defensible design rationale: a highly confident extraction does not compensate for a failed three-way match, since posting an invoice with no corresponding purchase order is a control failure regardless of how accurately the invoice's text was read. The

resulting composite confidence score is compared against calibrated governance thresholds, and the control-theoretic feedback loop described in Section 3.2 closes the loop by using human override outcomes, instances where a human reviewer's decision differs from what the governance layer would have predicted, to adjust those thresholds over time.

5. Research Methodology

This paper adopts a conceptual Design Science Research methodology, developing a reference architecture and its constituent formal mechanisms through synthesis of the extraction literature reviewed in Section 2 and the theoretical grounding developed in Section 3, illustrated through a worked scenario rather than empirical validation on production data. This methodological choice reflects the paper's objective directly: given that the extraction problem is well addressed by existing empirical literature, discussed in Section 2, the paper's contribution is architectural and governance-theoretic rather than a new extraction model requiring a labeled invoice dataset and accuracy benchmarking. No original data collection was undertaken; the worked scenario in Section 7 uses illustrative, clearly labeled synthetic values to demonstrate the governance mechanism's behavior, not to claim measured system performance.

Table 2 maps the paper's research questions to the methodological approach used to address each.

Research question	Methodological approach	Section
RQ1: architectural components for a complete invoice-to-pay pipeline	Literature synthesis; design science artifact development (reference architecture)	6
RQ2: confidence governance mechanism and failure modes	Formal specification (weakest-link aggregation, threshold routing); control-theoretic feedback design	4, 6.6
RQ3: practical and technological ERP integration requirements	Design science analysis; illustrative scenario	7, 8.3

Table 2. Mapping of research questions to methodological approach.

6. Proposed Reference Architecture

Figure 2 presents the proposed eight-layer reference architecture for AI-powered, ERP-integrated invoice processing.

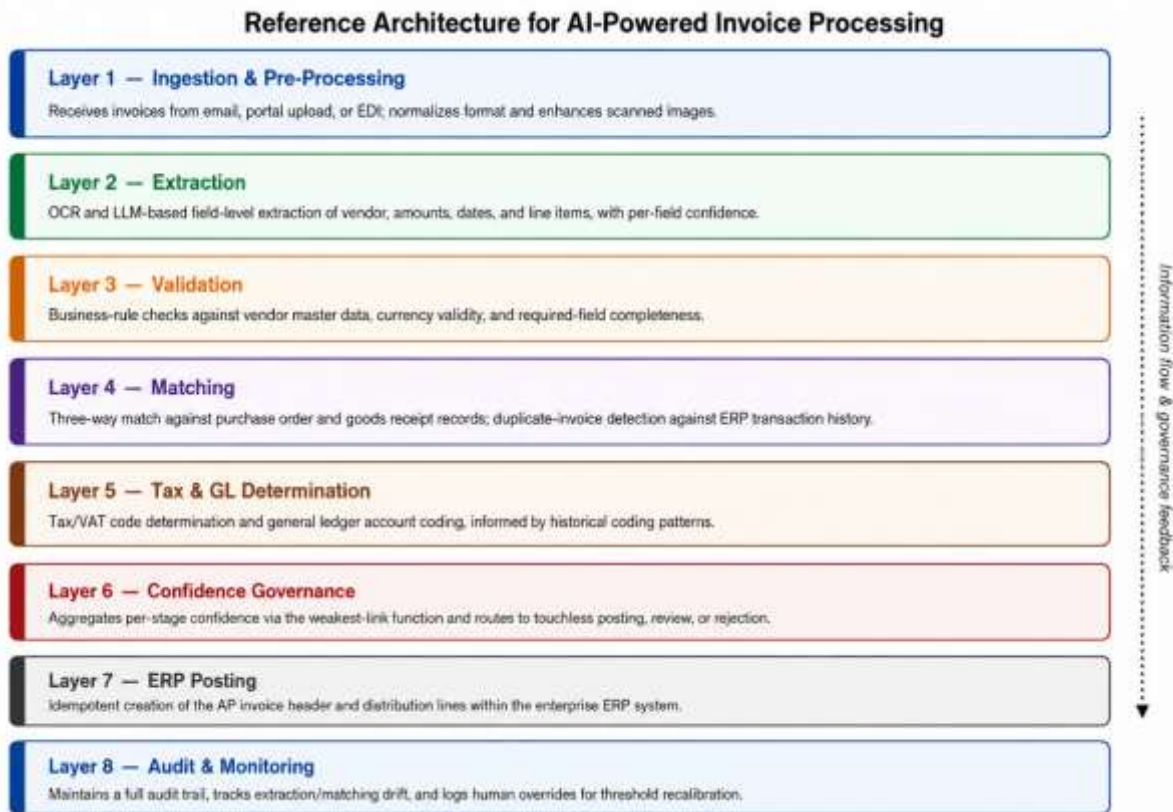


Figure 2. Eight-layer reference architecture for AI-powered invoice processing, from ingestion through ERP posting and audit, with a governance feedback loop from Layer 8 to the Layer 6 confidence thresholds.

6.1 Ingestion and Pre-Processing

This layer receives invoices from heterogeneous sources, including email attachments, vendor self-service portal uploads, and electronic data interchange (EDI) feeds, and normalizes them into a consistent internal representation, including image enhancement for scanned documents. This layer is functionally consistent with equivalent stages in the extraction literature reviewed in Section 2 and is not itself a novel contribution of this paper.

6.2 Extraction

This layer performs field-level extraction using OCR and LLM-based methods, consistent with the current state of the art documented in Section 2.3, producing structured field values together with a per-field extraction confidence score, c_{extract} . This paper is deliberately extraction-method-agnostic: the governance architecture developed in Section 6.6 is designed to consume a confidence score from any of the extraction approaches reviewed in Section 2, whether RNN-based (Palm, Winther, and Laws 2017), graph-based (Krieger et al. 2021), rather than proposing a competing extraction model.

6.3 Validation

This layer applies business-rule checks against extracted fields, including vendor master data matching, currency and date validity, and required-field completeness, producing a validation confidence score, c_{validate} . Vendor master mismatches, for instance an extracted vendor name that does not correspond to an approved vendor record, are treated as a validation-confidence penalty rather than an automatic rejection, since new or recently onboarded vendors are a legitimate, if less common, case.

6.4 Matching

This layer performs three-way matching, reconciling the invoice against the referenced purchase order and goods receipt records, and duplicate-invoice detection against existing ERP transaction history, producing a matching confidence score, c_{match} . This layer is, in the author's assessment, the most consequential and least formalized in the existing literature reviewed in Section 2, since a failure at this layer, an invoice with no corresponding purchase order or an apparent duplicate, represents a control and fraud-risk concern that extraction accuracy alone cannot surface.

6.5 Tax and General Ledger Determination

This layer determines the applicable tax or value-added tax (VAT) treatment and the general ledger (GL) account coding for the invoice, informed by historical coding patterns for the given vendor and goods or service category, producing a tax and GL confidence score, c_{tax} . In enterprise environments using external tax determination engines, this layer would interface with that engine's rate and taxability logic rather than reimplementing it, consistent with standard enterprise tax engine integration practice.

6.6 Confidence Governance

This is the architecture's central contribution. The composite confidence score for an invoice, $C(inv)$, is defined as the minimum of the four per-stage confidence scores,

$$C(inv) = \min(c_{extract}, c_{validate}, c_{match}, c_{tax})$$

reflecting the weakest-link design rationale introduced in Section 4: automation eligibility is gated by the least reliable pipeline stage, not averaged across stages. Two calibrated thresholds, $\theta_{touchless}$ and θ_{reject} , with $\theta_{reject} < \theta_{touchless}$, partition the composite confidence score into three routing zones.

Condition	Routing decision	Rationale
$C(inv) \geq \theta_{touchless}$	Auto-post (touchless)	All pipeline stages sufficiently confident; no control benefit from human review
$\theta_{reject} \leq C(inv) < \theta_{touchless}$	Route to human review queue	At least one stage below the touchless threshold; human judgment adds value
$C(inv) < \theta_{reject}$	Reject or return to vendor/requester	Confidence too low to responsibly present for review as a near-complete invoice

Table 3. Confidence governance routing rules.

Threshold calibration is not treated as a one-time configuration decision. Consistent with the control-theoretic framing in Section 3.2, Layer 8 (Section 6.8) logs every human override, cases where a reviewer confirms, corrects, or rejects an invoice routed to the review queue, or, more rarely, where a downstream error is discovered in a touchless-posted invoice. These override outcomes constitute an error signal usable to adjust $\theta_{touchless}$ and θ_{reject} over time: a pattern of reviewers routinely confirming invoices routed for review with no correction suggests $\theta_{touchless}$ may be set conservatively low relative to the pipeline's true reliability, while any confirmed error in a touchless-posted invoice is a strong signal to raise $\theta_{touchless}$. This paper does not specify a particular recalibration algorithm, treating this as an implementation detail appropriately left to Section 12, but the architectural requirement, that the governance layer's thresholds are recalibratable from Layer 8's audit signal rather than static, is a design commitment of the proposed architecture.

6.7 Enterprise Resource Planning Posting

This layer creates the AP invoice header and distribution lines within the enterprise ERP system for touchless-routed invoices, and for review-routed invoices following human confirmation. Consistent with idempotency principles

established in prior enterprise integration architecture work, posting operations should be designed to be safely retryable without creating duplicate ERP records, since network or system failures during posting are a realistic operational condition in high-volume environments.

6.8 Audit and Monitoring

This layer maintains a complete audit trail of every invoice's path through the pipeline, including per-stage confidence scores, the routing decision, and any human override, and monitors for extraction and matching drift, systematic degradation in confidence scores or increased override rates for a given vendor, document format, or time period. This layer's audit signal is the mechanism by which the control-theoretic feedback loop described in Section 3.2 and Section 6.6 is operationalized.

7. Illustrative Scenario

The following scenario is illustrative only, using clearly labeled synthetic values to demonstrate the governance mechanism's behavior; it does not represent measured performance from an implemented system, consistent with the methodological scope stated in Section 5.

Consider an invoice referencing purchase order PO-33021, for a shipment of specialized laboratory equipment, with a total amount of \$12,450.00. The extraction layer (Section 6.2) reports high confidence, $c_{\text{extract}} = 0.97$, having correctly read all header and line-item fields from a clean, digitally native PDF. The validation layer (Section 6.3) confirms the vendor matches an approved master record and all required fields are present, $c_{\text{validate}} = 0.96$. The matching layer (Section 6.4) finds an exact match against the referenced purchase order and a corresponding goods receipt, $c_{\text{match}} = 0.98$. The tax and GL determination layer (Section 6.5), however, encounters an ambiguity: the goods category for specialized laboratory equipment maps to two plausible tax treatments in the organization's historical coding data, and the layer reports a comparatively low confidence, $c_{\text{tax}} = 0.71$, reflecting genuine ambiguity rather than a data quality problem.

Under an averaging or weighted-sum aggregation scheme, this invoice's composite confidence might be computed as a weighted mean well above a typical touchless threshold, given three near-perfect scores and only one moderate score, potentially triggering automated posting with an incorrect or uncertain tax determination. Under the weakest-link aggregation specified in Section 6.6, however, $C(\text{inv}) = \min(0.97, 0.96, 0.98, 0.71) = 0.71$. If the organization's calibrated $\theta_{\text{touchless}}$ is, illustratively, 0.90, this invoice is routed to human review specifically because of the tax ambiguity, despite near-perfect extraction, validation, and matching confidence. This is the architecture's central design principle made concrete: a single weak dimension appropriately withholds full automation, directing human attention precisely to the dimension that requires it, tax determination, rather than requiring a reviewer to re-examine the entire invoice from scratch.

8. Analysis and Discussion

8.1 Theoretical Implications

The central theoretical contribution of this paper is the reframing of touchless processing eligibility as a multi-dimensional information quality problem, following Wang and Strong (1996), rather than a single extraction-accuracy threshold, which is the implicit framing throughout the extraction literature reviewed in Section 2. This has an implication beyond invoice processing specifically: any AI-augmented enterprise transaction workflow in which multiple independent evidentiary dimensions must all be sound for safe automation, such as expense report auditing, purchase requisition approval, or vendor onboarding, is a candidate for the same weakest-link governance framing developed here, suggesting the architecture's governance layer, rather than its invoice-specific pipeline layers, may generalize furthest.

8.2 Practical and Managerial Implications

For AP managers, the architecture offers a principled alternative to configuring touchless eligibility through ad hoc per-field thresholds, which is common in current practice and does not account for the interaction between extraction quality and downstream matching or tax risk. The weakest-link principle has a direct, explainable managerial implication: when an invoice is routed for review, the specific low-confidence dimension can be surfaced to the reviewer directly, reducing review time relative to an undifferentiated review queue in which reviewers must re-verify every field regardless of which one triggered the review.

8.3 Technological Implications

The architecture implies specific technological requirements beyond the extraction models reviewed in Section 2: a matching engine capable of querying live purchase order and goods receipt data from the ERP system, an audit data store capturing per-stage confidence and routing decisions at sufficient granularity to support the recalibration process described in Section 6.6, and an override-capture interface through which human reviewers' decisions are logged in a form usable as the control-theoretic feedback signal. None of these components is individually novel, but their integration into a unified, confidence-governed pipeline, rather than as separate, loosely coupled systems, is not, to the author's knowledge, addressed as a formal architecture in the literature reviewed in Section 2.

8.4 Policy and Compliance Implications

The confidence-governance architecture has a direct compliance implication: because every routing decision is grounded in an auditable, per-stage confidence score rather than an opaque single model output, the architecture supports the kind of explainability that internal and external auditors increasingly expect from AI-augmented financial processes. A rejected or review-routed invoice's audit trail shows precisely which pipeline stage fell below threshold, which is a materially stronger compliance posture than a system that reports only a single aggregate confidence or acceptance decision.

8.5 Comparison with Existing Approaches

Table 4 compares the proposed architecture against the two most closely related prior approaches: extraction-focused academic literature and current AP automation practice.

Dimension	Extraction-focused literature (Section 2.1–2.3)	Current AP automation practice	This paper
Confidence treated as multi-dimensional	No (extraction accuracy only)	Partially (ad hoc per-field rules)	Yes, formalized (Section 6.6)
Formal aggregation logic	Not applicable	Not typically formalized	Weakest-link function
Threshold recalibration	Not addressed	Rarely automated	Control-theoretic feedback (Section 3.2)
ERP-side matching and tax logic formalized	No	Yes, but as configuration, not architecture	Yes, as architectural layers

Table 4. Comparison of the proposed architecture with existing approaches.

9. Expected Findings (Theoretical Propositions)

As a conceptual paper, this work does not report empirical findings. The following propositions are logically derived from the architecture in Section 6 and are intended to guide the empirical validation identified in Section 12, not to be read as established results.

- Proposition 1: Weakest-link confidence aggregation will route a materially different, and smaller, set of invoices to touchless posting than an averaged or weighted-sum aggregation applied to the same underlying per-stage confidence scores, with the difference concentrated in invoices exhibiting exactly one low-confidence stage.
- Proposition 2: Touchless processing rate and post-posting error rate will exhibit a measurable trade-off as $\theta_{\text{touchless}}$ is varied, and this trade-off curve will differ meaningfully depending on whether weakest-link or averaged aggregation is used, with weakest-link aggregation producing a more favorable trade-off at comparable touchless rates.
- Proposition 3: Review time per invoice will be lower when reviewers are shown the specific low-confidence dimension that triggered routing, compared to an undifferentiated review interface, consistent with the practical implication discussed in Section 8.2.
- Proposition 4: Threshold recalibration driven by human override outcomes will converge toward stable threshold values over a bounded number of override observations, rather than oscillating indefinitely, under reasonable assumptions about the stability of the underlying extraction and matching model performance.

10. Contributions

Theoretical contribution

The paper introduces a multi-dimensional, information-quality-grounded reformulation of touchless processing eligibility, extending Wang and Strong's (1996) framework into a new enterprise AI governance context, and connects threshold calibration to control theory (Ivanov et al. 2018), a combination not previously applied to invoice processing specifically.

Practical contribution

The proposed reference architecture and its weakest-link governance mechanism (Section 6.6) provide AP automation practitioners with a principled, explainable alternative to ad hoc touchless-eligibility rules.

Managerial contribution

The architecture's audit trail design (Section 6.8) enables reviewers to be directed to the specific dimension requiring attention, with implications for review efficiency discussed in Section 8.2.

Technological contribution

The paper specifies concrete technological requirements for a unified, confidence-governed invoice pipeline (Section 8.3), identifying an integration gap between existing extraction, matching, and ERP-posting components.

Policy and compliance contribution

The paper's explicit, per-stage audit trail design (Section 6.8, Section 8.4) offers a stronger compliance posture for AI-augmented financial processes than architectures reporting only an aggregate acceptance decision.

11. Limitations

This paper's limitations follow directly from its conceptual, design-science scope. Methodologically, the architecture has not been empirically tested; the propositions in Section 9 are logically derived, not validated, and the illustrative scenario in Section 7 uses synthetic values explicitly for demonstration rather than measurement. Data-related

limitations follow from the absence of original data collection: no labeled invoice dataset was constructed or evaluated, and the extraction-layer confidence scores the architecture consumes are assumed to be well-calibrated outputs of the methods reviewed in Section 2, an assumption itself not tested here. Generalizability is correspondingly limited: the four-dimension confidence structure and weakest-link aggregation are architecturally motivated but their applicability may vary across ERP platforms with different native matching and tax-determination capabilities, and across industries with different invoice complexity profiles. Implementation limitations apply to the recalibration mechanism described in Section 6.6, which specifies a functional requirement, that thresholds be recalibratable from override data, without specifying a validated recalibration algorithm.

12. Future Research Directions

- Empirical validation: implementing the proposed architecture against a real or realistic invoice dataset with ground-truth purchase order and tax determination data, testing Propositions 1 and 2 directly by comparing weakest-link against averaged aggregation on touchless rate and error rate.
- Industry case studies: qualitative case studies of enterprises implementing confidence-governed touchless processing, examining how threshold recalibration behaves in practice and whether Proposition 4's convergence expectation holds.
- System implementation: development of a reference implementation of the eight-layer architecture (Section 6) integrated with a specific ERP platform, as a precondition for the empirical validation above.
- AI and machine learning extensions: investigating whether the per-stage confidence scores themselves, rather than only their aggregation, can be jointly calibrated using techniques from uncertainty quantification, extending beyond the extraction-model-agnostic treatment adopted in this paper.
- Cross-industry and cross-platform comparison: examining whether optimal governance thresholds and recalibration dynamics differ systematically across industries with different invoice complexity, such as services versus goods procurement, or across ERP platforms with different native matching capabilities.

13. Conclusion

This paper has argued that the rapidly maturing literature on AI-powered invoice extraction, while substantially advancing the accuracy with which structured data can be pulled from unstructured invoice documents, has left the architecture connecting that extraction to trustworthy, automated ERP posting comparatively underdeveloped. Grounding the problem in information quality theory (Wang and Strong 1996) and control theory (Ivanov et al. 2018), this paper proposed an eight-layer reference architecture extending invoice extraction into a complete invoice-to-pay pipeline, with a confidence-governance layer as its central contribution: a weakest-link aggregation of extraction, validation, matching, and tax and GL determination confidence, routed through calibrated, recalibratable thresholds to touchless posting, human review, or rejection. An illustrative scenario demonstrated the architecture's central design principle concretely, that a single low-confidence dimension appropriately withholds full automation even when other pipeline stages are highly confident. For researchers, the paper's primary contribution is a formal governance mechanism addressing a gap between the mature extraction literature and enterprise financial control requirements; for practitioners, its primary contribution is an explainable, auditable alternative to ad hoc touchless-eligibility configuration. Empirical validation of the propositions advanced in Section 9, through the reference implementation and case studies identified in Section 12, remains the necessary next step before this architecture's practical value can be established with confidence.

References

- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Ivanov, D., Sethi, S., Dolgui, A., & Sokolov, B. (2018). A survey on control theory applications to operational systems, supply chain management, and Industry 4.0. *Annual Reviews in Control*, 46, 134–147. <https://doi.org/10.1016/j.arcontrol.2018.10.003>
- Krieger, F., Drews, P., Funk, B., & Wobbe, T. (2021). Information extraction from invoices: A graph neural network approach for datasets with high layout variety. In F. Ahlemann, R. Schütte, & S. Stieglitz (Eds.), *Innovation through information systems: Volume II: A collection of latest research on technology issues* (pp. 5–20). Springer. https://doi.org/10.1007/978-3-030-86797-3_1
- Liu, X., Gao, F., Zhang, Q., & Zhao, H. (2019). Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)* (pp. 32–39). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-2005>
- Palm, R. B., Laws, F., & Winther, O. (2019). Attend, copy, parse: End-to-end information extraction from documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 329–336). IEEE. <https://doi.org/10.1109/ICDAR.2019.00060>
- Palm, R. B., Winther, O., & Laws, F. (2017). CloudScan—A configuration-free invoice analysis system using recurrent neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 406–413). IEEE. <https://doi.org/10.1109/ICDAR.2017.74>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Saout, T., Lardeux, F., & Saubion, F. (2024). An overview of data extraction from invoices. *IEEE Access*, 12, 19872–19886. <https://doi.org/10.1109/ACCESS.2024.3360528>
- 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>